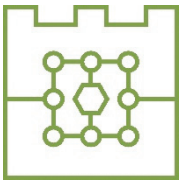
Politechnika Krakowska
im. Tadeusza Kościuszki

Wydział Informatyki i Matematyki



Rozprawa doktorska

Grzegorz Nowakowski

Formalne i algorytmiczne metody przetwarzania informacji nieprecyzyjnej w systemach informatycznych

Formal and algorithmic methods for processing imprecise information in information systems

Promotor:
dr hab. inż. Paweł Pławiak, prof. PK

Kraków, 2026

Pragnę złożyć serdeczne podziękowania

Panu ***dr hab. inż. Pawłowi Pławiakowi, prof. PK*** *za to, że pomógł mi dostrzec drogę tam, gdzie początkowo była tylko myśl. Podsunięta we właściwym momencie możliwość stała się początkiem drogi, która nabrała własnego kształtu i doprowadziła do powstania niniejszej rozprawy doktorskiej.*

Panu ***prof. dr hab. inż. Sergiiu Telenykowi*** *za naukową drogę, która dzięki Jego obecności przez lata była dla mnie źródłem rozwoju, odwagi i poczucia, że w nauce najważniejsze rzeczy rodzą się nie tylko z intelektualnego wysiłku, lecz także z ludzkiej życzliwości.*

Pani ***Marii Madej*** *za uważność i wnikliwość, z jakimi czytała kolejne fragmenty rozprawy, za cierpliwą obecność na tej naukowej drodze oraz za pomoc, dzięki której moje myśli mogły zyskać większą przejrzystość i dojrzalszy wyraz, a także za przyjaźń i życzliwość, na które zawsze mogłem liczyć.*

Pani ***Katarzynie Pol*** *za to, że potrafiła nadać bieg temu, co dopiero miało stać się drogą, lecz przez długi czas pozostawało w bezruchu. To dzięki Niej doszło do spotkania z Promotorem, od którego ta droga mogła się rozpocząć.*

Podziękowania składam również wszystkim, którzy na różnych etapach tej drogi inspirowali mnie i motywowali do dalszego wysiłku.

Babci Helenie - in memoriam

Streszczenie

Celem rozprawy „Formalne i algorytmiczne metody przetwarzania informacji nieprecyzyjnej w systemach informatycznych” było opracowanie oraz weryfikacja rozwiązań umożliwiających przejście od informacji nieprecyzyjnej, rozumianej szeroko jako nieostrość, niepełność, zaszumienie lub kontekstowość, do postaci formalnie określonej i algorytmicznie przetwarzalnej w realnych środowiskach obliczeniowych. Przedstawiony cykl publikacji potwierdza, że dla różnych postaci nieprecyzyjności możliwe jest budowanie spójnych rozwiązań spełniających trzy warunki: (I) jawna formalizacja sposobu interpretacji informacji rozumiana jako określenie modelu i semantyki przetwarzania albo jednoznacznej funkcji celu, (II) algorytmiczna realizacja formalizacji w postaci procedury możliwej do uruchomienia w systemie informatycznym (np. transformacja, wnioskowanie, rekonstrukcja lub uczenie) oraz (III) empiryczna weryfikacja jakości i kosztu przetwarzania w warunkach systemowych i/lub eksperymentalnych.

Weryfikację tezy sformułowanej w podrozdziale 2.2 przeprowadzono w czterech obszarach badawczych I–IV. W obszarze I przedstawiono wykonalność zapytań rozmytych w relacyjnych bazach danych poprzez formalizację semantyki stopniowej oraz transformację do standardowego SQL. W obszarze II ustalono, że niepełna specyfikacja przebiegu procesu przetwarzania może zostać uzupełniona przez formalny opis usług, wnioskowanie i rekonstrukcję planu wykonania. W obszarze III zaprezentowano możliwość redukcji nieprecyzyjności wymagań projektowych przez formalizację opisu komponentów i reguł ich łączenia oraz dedukcyjną syntezę struktury aplikacji. W obszarze IV potwierdzono, że ten sam schemat badawczy: formalizacja → algorytm → weryfikacja pozostaje spójny również dla innych nośników nieprecyzyjności, w tym sygnałów zewnętrznych, braków struktury w tekście oraz formalizacji optymalizacyjnej prowadzącej do kompresji informacji w danych syntetycznych. W każdym z obszarów ocena obejmowała weryfikację jakości oraz, tam gdzie było to istotne, analizę kosztu obliczeniowego z użyciem miar adekwatnych do danej domeny.

Rozprawa dostarcza spójnego ujęcia przetwarzania informacji nieprecyzyjnej, w którym jawna semantyka lub jednoznaczna funkcja celu jest konsekwentnie przekładana na wykonalne procedury algorytmiczne i poddawana empirycznej weryfikacji jakości oraz kosztu. Wkład własny autora obejmuje opracowanie metod uruchamiania semantyki stopniowej w standardowym SQL, sformułowanie formalno-algorytmicznego przejścia od celu użytkownika do wykonalnego planu integracji usług, zaproponowanie mechanizmu syntezy

struktury aplikacji w procesie projektowania oraz rozszerzenie schematu badawczego na różne nośniki nieprecyzyjności, w tym o formalizację optymalizacyjną w uczeniu maszynowym.

Wyniki rozprawy potwierdzają postawioną tezę, zgodnie z którą formalne modele oraz algorytmiczne metody przetwarzania informacji nieprecyzyjnej umożliwiają skuteczne i obliczeniowo efektywne wykorzystanie informacji niejednoznacznych, niepełnych, zaszumionych i kontekstowych w systemach informatycznych, pod warunkiem jawnego zdefiniowania semantyki przetwarzania lub funkcji celu oraz empirycznej weryfikacji jakości i kosztu.

Słowa kluczowe: *informacja nieprecyzyjna, formalizacja semantyki, algorytmy przetwarzania informacji, zapytania rozmyte i transformacja do SQL, integracja usług i plan wykonania, baza wiedzy i wnioskowanie, formalizacja optymalizacyjna i destylacja danych*

Summary

The aim of the dissertation ‘Formal and Algorithmic Methods for Processing Imprecise Information in Information Systems’ was to develop and validate solutions that enable a transition from imprecise information − understood broadly as fuzziness, incompleteness, noise, or context dependence − to a form that is formally specified and algorithmically processable in real computational environments. The presented publication cycle confirms that, for different manifestations of imprecision, coherent solutions can be constructed that satisfy three conditions: (I) an explicit formalisation of how information is interpreted understood as defining a processing model and semantics or an unambiguous objective function, (II) an algorithmic realisation of that formalisation as an executable procedure within an information system (e.g., transformation, inference, reconstruction, or learning) and (III) empirical verification of processing quality and cost under system and/or experimental conditions.

The thesis formulated in Section 2.2 was verified in four research areas (I–IV). In Area I, the feasibility of fuzzy queries in relational databases was presented by formalising graded semantics and transforming fuzzy queries into standard SQL. In Area II, it was established that an incomplete specification of a processing workflow can be complemented by a formal description of services, inference, and reconstruction of an execution plan. In Area III, the possibility of reducing imprecision in design requirements was demonstrated through formalisation of component descriptions and composition rules, followed by deductive synthesis of the application structure. In Area IV, it was confirmed that the same research pattern: formalisation → algorithm → verification remains coherent for other carriers of imprecision, including external signals, missing structure in text, and optimisation-based formalisation leading to compression of information in synthetic data. In each area, the evaluation included quality verification and, where relevant, analysis of computational cost using metrics appropriate to the given domain.

The dissertation provides a coherent framework for processing imprecise information in which explicit semantics or an unambiguous objective function is consistently translated into executable algorithmic procedures and subjected to empirical evaluation of quality and cost. The author’s contributions include methods for executing graded semantics in standard SQL, a formal-and-algorithmic transition from a user goal to an executable service-integration plan, a mechanism for synthesising application structure at the design stage, and an extension of the research pattern to multiple carriers of imprecision, including optimisation-based formalisation in machine learning.

The results confirm the dissertation's thesis that formal models and algorithmic methods for processing imprecise information enable effective and computationally efficient use of ambiguous, incomplete, noisy, and context-dependent information in information systems, provided that the processing semantics or objective function is explicitly defined and that quality and cost are empirically verified.

Keywords**:** *imprecise information, semantic formalisation, information processing algorithms, fuzzy queries and SQL transformation, service integration and execution plan, knowledge base and inference, optimisation-based formalisation and dataset distillation*

Spis treści

1. Wykaz publikacji stanowiących podstawę rozprawy

[1] Nowakowski G., *Fuzzy queries on relational databases.*

Publikacja: 2018 International Interdisciplinary PhD Workshop (IIPhDW 2018), 2018, Institute of Electrical and Electronics Engineers, IEEE, pp. 293-299, ISBN 978-1-5386-6143-7, doi: 10.1109/IIPHDW.2018.8388376.

Typ: publikacja konferencyjna (artykuł w materiałach konferencyjnych). **Rok**: 2018. **MNiSW**: 20 pkt. **Cytowania**: Scopus − 11; Google Scholar − 14.

Wkład autora rozprawy: 100%. **Rola**: autor korespondencyjny.

Upowszechnienie: referat konferencyjny, 8th IIPhDW 2018, 9–12.05.2018, Świnoujście, Polska.

Zakres wkładu: sformułowanie problemu i celu pracy; opracowanie formalnego ujęcia zapytań rozmytych w relacyjnych bazach danych, w tym semantyki stopniowej oraz miary dopasowania krotek; zaprojektowanie sposobu wykorzystania tej miary do filtrowania wyników (przekroje α) oraz ich porządkowania według stopnia spełnienia warunków; przygotowanie zestawu reprezentatywnych klas zapytań; przeprowadzenie oceny wykonalności i narzutu obliczeniowego; opracowanie i interpretacja wyników; przygotowanie manuskryptu oraz prezentacja wyników w formie referatu na konferencji IIPhDW 2018.

[2] Nowakowski G., *Methodology of transformation of fuzzy queries into queries in the SQL standard*.

Publikacja: IDAACS 2019: proceedings of the 2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), vol. 2, 2019, Institute of Electrical and Electronics Engineers, IEEE, pp. 674-679, ISBN 978-1-7281-4069-8, doi: 10.1109/IDAACS.2019.8924312.

Typ: publikacja konferencyjna (artykuł w materiałach konferencyjnych). **Rok**: 2019. **MNiSW**: 20 pkt. **Cytowania**: Web of Science – 1; Scopus – 3; Google Scholar − 5.

Wkład autora rozprawy: 100%. **Rola**: autor korespondencyjny.

Upowszechnienie: referat konferencyjny, IDAACS 2019, 18–21.09.2019, Metz, Francja.

Zakres wkładu: sformułowanie celu pracy jako metodyki przejścia od zapytań rozmytych do postaci wykonywalnej w standardowym SQL oraz opracowanie spójnego opisu dwóch wariantów realizacyjnych, tj. realizacji po stronie DBMS oraz po stronie aplikacji; opracowanie algorytmicznego ujęcia analizy warunków zapytania, w tym klauzuli WHERE

obejmującego konstrukcję drzewa wyrażeń oraz rekurencyjne wyznaczanie stopni spełnienia dla warunków złożonych zgodnie z przyjętą semantyką operatorów łączenia ($T-\text{norma}$ / $S-\text{norma}$); opracowanie części porównawczej i weryfikacyjnej metodyki; przygotowanie manuskryptu oraz prezentacja wyników w formie referatu na konferencji IDAACS 2019.

[3] Telenyk S., **Nowakowski G.**, Yefremov K., Khmeliuk V., *Logics based application integration for interdisciplinary scientific investigations*.

Publikacja: IDAACS 2017: proceedings of the 2017 IEEE 9th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), vol. 2, 2017, Institute of Electrical and Electronics Engineers, IEEE, pp. 1026–1031, ISBN 978-1-5386-0697-1, doi: 10.1109/IDAACS.2017.8095241.

Typ: publikacja konferencyjna (artykuł w materiałach konferencyjnych). **Rok**: 2017. **MNiSW**: 20 pkt. **Cytowania**: Web of Science – 3; Scopus − 6; Google Scholar − 9.

Wkład autora rozprawy: 55%. **Rola**: autor korespondencyjny.

Upowszechnienie: referat konferencyjny, IDAACS 2017, 21–23.09.2017, Bukareszt, Rumunia.

Zakres wkładu: współtworzenie koncepcji integracji usług przy niepełnej specyfikacji celu użytkownika, w tym w zakresie opracowania aparatu formalnego metody; przygotowanie kluczowych elementów formalizacji (definicje, relacje oraz reguły wnioskowania) oraz opracowanie opisu elementów algorytmicznych prowadzących od specyfikacji celu do konstrukcji planu wykonania, w tym mechanizmu rekonstrukcji struktury rozwiązania na podstawie wnioskowania; przygotowanie znaczącej części manuskryptu w obszarze formalizacji i algorytmów; integracja uwag współautorów; pełnienie roli autora korespondencyjnego i koordynacja procesu publikacyjnego; reprezentowanie zespołu oraz prezentacja wyników w formie referatu na konferencji IDAACS 2017.

[4] **Nowakowski G.**, Telenyk S., Yefremov K., Khmeliuk V., *The approach to applications integration for World Data Center interdisciplinary scientific investigations*.

Publikacja: Proceedings of the 2019 Federated Conference on Computer Science and Information Systems (FedCSIS), Annals of Computer Science and Information Systems (ACSIS), vol. 18, 2019, New York, Institute of Electrical and Electronics Engineers, pp. 539–545, ISBN 978-83-955416-0-5, doi: 10.15439/2019F71.

Typ: publikacja konferencyjna (artykuł w materiałach konferencyjnych). **Rok**: 2019. MNiSW: 70 pkt. **Cytowania**: Web of Science – 3; Scopus − 6; Google Scholar − 10.

Wkład autora rozprawy: 75%. **Rola**: autor korespondencyjny.

Upowszechnienie: referat konferencyjny, 14th FedCSIS 2019, 1–4.09.2019, Lipsk, Niemcy.

Zakres wkładu: współtworzenie koncepcji integracji aplikacji i usług w środowisku World Data System i World Data Centers (WDS/WDC), w tym doprecyzowanie problemu oraz założeń metody; opracowanie i opis architektury rozwiązania obejmującej podział na warstwę usług oraz warstwę orkiestracji agentowej; rozwinięcie formalizmu logicznego stanowiącego podstawę wnioskowania i dedukcyjnej konstrukcji planu; przygotowanie mechanizmu rekonstrukcji struktury rozwiązania oraz zasad wykonania planu, w tym obsługi rozgałęzień i fragmentów możliwych do uruchomienia równoległego; przygotowanie zasadniczej części manuskryptu w obszarze formalizacji, algorytmów i architektury; integracja uwag współautorów; pełnienie roli autora korespondencyjnego i koordynacja procesu publikacyjnego; reprezentowanie zespołu oraz prezentacja wyników w formie referatu na konferencji FedCSIS 2019.

[5] **Nowakowski G.**, Telenyk S., Yefremov K., Khmeliuk V., *Simple and flexible way to integrate heterogeneous information systems and their services into the world data system*.

Publikacja: Journal of Automation, Mobile Robotics and Intelligent Systems (JAMRIS), vol. 15, 2021, no. 4, pp. 76-90, doi: 10.14313/JAMRIS/4-2021/29.

Typ: publikacja w czasopiśmie. **Rok**: 2022. **MNiSW**: 70 pkt.

Cytowania: Scopus − 1; Google Scholar − 1.

Wkład autora rozprawy: 75%. **Rola**: autor korespondencyjny.

Zakres wkładu: wiodący udział w sformułowaniu problemu i założeń integracji heterogenicznych systemów i usług w kontekście WDS/WDC oraz opracowanie architektury rozwiązania; opracowanie opisu pełnej ścieżki od celu użytkownika do wykonania, tj. przejścia: cel użytkownika → wnioskowanie → plan wykonania → wykonanie; rozwinięcie formalizacji logicznej oraz kluczowych procedur algorytmicznych prowadzących do konstrukcji planu wykonania; przygotowanie przykładu demonstracyjnego pokazującego rekonstrukcję struktury rozwiązania jako kompozycji usług oraz uruchomienie tak wyznaczonego planu zgodnie z zależnościami danych;

przygotowanie i redakcja zasadniczej części manuskryptu; integracja uwag współautorów; koordynacja procesu publikacyjnego jako autor korespondencyjny.

[6] Telenyk S., **Nowakowski G.**, Zharikov E., Vovk J., *Conceptual foundations of the use of formal models and methods for the rapid creation of web applications*.
Publikacja: IDAACS 2019: proceedings of the 2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), vol. 2, 2019, Institute of Electrical and Electronics Engineers, IEEE, pp. 512-518, ISBN 978-1-7281-4069-8, doi: 10.1109/IDAACS.2019.8924416.
Typ: publikacja konferencyjna (artykuł w materiałach konferencyjnych). **Rok**: 2019. **MNiSW**: 20 pkt. **Cytowania**: Web of Science − 1; Scopus − 3; Google Scholar − 9.
Wkład autora rozprawy: 55%. **Rola**: autor korespondencyjny.
Upowszechnienie: referat konferencyjny, IDAACS 2019, 18–21.09.2019, Metz, Francja.
Zakres wkładu: współtworzenie koncepcji artykułu i sposobu ujęcia problemu, w tym sformułowanie problemu badawczego, określenie zakresu oraz uzasadnienie przyjętych założeń; opracowanie formalnego ujęcia przejścia od specyfikacji i wymagań do struktury aplikacji oraz schematu wykonania i współdziałania modułów, stanowiącego rdzeń zależności: formalizacja → konstrukcja rozwiązania; uzgodnienie i opis metodycznych zasad doboru oraz łączenia komponentów aplikacji webowej, w tym wariantu szablonowego; przygotowanie zasadniczej części treści artykułu; redakcja i integracja całości oraz dopracowanie narracji naukowej; pełnienie roli autora korespondencyjnego i koordynacja procesu publikacyjnego; reprezentowanie zespołu oraz prezentacja wyników w formie referatu na konferencji IDAACS 2019.

[7] Telenyk S., **Nowakowski G.**, Gavrilenko O., Miahkyi M., Khalus O., *An analysis of the influence of famous people's posts on social networks on the cryptocurrency exchange rate*.
Publikacja: Bulletin of the Polish Academy of Sciences: Technical Sciences, vol. 72, no. 4, 2024, pp. 1-9, doi: 10.24425/bpasts.2024.150117.
Typ: publikacja w czasopiśmie. **Rok**: 2024. **MNiSW**: 100 pkt.
Wskaźniki bibliometryczne: Scopus CiteScore 2024 – 2.7; JCR Impact Factor 2024 – 1.1.
Cytowania: Web of Science – 1; Scopus – 1; Google Scholar − 4.
Wkład autora rozprawy: 55%. **Rola**: autor korespondencyjny.
Zakres wkładu: wiodący udział w sformułowaniu problemu badawczego oraz metodyki walidacji statystycznej wpływu sygnału zewnętrznego (publikacji w mediach

społecznościowych) na kurs kryptowaluty; zaprojektowanie porównania metod predykcyjnych (z użyciem ATAPSN *(ang. algorithm for forecasting the cryptocurrency exchange rate taking into account posts on social networks)* oraz metod klasycznych: ARIMA *(ang. autoregressive integrated moving average)* i wygładzania wykładniczego *(ang. exponential smoothing)*) oraz przygotowanie procedury oceny obejmującej miary błędu prognozy, korelacje Spearmana i Pearsona oraz test t – Studenta; współkształtowanie narzędzia obliczeniowego do pozyskiwania danych i uruchamiania obliczeń; opracowanie zasadniczej części manuskryptu; integracja uwag współautorów; koordynacja procesu publikacyjnego jako autor korespondencyjny.

[8] Shymkovych V., **Nowakowski G.**, Telenyk S., *Joint Punctuation Restoration and Text Capitalisation with a Hybrid XLM-RoBERTa–LSTM Model*.

Publikacja: IDAACS 2025: proceedings of the 13th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS), Vol. 1, 2025, Institute of Electrical and Electronics Engineers, IEEE, pp. 990-997, ISBN 979-8-3315-8045-2, doi: 10.1109/IDAACS68557.2025.11322226.

Typ: publikacja konferencyjna (artykuł w materiałach konferencyjnych). **Rok:** 2025. **MNiSW:** 20 pkt.

Wkład autora rozprawy: 55%. **Rola:** autor korespondencyjny.

Upowszechnienie: referat konferencyjny, IDAACS 2025, 04–06.09.2025, Gliwice, Polska.

Zakres wkładu: wiodący udział w sformułowaniu problemu rekonstrukcji brakującej interpunkcji i kapitalizacji jako nośnika nieprecyzyjności w danych tekstowych oraz określenie celu i zakresu eksperymentów; formalizacja zadania i metodyki oceny, w tym ujęcie problemu jako klasyfikacji tokenów w przestrzeni klas łączących wariant interpunkcji z informacją o kapitalizacji oraz zastosowanie masek (tokeny końca słowa oraz tokeny wypełniające sekwencję, padding); opracowanie sposobu raportowania wyników z uwzględnieniem niezbalansowania klas, obejmującego metryki: accuracy, precision, recall i F1, a także macierz pomyłek oraz tabelaryczne porównania wyników z literaturą; współprojektowanie architektury hybrydowej (enkoder XLM – RoBERTa, warstwa BiLSTM, klasyfikator liniowy) oraz opracowanie kluczowych elementów opisu metody i interpretacji wyników; przygotowanie zasadniczej części manuskryptu; integracja uwag współautorów; pełnienie roli autora korespondencyjnego i koordynacja

procesu publikacyjnego; reprezentowanie zespołu oraz prezentacja wyników w formie referatu na konferencji IDAACS 2025.

[9] Gordienko Y., **Nowakowski G.**, Kochura Y., Taran V., Stirenko S., *Generative Data Augmentation by Dataset Distillation*.

Publikacja: Artificial Intelligence for Biomedical Data: First International Workshop, AIBio 2025, Held in Conjunction with the European Conference on Artificial Intelligence, ECAI 2025, Communications in Computer and Information Science, vol. 2696, 2026, Cham, Springer, pp. 105-118, ISBN 978-3-032-17215-0. DOI: 10.1007/978-3-032-17216-7_9.

Typ: publikacja konferencyjna (artykuł w materiałach konferencyjnych). **Rok:** 2026. **MNiSW:** 140 pkt.

Wkład autora rozprawy: 40%.

Upowszechnienie: referat konferencyjny, First International Workshop, AIBio 2025, Held in Conjunction with the European Conference on Artificial Intelligence, ECAI 2025, 25–26.10.2025, Bolonia, Włochy.

Zakres wkładu: wiodący udział w zdefiniowaniu problemu oraz w sformułowaniu formalizacji optymalizacyjnej dla generatywnej augmentacji przez destylację zbioru danych w podejściu GDADD (*ang. generative data augmentation by dataset distillation*), w którym formalizacja przyjmuje postać jednoznacznie określonej funkcji celu sterującej konstrukcją danych syntetycznych; wiodący udział w ujęciu celu jako dopasowania trajektorii uczenia DDMTT (*ang. dataset distillation by matching training trajectories*) z wykorzystaniem trajektorii eksperckich jako odniesienia dla procesu destylacji; współopracowanie metodyki badań i protokołu ewaluacji, w tym dobór zbiorów danych i scenariuszy porównawczych, konfiguracja parametrów oraz sposób raportowania wyników; wiodący udział w interpretacji rezultatów wraz z omówieniem ograniczeń systemowych, takich jak koszt obliczeniowy destylacji, zależność od jakości trajektorii uczenia, wrażliwość na niezbalansowanie klas oraz potencjalne obciążenia danych; przygotowanie znaczącej części manuskryptu; integracja wkładów współautorów; redakcja wersji finalnej; reprezentowanie zespołu oraz prezentacja wyników w formie referatu na konferencji ECAI 2025.

2. Wstęp

2.1. Motywacja i kontekst pracy

Współczesne systemy informatyczne coraz rzadziej operują na danych, których interpretacja jest jednoznaczna i kompletna. W praktyce dominują informacje nieostre (np. wyrażane językiem naturalnym i obarczone nieostrością pojęć), niepełne (braki danych, luki pomiarowe), zaszumione (błędy rejestracji, degradacja sygnału) oraz kontekstowe (zależne od użytkownika, celu i sytuacji). Skuteczne przetwarzanie takiej informacji wymaga podejścia łączącego formalizację − tak, aby własności istotne z punktu widzenia interpretacji i przetwarzania (nieostrość, niepełność, zaszumienie, kontekstowość) były jawnie modelowane i kontrolowalne − oraz algorytmizację − tak, aby przetwarzanie było wykonalne i efektywne obliczeniowo w rzeczywistych systemach. W niniejszej rozprawie termin *informacja nieprecyzyjna* rozumiany jest szeroko jako informacja, której interpretacji nie da się w pełni opisać za pomocą predykatów dwuwartościowych bez utraty istotnych aspektów znaczenia. Pojęcie to obejmuje cztery podstawowe postacie: (I) nieostrość pojęć *(ang. fuzziness)* typową dla terminów lingwistycznych, (II) niepełność *(ang. incompleteness)* opisu danych lub procesu, (III) zaszumienie *(ang. noise)* wynikające z błędów rejestracji i degradacji sygnału oraz (IV) kontekstowość *(ang. context dependence)* znaczenia zależnego od użytkownika i celu. Przyjęta perspektywa zakłada, że formalny opis semantyki oraz algorytmiczna realizacja nie są konkurencyjnymi podejściami, lecz komplementarnymi warunkami tworzenia rozwiązań, które są jednocześnie interpretowalne i wykonalne obliczeniowo. Ten dualizm stanowi oś tematyczną rozprawy „Formalne i algorytmiczne metody przetwarzania informacji nieprecyzyjnej w systemach informatycznych”.

Naturalnym punktem wyjścia do rozważań o formalnym przetwarzaniu informacji nieprecyzyjnej są systemy bazodanowe, ponieważ formalizm (model danych, algebra relacyjna, logika predykatów) jest w nich szczególnie silnie ugruntowany [10]–[13]. Klasyczny model relacyjny Codda stanowi fundament projektowania i teorii relacyjnych baz danych [10]. Równolegle standard SQL − jako dominujący język zapytań [13], [14] − został sformalizowany i jest rozwijany w postaci norm ISO/IEC, w tym w aktualnej edycji standardu (SQL:2023 / ISO/IEC 9075:2023) [15]. Zarówno model relacyjny, jak i standardowy SQL zostały zaprojektowane przede wszystkim z myślą o przetwarzaniu informacji ostrych *(ang. crisp)*. Warunki selekcji są oceniane logicznie (w klasycznym ujęciu: prawda/fałsz; w SQL dodatkowo występuje wartość UNKNOWN wynikająca z obecności NULL), natomiast semantyka zapytania SQL nie przewiduje oceny stopniowej [13], [15]. W konsekwencji klasyczne systemy

bazodanowe nie umożliwiają bezpośredniego wyrażania preferencji użytkownika w postaci wymagań stopniowych, co w literaturze ujmowane jest jako problem sztywności (*ang. rigidity*) w kontekście elastycznego wyszukiwania [16]–[21]. Wymagania użytkowników są często formułowane za pomocą predykatów lingwistycznych (np. „wysoki", „podobny", „blisko", „w przybliżeniu", „raczej"), których znaczenie ma charakter nieostry. Sprowadzenie ich do predykatów dwuwartościowych wymaga zatem przyjęcia ostrych progów decyzyjnych, czego skutkiem może być utrata intencji zapytania [18], [22]–[24]. Zastosowanie takich progów powoduje ponadto efekty brzegowe polegające na tym, że krotki (*ang. tuples*) o wartościach nieznacznie przekraczających ustaloną granicę są odrzucane w takim samym stopniu jak krotki wyraźnie niespełniające kryterium, mimo że z perspektywy użytkownika mogą nadal odpowiadać poszukiwanym cechom [1], [2], [25]. W rezultacie sztywność klasycznych zapytań skutkuje brakiem mechanizmu różnicowania i porządkowania odpowiedzi zgodnie z preferencjami użytkownika oraz nadwrażliwością na dobór progów, która może eliminować krotki leżące blisko granicy kryterium, a w skrajnym przypadku prowadzić do wyniku pustego i wymuszać iteracyjne dostrajanie parametrów zapytania [1], [25]. Z tego powodu w literaturze sformułowano formalne modele reprezentacji nieostrości, m.in. oparte na teorii zbiorów rozmytych [26]. Ujęcie to umożliwia przejście od logiki binarnej do opisu stopniowego poprzez funkcje przynależności i stanowi podstawę logiki rozmytej, w której terminy lingwistyczne oraz operatory logiczne interpretuje się w sposób ciągły [26]–[29]. Kolejnym krokiem było pogłębienie i uporządkowanie zagadnień semantyki oraz wnioskowania w warunkach niepewności [16], [20]. Na gruncie baz danych doprowadziło to do powstania języków zapytań rozmytych (*ang. fuzzy query*), w których warunki selekcji i/lub dopasowania są oceniane stopniem spełnienia [25], [30] (np. z użyciem funkcji przynależności), a wyniki mogą być porządkowane według stopnia spełnienia warunków zapytania w ujęciu rozmytym [18], [19], [22]–[24], [31]. Logika rozmyta była wielokrotnie wskazywana jako narzędzie zwiększające elastyczność języków zapytań do baz danych, co znajduje odzwierciedlenie w pracach m.in. [20], [22], [24], [32]–[40].

Wśród klasycznych propozycji wskazuje się m.in. relacyjny język SQLf dla zapytań rozmytych [41], który stanowi przykład formalnego rozszerzenia semantyki zapytań przy zachowaniu relacyjnego rdzenia. W takim ujęciu dopasowanie krotki do warunku nie jest już traktowane dychotomicznie, lecz stopniowo, a wynik zapytania można interpretować jako relację wzbogaconą o atrybut opisujący stopień spełnienia w przedziale [0,1]. Umożliwia to porządkowanie (rankingowanie) odpowiedzi zgodnie z preferencjami użytkownika.

Istotnym wyzwaniem pozostaje przełożenie formalnego opisu nieostrości na rozwiązanie możliwe do praktycznej realizacji. Dotyczy to pytania, jak realizować zapytania rozmyte w systemach, których rdzeń opiera się na standardowym SQL oraz klasycznych mechanizmach wykonania i optymalizacji zapytań relacyjnych, bez modyfikacji silnika DBMS. Z tego względu w niniejszej rozprawie wyodrębniono obszar badawczy (publikacje [1], [2]) dotyczący metod transformacji zapytań rozmytych rozumianych jako przekształcenia pozwalające zachować formalną semantykę zapytania nieostrego, a jednocześnie umożliwiające jego wykonanie w istniejących systemach relacyjnych bez kosztownej przebudowy DBMS. Uzasadnieniem takiego podejścia jest m.in. fakt, że w literaturze podkreśla się ostrożność wobec zapytań rozmytych ze względu na koszt ich przetwarzania. W przeciwieństwie do selekcji dwuwartościowej wykonanie zapytania rozmytego wymaga obliczenia stopnia spełnienia warunków dla każdej krotki oraz uporządkowania wyników według tej wartości. Powoduje to dodatkowy narzut obliczeniowy i etap porządkowania, utrudniając tym samym optymalizację zapytania i może znacząco zwiększać koszt jego wykonania w porównaniu z selekcją klasyczną [18], [21].

W tym kontekście warto zaznaczyć, że problematyka przetwarzania zapytań rozmytych pozostaje aktualnym obszarem badań rozwijanym w kilku uzupełniających się kierunkach. Obejmują one m.in. (I) przeglądy systematyzujące, które porządkują podejścia zarówno do modelowania danych rozmytych, jak i do formułowania zapytań rozmytych nad danymi ostrymi w różnych modelach danych, wskazując jednocześnie trendy oraz otwarte problemy badawcze [42]; (II) nurt związany z traktowaniem zapytań rozmytych jako elementu systemów wyszukiwania, w których rekordy są oceniane i porządkowane według stopnia spełnienia warunków; stopień ten jest wyznaczany na podstawie zdefiniowanych predykatów i obsługiwany przez widoki, zamiast być każdorazowo liczony podczas wykonania zapytania; upraszcza to użycie konstrukcji rozmytych i może skracać czas odpowiedzi [43]; (III) rozwój formalnych modeli relacyjnych integrujących różne postacie niedoskonałości informacji, takie jak rozmytość wartości atrybutów oraz niepewność na poziomie krotek, tj. przypisanie rekordom prawdopodobieństwa (stopnia pewności) ich istnienia lub poprawności, co wzmacnia podstawy teoretyczne przetwarzania zapytań rozmytych w paradygmacie relacyjnym [44] oraz (IV) mechanizmy wspierające zapytania rozmyte w środowiskach danych semi-ustrukturyzowanych (np. JSON/NoSQL) obejmujące formalizację oceny stopnia spełnienia warunków i agregacji rozmytych w językach zapytań, a także meta-model mechanizmów oceny oraz konstrukcje wspierające ich definicję i użycie [45]. Na tym tle metody transformacji zapytań rozmytych do standardowego SQL przedstawione w publikacjach [1], [2] można

traktować jako podejście wykonawcze realizowane w warstwie bazy danych w oparciu o standardowy SQL zachowujące formalną semantykę rozmytości przy jednoczesnym wykorzystaniu istniejących mechanizmów wykonania i optymalizacji zapytań.

Z perspektywy implementacyjnej można wyróżnić dwa komplementarne podejścia. Pierwsze polega na realizacji zapytań rozmytych po stronie bazy danych. Do wyznaczania stopni spełnienia warunków oraz porządkowania wyników według stopnia dopasowania wykorzystuje się standardowe mechanizmy DBMS, takie jak funkcje i procedury. Drugie podejście przenosi część logiki na poziom aplikacji komunikującej się z bazą, co wymaga algorytmicznego przetwarzania warunków zapytania oraz kontroli kosztu obliczeniowego związanego z wyznaczaniem stopni spełnienia i porządkowaniem odpowiedzi. W publikacjach [1], [2] rozwinięto metodyki i algorytmy transformacji zapytań rozmytych do postaci wykonywalnej w standardowym SQL, przy zachowaniu semantyki stopniowej oraz weryfikacji wykonalności i kosztu przetwarzania.

Inny kontekst badawczy rozprawy dotyczy integracji aplikacji i źródeł danych w systemach informatycznych, zwłaszcza w warunkach niejednorodności technologicznej i semantycznej [46]–[49]. W środowiskach badawczych, a także w wielu zastosowaniach przemysłowych, typowym scenariuszem jest konstruowanie procesu przetwarzania prowadzącego do uzyskania określonego rezultatu, obejmującego liczne narzędzia, formaty danych i usługi, przy czym ich dobór oraz sposób połączenia nie są zadane wprost na wejściu. W takiej sytuacji nieprecyzyjność ujawnia się nie tylko na poziomie danych, lecz także na poziomie opisu procesu, m.in. w zakresie doboru usług i ich kolejności. Użytkownik określa cel obliczeniowy w języku dziedziny, opisując oczekiwany rezultat, natomiast przepływ pracy (*ang. workflow*), rozumiany jako konkretna sekwencja kroków przetwarzania, nie jest podany i musi zostać skonstruowany przez system w postaci planu wykonania [3]–[5]. Problem ten jest szeroko omawiany w literaturze dotyczącej kompozycji przepływów pracy, w której podkreśla się, że w wielu dziedzinach nadal dominuje ręczne łączenie heterogenicznych narzędzi, a zwiększenie poziomu automatyzacji wymaga jawnych opisów semantycznych i mechanizmów wnioskowania [46]–[50].

Ten typ nieprecyzyjności ma charakter niepełności specyfikacji. System powinien uzupełnić brakujące elementy procesu poprzez wyznaczenie wykonalnego planu przy zachowaniu poprawności. Z tego względu uzasadnione staje się zastosowanie metod formalnych, które pozwalają przejść od specyfikacji celu użytkownika do algorytmicznie skonstruowanego planu wykonania. W ujęciu logiczno-formalnym zaproponowanym w [3]–[5] usługi są opisywane jawnie za pomocą warunków wstępnych i końcowych (*ang. preconditions / postconditions*)

oraz aksjomatów wiążących metody z danymi. Następnie system realizuje wnioskowanie prowadzące do konstrukcji dowodu, z którego rekonstruowany jest plan wykonania w postaci drzewa wykonania (*ang. execution tree*) [3], [4]. W praktycznej realizacji plan wykonania ma postać acyklicznego grafu skierowanego (DAG, *ang. directed acyclic graph*), w którym obok kroków sekwencyjnych mogą występować operacje rozdzielania i scalania danych, co pozwala identyfikować fragmenty możliwe do wykonania równolegle. Uwzględnia się również kontrolę złożoności obliczeniowej, aby ograniczyć przyrost liczby możliwych rozwiązań. W tym celu wnioskowanie jest prowadzone z wykorzystaniem abstrakcji typów (*ang. typification abstraction*), dzięki czemu możliwe jest najpierw uzyskanie struktury rozwiązania w przestrzeni bardziej abstrakcyjnej, a następnie wykorzystanie jej do sterowania wyszukiwaniem rozwiązania szczegółowego. Takie prowadzenie dowodzenia pozwala już na wczesnym etapie eliminować nieperspektywiczne gałęzie i ograniczać koszt przeszukiwania [4], [5].

Warto zauważyć, że we współczesnych podejściach do integracji na poziomie infrastruktury procesy obliczeniowe są coraz częściej traktowane jako artefakty, które powinny być nie tylko wykonalne, lecz także współdziałające i możliwe do ponownego użycia. Taki sposób myślenia jest spójny m.in. z koncepcją FAIR Hourglass, która porządkuje wdrażanie zasad FAIR *(ang. Findable, Accessible, Interoperable, Reusable)* poprzez rozróżnienie etapów FAIRification oraz FAIR Orchestration [51], [52]. W tym kontekście korespondują również publikacje postulujące budowę ekosystemu usług wspierających przepływy pracy zgodne z FAIR i traktowane jako pełnoprawne obiekty badawcze [53]–[55]. Ponadto w literaturze podkreśla się rosnącą rolę rozwiązań agentowych w zarządzaniu i ewolucji przepływów pracy [56]–[58]. W rezultacie akcentuje się znaczenie mechanizmów planowania, koordynacji i kontroli, które pozwalają utrzymać spójność i wykonalność procesów w heterogenicznych środowiskach narzędzi i usług.

Wkład obszaru badawczego II obejmującego publikacje [3]–[5] polega na formalizacji opisu usług za pomocą warunków wstępnych i końcowych oraz aksjomatów, na dedukcyjnej konstrukcji dowodu realizowalności celu użytkownika, a następnie na rekonstrukcji planu wykonania. Ujęcie to obejmuje także mechanizm ograniczania złożoności, w którym wnioskowanie prowadzi się najpierw na poziomie uogólnionym, z wykorzystaniem abstrakcji typów, a uzyskany wynik służy do ukierunkowania wyszukiwania rozwiązania szczegółowego. W efekcie nieprecyzyjność w systemie zintegrowanym dotyczy nie tylko treści danych, lecz także opisu procesu i reguł kojarzenia jego elementów. Zastosowanie metod formalnych

pozwala tę niejednoznaczność jawnie reprezentować oraz przetwarzać w sposób kontrolowany i obliczeniowo wykonalny.

Kolejny ważny kontekst rozprawy dotyczy nieprecyzyjności ujawniającej się na etapie projektowania systemów informatycznych, zwłaszcza gdy wymagania są formułowane w sposób niepełny, ewoluujący lub wieloznaczny, a czas wytworzenia rozwiązania ma kluczowe znaczenie. W takim ujęciu informacja nieprecyzyjna nie dotyczy wyłącznie danych wejściowych, lecz także opisu projektowanego systemu, obejmującego model dziedziny, reguły biznesowe oraz zależności między elementami. Opis ten musi zostać sformalizowany na tyle, aby umożliwić algorytmiczne wyprowadzenie spójnej struktury aplikacji, w tym szkieletu aplikacji webowej złożonej z komponentów o jawnie określonych wejściach i wyjściach. Wyniki badań w obszarze badawczym III przedstawiono w publikacji [6].

W [6] skoncentrowano się na skróceniu procesu przejścia od niepełnej specyfikacji użytkownika do spójnej konstrukcji rozwiązania poprzez przyjęcie standardowej architektury aplikacji webowych oraz formalny dobór jej komponentów zgodnie z wymaganiami. Projektowanie jest w tym ujęciu traktowane jako proces formalny, ponieważ skonstruowanie rozwiązania odpowiada wyprowadzeniu opisu określającego sposób działania i współdziałania modułów systemu. Jest to zatem przejście od niepełnej specyfikacji do jednoznacznie zdefiniowanej struktury aplikacji, która może stanowić podstawę implementacji [6]. Formalny opis wspiera przy tym automatyczne zestawianie warstw aplikacji oraz wykorzystanie gotowych szablonów i wzorców projektowych typowych dla aplikacji webowych [6]. Na poziomie inżynierii oprogramowania podejście to wpisuje się w szerszy nurt modelowania i automatyzacji wytwarzania aplikacji webowych [59]–[62]. W publikacji [6] przyjęto założenie, że formalizacja nie pełni wyłącznie funkcji opisowej, lecz ma charakter operacyjny, ponieważ umożliwia dobór i integrację komponentów systemu w spójną całość, a tym samym stanowi podstawę algorytmicznej syntezy rozwiązania. W takim ujęciu nieprecyzyjność wymagań nie jest eliminowana heurystycznie, lecz redukowana poprzez kontrolowany mechanizm formalny, który wspiera przejście od specyfikacji do artefaktów projektowych i implementacyjnych.

Perspektywa przedstawiona w [6] znajduje potwierdzenie w nowszych badaniach, w których zwraca się uwagę na systemowy charakter nieprecyzyjności specyfikacji. W obszarze inżynierii wymagań akcentuje się to, że metody formalne realizowane w postaci narzędzi wspierających analizę pozwalają wcześniej wykrywać typowe defekty specyfikacji, takie jak niespójności, sprzeczności czy niedookreślenia i dzięki temu ograniczać ryzyko kosztownych poprawek na późnych etapach wytwarzania [63]–[65]. Systematyczne przeglądy w obszarze inżynierii

aplikacji webowych pokazują ponadto, że automatyzacja projektowania oraz generowanie artefaktów implementacyjnych pozostają aktywnym kierunkiem badań, a jako najczęściej wskazywane korzyści raportuje się wzrost produktywności i usprawnienie procesu wytwórczego [66]–[68]. Tendencja ta obejmuje także współczesne, złożone domeny, w tym środowiska chmurowe i Internet Rzeczy, w których projektowanie jest ujmowane jako proces iteracyjnego identyfikowania i doprecyzowywania wymagań, a część informacji potrzebnej do poprawnej konstrukcji rozwiązania musi zostać uzupełniona na etapie konfiguracji, ponieważ nie wynika bezpośrednio z samego modelu [69]–[71]. W tym sensie wkład [6] traktujący formalizację jako mechanizm operacyjny wspierający przejście od niepełnej specyfikacji do jednoznacznej struktury rozwiązania, pozostaje spójny z aktualnymi kierunkami badań.

W szerszym kontekście współczesnej informatyki coraz większe znaczenie ma przetwarzanie informacji pochodzącej ze źródeł nieustrukturyzowanych oraz sygnałów pośrednich, które są trudne do jednoznacznej interpretacji, ale mogą mieć wysoką wartość decyzyjną. Informacje dostępne w Internecie są często zaszumione, a ich użyteczność zależy od zdolności systemu do szybkiej selekcji i analizy treści istotnych dla danego problemu. Jak pokazano w [7] na przykładzie analizy wpływu publikacji w mediach społecznościowych na zjawiska ekonomiczne, istotne jest szybkie uwzględnienie tego sygnału informacyjnego, ponieważ oddziaływanie informacji zewnętrznej bywa krótkotrwałe. Jednocześnie zachodzi ono równolegle z wieloma innymi bodźcami, dlatego obserwowany efekt zależy od aktualnej sytuacji rynkowej i informacyjnej, utrudniając tym samym jednoznaczną interpretację [7]. W tym ujęciu nieprecyzyjność ma postać sygnału kontekstowego, którego znaczenie wymaga formalizacji umożliwiającej jego algorytmiczne wykorzystanie, a następnie empirycznej oceny wpływu na wynik prognozy [7].

Perspektywa ta jest spójna ze współczesnymi kierunkami badań, tym samym [7] wpisuje się w ten nurt, ujmując sygnały z mediów społecznościowych jako krótkookresowe impulsy informacyjne, którym mogą towarzyszyć wzrosty zmienności i inne reakcje rynku. W centrum uwagi pozostają m.in. wyznaczenie horyzontu czasowego oddziaływania, odporność wniosków na szum oraz uwzględnianie czynników zakłócających. Literatura ostatnich lat wskazuje na dwa uzupełniające się kierunki badań. Pierwszy obejmuje analizę wpływu aktywności pojedynczych, silnie oddziałujących kont na rynki kryptowalut [72]–[74]. Drugi dotyczy rozwoju systemów wspomagania decyzji i prognozowania, które łączą sygnały z mediów społecznościowych i serwisów informacyjnych [75]–[77]. Równolegle podejmowane są także prace przeglądowe porządkujące metody integracji danych finansowych, on-chain oraz tekstowych w zadaniach predykcyjnych [78]–[80].

W przetwarzaniu języka naturalnego (NLP, *ang. natural language processing*) nieprecyzyjność w wielu przypadkach wynika z niepełnej struktury tekstu. Dotyczy to zwłaszcza transkrypcji mowy oraz treści tworzonych przez użytkowników, w których często brakuje interpunkcji i wielkich liter. W efekcie tekst jest trudniejszy zarówno do czytania, jak i do dalszego automatycznego przetwarzania. W takim ujęciu celem jest odtworzenie brakujących elementów struktury, czyli automatyczne przywrócenie interpunkcji i kapitalizacji [81]. Zadanie można formalnie opisać jako etykietowanie sekwencji, w którym dla każdego tokenu model przewiduje etykietę określającą, czy po danym tokenie powinien wystąpić znak interpunkcyjny oraz czy token powinien rozpoczynać się wielką literą. W procesie uczenia i oceny należy pominąć tokeny techniczne, takie jak wypełnienie sekwencji (*ang. padding*) oraz znaczniki specjalne, aby modele były trenowane i oceniane wyłącznie na fragmentach niosących treść.

Oba przykłady pokazują, że nieprecyzyjność nie ogranicza się do klasycznej nieostrości rozmytej. Może ona wynikać z kontekstowości sygnałów zewnętrznych i krótkotrwałości ich wpływu, a także z braków strukturalnych w danych językowych, takich jak brak interpunkcji i kapitalizacji. Z perspektywy niniejszej rozprawy istotne jest, że również w takich zadaniach realizuje się ten sam schemat badawczy obejmujący formalizację problemu, ujęcie algorytmiczne oraz empiryczną weryfikację skuteczności proponowanego rozwiązania [7], [8].

Wątek ten jest konsekwentnie rozwijany w literaturze ostatnich lat, w tym w pracach powstałych w zbliżonym okresie do [8], w których rekonstrukcję interpunkcji i kapitalizacji traktuje się jako etap normalizacji i ustrukturyzowania tekstu przed dalszym przetwarzaniem [82]–[87]. Rekonstrukcja ta bywa ujmowana jako moduł postprocessingu w łańcuchach rozpoznawania mowy [86], [87] oraz jako etap przygotowania tekstu pochodzącego z domen nieformalnych, takich jak wpisy użytkowników [84], [85], gdzie brak tych elementów strukturalnych jest zjawiskiem częstym. Prace te wskazują na dominację podejść opartych na transformerach [82]–[87], jednocześnie akcentując aspekty istotne z punktu widzenia wdrożeń. Po pierwsze, w scenariuszach strumieniowych kluczowe są wymagania wydajnościowe i niskie opóźnienie, co prowadzi do poszukiwania rozwiązań o niskim koszcie obliczeniowym i działających online [86], [87]. Po drugie, ocena jakości powinna uwzględniać nierównowagę klas oraz trudność predykcji rzadkich znaków interpunkcyjnych, ponieważ te czynniki silnie wpływają na interpretację wyników w praktyce [82]–[87].

Współczesne systemy oparte na uczeniu maszynowym wprowadzają dodatkowy wymiar nieprecyzyjności. Informacja o zjawisku jest pozyskiwana pośrednio, na podstawie danych empirycznych, a zbiór treningowy stanowi jedynie przybliżenie tego zjawiska i może

odzwierciedlać szum, ograniczoną liczbę danych, nierównowagę klas oraz obciążenia wynikające z procesu pozyskiwania i uczenia. Jednocześnie rosnące koszty obliczeniowe i pamięciowe uczenia głębokiego sprzyjają metodom, które pozwalają kompresować informację zawartą w danych, przy zachowaniu jakości generalizacji. W publikacji [9] rozpatrzono w tym kontekście podejście GDADD oparte na destylacji zbioru danych i generowaniu danych syntetycznych, w którym formalizacja ma charakter optymalizacyjny. Zbiór syntetyczny wyznacza się tak, aby uczenie modelu na danych syntetycznych prowadziło do przebiegu możliwie zbliżonego do uczenia na danych rzeczywistych, rozumianego jako sekwencja kolejnych stanów modelu (trajektoria w przestrzeni parametrów). Cel ten jest formalizowany przez kryterium dopasowania trajektorii uczenia DDMTT. W tym ujęciu wiedza nie jest zapisana w postaci reguł lub semantyki logicznej, lecz jest utrwalona pośrednio w danych syntetycznych, a proces jej „kompresji" jest sterowany jednoznacznie zdefiniowaną funkcją celu. Podejście to jest spójne z tezą niniejszej rozprawy, ponieważ również tutaj realizowany jest schemat badawczy: formalizacja → algorytm → weryfikacja empiryczna. Formalizacja przyjmuje postać celu optymalizacji, na jej podstawie konstruowany jest algorytm destylacji, a skuteczność rozwiązania jest oceniana eksperymentalnie na różnych zbiorach danych, z uwzględnieniem kosztów obliczeniowych i ograniczeń systemowych.

Zagadnienie destylacji określane również jako kondensacja zbiorów danych, stanowi obecnie aktywny i szybko rozwijany kierunek badań [88]–[92]. Kluczowe znaczenie ma w nim precyzyjne zdefiniowanie celu kompresji informacji oraz zaprojektowanie stabilnej procedury optymalizacji, która pozwala ograniczyć koszt uczenia przy zachowaniu jakości generalizacji na nowych danych [92]. W literaturze rozwijane są podejścia, w których zbiór danych syntetycznych jest konstruowany poprzez bezpośrednią optymalizację tak, aby możliwie szybko przybliżać efekt treningu na danych rzeczywistych [88], jak i metody wzmacniające stabilność procesu przez dopasowanie przebiegów uczenia rozumianych jako trajektorie treningu [93], lub przez włączenie różniczkowalnych mechanizmów augmentacji do pętli optymalizacyjnej [94]. Równolegle powstają przeglądy porządkujące warianty metod, ich założenia oraz ograniczenia, takie jak stabilność optymalizacji, zależność od architektury i możliwość transferu między zadaniami [92]. W tym kontekście publikacja [9] odnosi się do ukształtowanego stanu badań i rozwija podejście oparte na destylacji danych oraz generowaniu zbioru syntetycznego.

Podsumowując, przedstawiony cykl publikacji można ująć w cztery uzupełniające się obszary badawcze. Obejmują one: (I) formalizację nieostrości w danych i zapytaniach oraz transformację zapytań nieprecyzyjnych do postaci wykonywalnej w SQL, (II) formalno-

algorytmiczną konstrukcję planów integracji usług przy niepełnej specyfikacji celu, (III) formalizację nieprecyzyjnej specyfikacji i wymagań na etapie projektowania w modelach wspierających szybkie tworzenie aplikacji oraz (IV) rozszerzenia obejmujące inne nośniki nieprecyzyjności, takie jak informacja zewnętrzna, niepełność struktury danych językowych oraz pośrednia reprezentacja empiryczna w danych syntetycznych. Obszar IV pełni rolę weryfikacyjną, pokazując, że przyjęty w rozprawie schemat badawczy: formalizacja → algorytm → weryfikacja pozostaje zasadny również poza klasycznymi zapytaniami rozmytymi i przepływami pracy.

Motywację do napisania niniejszej rozprawy stanowił fakt, że w wielu typach systemów informatycznych, od relacyjnych baz danych, przez platformy integracyjne i środowiska usługowe, po analitykę informacji zewnętrznej i systemy oparte na uczeniu maszynowym, informacja nieprecyzyjna jest regułą, a nie wyjątkiem. Kluczowym wyzwaniem nie jest samo dopuszczenie nieprecyzyjności, lecz zaprojektowanie metod, które formalizują jej znaczenie, na przykład przez logikę, modele semantyczne, jawne reguły lub jednoznacznie zdefiniowane cele optymalizacji oraz dostarczają algorytmów możliwych do implementacji i oceny w realnych środowiskach obliczeniowych, takich jak standardowe systemy zarządzania bazą danych oraz systemy integracji oparte na usługach i agentach. Wymaga to także uwzględnienia jakości wyników, kosztu obliczeniowego oraz interpretowalności.

W takim ujęciu przedstawiony cykl publikacji stanowi spójny wkład naukowy w rozwój formalnych i algorytmicznych metod przetwarzania informacji nieprecyzyjnej w systemach informatycznych.

2.2. Cele i teza pracy

Przedmiotem rozprawy są formalne i algorytmiczne metody przetwarzania informacji nieprecyzyjnej w systemach informatycznych. Definicję *informacji nieprecyzyjnej* przedstawiono w podrozdziale 2.1. Przyjęto takie ujęcie nieprecyzyjności, w którym semantyka informacji jest jawnie określona w modelu, przetwarzanie jest wykonalne w realnych środowiskach obliczeniowych, a jakość i koszt przetwarzania podlegają weryfikacji empirycznej. Z tego wynika połączenie perspektywy formalnej (modele, reguły i semantyka) z perspektywą algorytmiczną (transformacje, procedury wykonania oraz kontrola złożoności i kosztu).

Weryfikację tezy przeprowadzono w czterech obszarach badawczych (I–IV) odpowiadających cyklowi publikacji przedstawionych syntetycznie w rozdziale trzecim.

Obejmują one: (I) zapytania rozmyte w relacyjnych bazach danych i ich wykonanie w środowisku SQL [1], [2], (II) integrację heterogenicznych usług przy niepełnej specyfikacji celu użytkownika [3]–[5], (III) formalizację nieprecyzyjnej specyfikacji na etapie projektowania oraz dedukcyjna synteza struktury aplikacji [6] oraz (IV) rozszerzenia obejmujące inne nośniki nieprecyzyjności, takie jak informacja zewnętrzna, braki strukturalne w tekście oraz formalizacja optymalizacyjna prowadząca do kompresji informacji w danych syntetycznych [7]–[9].

Cel główny zakłada opracowanie oraz empiryczną weryfikację formalnych i algorytmicznych metod przetwarzania informacji nieprecyzyjnej w systemach informatycznych. Metody te mają zapewniać przetwarzanie skuteczne i kontrolowane, oparte na jawnie zdefiniowanej semantyce, wykonalne w praktycznych środowiskach obliczeniowych oraz umożliwiające empiryczną ocenę jakości i kosztu przetwarzania.

Cele szczegółowe:

C1. Formalizacja semantyki zapytań rozmytych w relacyjnych bazach danych poprzez takie ujęcie warunków zapytania, aby klasyczną selekcję dwuwartościową zastąpić semantyką stopniową oraz uzyskać wynik interpretowalny i możliwy do uporządkowania według stopnia spełnienia warunków.

C2. Opracowanie metodyk i algorytmów transformacji zapytań rozmytych do postaci wykonywalnej w standardowym SQL w taki sposób, aby formalna semantyka zapytania mogła być realizowana w istniejących systemach relacyjnych bez ingerencji w ich rdzeń, z zachowaniem przenośności rozwiązania między systemami DBMS oraz wykonalności obliczeniowej (kontroli kosztu).

C3. Opracowanie formalno-algorytmicznego podejścia do integracji heterogenicznych usług i źródeł danych przy niepełnej specyfikacji celu użytkownika, w którym system na podstawie formalnych opisów usług i mechanizmu wnioskowania konstruuje plan wykonania w postaci drzewa wykonania, z uwzględnieniem ograniczeń złożoności wyszukiwania i kompozycji.

C4. Ujęcie nieprecyzyjności specyfikacji na etapie projektowania oraz skrócenie drogi od modelu do implementacji poprzez wykazanie, że formalizacja wymagań umożliwia dedukcyjne wyprowadzenie (syntezę) spójnej struktury aplikacji, w tym aplikacji webowej, a tym samym ogranicza skutki niepełnych i ewoluujących wymagań przy zachowaniu kontroli semantycznej rozwiązania.

C5. Weryfikacja zakresu stosowalności schematu: formalizacja → algorytm → weryfikacja dla innych nośników nieprecyzyjności poprzez wykazanie, że podejście pozostaje zasadne również wtedy, gdy nieprecyzyjność wynika z sygnałów zewnętrznych (kontekstowości i krótkotrwałości oddziaływania), z braków struktury w tekście, które wymagają rekonstrukcji oraz z formalizacji optymalizacyjnej prowadzącej do kompresji informacji w danych syntetycznych.

Teza rozprawy

Formalne modele oraz algorytmiczne metody przetwarzania informacji nieprecyzyjnej umożliwiają wykonalną obliczeniowo, skuteczną i kontrolowaną integrację oraz analizę danych niepełnych, zaszumionych i niejednoznacznych w systemach informatycznych, pod warunkiem jednoznacznego określenia semantyki oraz empirycznej weryfikacji jakości i kosztu przetwarzania.

2.3. Zawartość rozprawy

Niniejsza rozprawa doktorska ma formę cyklu publikacji. Obejmuje streszczenie w języku polskim i angielskim, część syntetyczną prezentującą uzyskane wyniki, część źródłową zawierającą pełne teksty publikacji oraz część uzupełniającą, w której zamieszczono wykaz literatury, spis tabel i wymagane oświadczenia. Zasadnicza część rozprawy składa się z pięciu rozdziałów, których krótki opis przedstawiono poniżej.

Rozdział 1. Wykaz publikacji stanowiących podstawę rozprawy − przedstawia listę publikacji wchodzących w skład cyklu wraz z danymi bibliograficznymi, informacją o wkładzie autora oraz wybranymi wskaźnikami bibliometrycznymi i punktacją ministerialną.

Rozdział 2. Wstęp − przedstawia motywację i kontekst pracy, cele i tezę rozprawy oraz opis jej zawartości.

Rozdział 3. Prace badawcze − stanowi część syntetyczną i przedstawia wyniki badań zaprezentowane w publikacjach [1]–[9] składających się na cykl w układzie czterech obszarów badawczych I–IV. Zawiera również pozycjonowanie wybranych wyników na tle literatury w obszarze I (podrozdział 3.1.7) oraz w obszarze IV (podrozdział 3.4.4).

Rozdział 4. Teksty publikacji stanowiących podstawę rozprawy − zawiera pełne teksty publikacji wchodzących w skład cyklu.

Rozdział 5. Podsumowanie − przedstawia przekrojowe podsumowanie rozprawy w odniesieniu do celów i tezy, w tym oryginalne elementy rozprawy, ograniczenia przedstawionych rozwiązań, końcowe wnioski oraz kierunki dalszych badań.

3. Prace badawcze

W Rozdziale 3 przedstawiono opis badań zaprezentowany w publikacjach [1]–[9], które tworzą spójny cykl. Wyodrębniono cztery obszary badawcze obejmujące różne nośniki i poziomy występowania nieprecyzyjności w systemach informatycznych. W każdym z obszarów zastosowano tę samą ścieżkę badawczą: (I) sformułowanie problemu i jego formalizację, w tym jawne określenie semantyki, (II) opracowanie metody lub algorytmu wraz z realizacją wykonalną w realnym środowisku obliczeniowym oraz (III) empiryczną weryfikację jakości i kosztu przetwarzania. Taki układ pozwala zestawić cele C1–C5 z rezultatami uzyskanymi w czterech obszarach oraz porównywać przyjęte rozwiązania w różnych kontekstach zastosowań.

Przekrojowe podsumowanie wyników, wyodrębnienie oryginalnych elementów rozprawy, ograniczeń oraz kierunków dalszych badań przedstawiono w Rozdziale 5.

3.1. Obszar I: Nieprecyzyjność w danych i w zapytaniach relacyjnych – zapytania rozmyte oraz ich wykonanie i transformacja do SQL

Kryteria wyszukiwania formułowane przez użytkowników mają często charakter nieostry, natomiast relacyjny model danych i standardowy SQL operują warunkami ostrymi, czyli dwuwartościowymi. W konsekwencji użytkownik jest zmuszony do jednoznacznego doprecyzowania pojęć takich jak *tani, nowy, wysoki* lub *podobny* poprzez dobór progów i zakresów wartości. Prowadzi to do efektów brzegowych, na przykład arbitralnego odrzucenia krotek minimalnie przekraczających próg oraz do ryzyka uzyskania wyniku pustego albo wyniku o ograniczonej użyteczności, co wymusza iteracyjne modyfikowanie warunków zapytania.

Zapytania rozmyte (*ang. fuzzy query*) stanowią odpowiedź na tę sprzeczność, ponieważ pozwalają przenieść nieostrość z poziomu języka naturalnego do formalnej reprezentacji, na przykład w postaci zbiorów rozmytych i funkcji przynależności, a następnie przetwarzać ją algorytmicznie w środowisku relacyjnym. Dzięki temu warunki zapytania mogą być spełniane w różnym stopniu, a wynik może być interpretowany oraz porządkowany według stopnia spełnienia warunków rozumianego jako miara dopasowania do kryteriów użytkownika.

W polskiej literaturze spotyka się różne określenia odnoszące się do zapytań formułowanych z użyciem aparatu logiki rozmytej (m.in. zapytania nieprecyzyjne, zapytania rozmyte). W niniejszej rozprawie przyjęto termin zapytanie rozmyte jako odpowiednik angielskiego *fuzzy query*, natomiast określenie zapytanie nieprecyzyjne traktowane jest jako pojęcie nadrzędne, obejmujące szerszą klasę zjawisk nieprecyzyjności.

Obszar badawczy I opiera się na wynikach przedstawionych w [1], [2] i koncentruje się na sformalizowaniu semantyki stopniowej warunków rozmytych w zapytaniach relacyjnych oraz na metodyce ich realizacji w standardowym SQL, w tym na algorytmach wyznaczania stopni spełnienia dla warunków prostych i złożonych oraz na transformacji zapytania do postaci wykonywalnej.

3.1.1. Formalizacja zapytań rozmytych i stopnia spełnienia warunków

W przedstawionej formalizacji warunek zapytania wyrażony terminem lingwistycznym jest modelowany jako zbiór rozmyty opisany funkcją przynależności. Dopasowanie krotki do warunku nie ma charakteru dychotomicznego (spełnia / nie spełnia), lecz jest wyrażane stopniem spełnienia w przedziale [0,1]. W przypadku warunków złożonych stopnie cząstkowe są agregowane operatorami odpowiadającymi spójnikom logicznym. Wykorzystuje się przy tym operatory agregacji zdefiniowane na bazie $\mathrm{T-norma}$ i $\mathrm{S-norm}$ [95]–[101], dzięki czemu możliwe jest zachowanie spójnej semantyki konstrukcji AND i OR oraz precyzyjne zdefiniowanie operacji negacji dla NOT. Wybór konkretnej $\mathrm{T-normy}$ i odpowiadającej jej $\mathrm{S-normy}$ (np. $\mathrm{MIN/MAX}$ lub PROD i suma algebraiczna) determinuje sposób agregacji stopni spełnienia w warunkach złożonych. Tak zdefiniowana semantyka stopniowa prowadzi do dwóch podstawowych mechanizmów użytkowych. Pierwszym jest porządkowanie wyników według stopnia spełnienia warunków, drugim ograniczanie zbioru wynikowego przez zastosowanie przekrojów α, czyli przyjęcia minimalnego wymaganego stopnia spełnienia warunków.

Na poziomie wykonania zapytania istotne jest, by semantyka ta była nie tylko formalnie poprawna, lecz także jednoznacznie interpretowalna w kontekście relacji wynikowej. Dlatego wynik zapytania jest traktowany jako relacja wzbogacona o jawnie wyznaczaną miarę dopasowania, na przykład atrybut FUZZY_DEGREE [1], [2], [25], która może być następnie wykorzystywana zarówno do filtrowania, jak i do porządkowania odpowiedzi.

3.1.2. Reprezentacja terminów lingwistycznych w relacyjnej bazie danych

Aby umożliwić praktyczne wykonywanie zapytań rozmytych w systemie relacyjnym, terminy lingwistyczne powiązano z reprezentacją danych możliwą do przechowywania w tabelach. Parametry funkcji przynależności, na przykład funkcji trapezowej, są przechowywane w tabelach metadanych. Umożliwia to jednoznaczne wiązanie terminów z konkretnymi atrybutami relacji oraz zarządzanie ich definicjami (modyfikację i ponowne

wykorzystanie) w oparciu o standardowe mechanizmy DBMS, takie jak tabele, zapytania SQL i kontrola uprawnień.

W implementacji referencyjnej funkcje przynależności zaimplementowano w języku PL/SQL w środowisku Oracle Database 11g Express Edition (Oracle 11g XE). W tym samym środowisku przeprowadzono wykonanie zapytań oraz pomiary czasu. Wynik zapytania wzbogacono o jawnie wyznaczaną miarę dopasowania (np. atrybut FUZZY_DEGREE), która jest dalej wykorzystywana do filtrowania i porządkowania odpowiedzi.

W takim ujęciu słownik terminów lingwistycznych stanowi formalne powiązanie między sposobem formułowania potrzeb informacyjnych przez użytkownika a semantyką obliczeniową realizowaną w systemie. Umożliwia on także kontrolowaną parametryzację znaczeń. To samo pojęcie może być różnie zdefiniowane w zależności od kontekstu dziedziny lub preferencji użytkownika, co sprzyja adaptacji semantyki zapytań w zastosowaniach praktycznych. Szczegóły implementacyjne przedstawiono w [1], [2].

3.1.3. Transformacja zapytań rozmytych do postaci wykonywalnej w standardowym SQL

Kluczowym problemem systemowym jest przejście od formalnego opisu nieostrości do rozwiązania, które może zostać wykonane w istniejącym środowisku relacyjnym, czyli w systemie realizującym standardowy SQL wraz z typowymi mechanizmami wykonania i optymalizacji zapytań. W ramach obszaru I przedstawiono metodyki transformacji, które sprowadzają konstrukcje rozmyte do postaci wyrażeń SQL wykonywalnych bez ingerencji w rdzeń systemu zarządzania bazą danych, przy zachowaniu jednoznacznej interpretacji semantyki stopniowej.

W ujęciu ogólnym transformacja polega na jawnym obliczaniu wartości funkcji przynależności dla atrybutów występujących w warunkach lingwistycznych oraz na wyznaczeniu miary dopasowania krotki i jej propagacji w warunkach złożonych. Spójniki logiczne są odwzorowywane na operacje agregujące stopnie spełnienia zgodnie z przyjętą parą operatorów $T-$normy i $S-$normy. Dodatkowo zbiór wynikowy może zostać ograniczony przez zastosowanie przekrojów α, czyli przyjęcie minimalnego wymaganego stopnia spełnienia, a wyznaczona miara dopasowania jest następnie wykorzystywana do porządkowania wyników zapytania.

Zakres transformacji obejmuje również konstrukcje związane z grupowaniem i agregacją. Rozpatrzono rozmyte odpowiedniki wybranych agregatów i sformułowano ich definicje.

Niech G oznacza zbiór krotek tworzących rozpatrywaną grupę (np. po GROUP BY), a dla każdej krotki $t \in G$ niech $\mu(t) \in [0,1]$ oznacza stopień spełnienia warunku rozmytego przez t (w implementacji wyznaczany jako miara dopasowania, np. atrybut FUZZY_DEGREE). Rozmyte zliczanie definiuje się jako sumę stopni spełnienia:

$$\text{FUZZY\_COUNT}(G) = \sum_{t \in G} \mu(t) \tag{1}$$

Z (1) wynika, że wartość zliczania nie musi być liczbą całkowitą, ponieważ wkład każdej krotki jest proporcjonalny do stopnia spełnienia warunku.

Rozmytą średnią wartości atrybutu x definiuje się jako średnią ważoną, w której wagami są stopnie spełnienia:

$$\text{FUZZY\_AVG}(G, x) = \frac{\sum_{t \in G} \mu(t) * x(t)}{\sum_{t \in G} \mu(t)} \tag{2}$$

Z równania (2) wynika, że wkład krotek do wartości FUZZY_AVG jest ważony stopniem spełnienia $\mu(t)$: krotki o wyższym $\mu(t)$ mają większy udział w wyniku, natomiast krotki o niskim $\mu(t)$ są uwzględniane w mniejszym stopniu.

3.1.4. Dwa warianty realizacji transformacji w DBMS i po stronie aplikacji

W celu zapewnienia wykonalności w zróżnicowanych środowiskach przedstawiono dwa komplementarne warianty realizacji transformacji zapytań rozmytych. Pierwszy wariant zakłada realizację po stronie systemu zarządzania bazą danych (DBMS). W tym ujęciu obliczanie stopni spełnienia oraz elementy transformacji są wykonywane w bazie danych z wykorzystaniem programowalnych mechanizmów DBMS, tj. zdefiniowanych w bazie funkcji i procedur składowanych. Rozwiązanie to pozwala wykonywać obliczenia w bezpośrednim sąsiedztwie danych oraz korzystać z natywnych mechanizmów wykonania i optymalizacji zapytań relacyjnych. Drugi wariant zakłada realizację po stronie aplikacji komunikującej się z bazą danych. W tym przypadku część logiki związanej z interpretacją i przekształceniem zapytania jest wykonywana poza DBMS; dotyczy to zwłaszcza warunków w klauzuli WHERE. Warunki te są najpierw rozkładane na elementy składniowe (operandy, operatory i nawiasy), a następnie przekształcane do postaci umożliwiającej konstrukcję drzewa wyrażeń z wykorzystaniem algorytmu stosowego, tj. procedury opartej na stosie, uwzględniającej priorytety operatorów i nawiasy. Następnie redukcja jest prowadzona z zachowaniem przyjętych operatorów rozmytych, tj. $T-\text{norm}$ i $S-\text{norm}$, zapewniając tym samym spójne wyznaczanie stopnia spełnienia dla warunków złożonych zgodnie ze strukturą

drzewa, np. w sposób rekurencyjny. W efekcie uzyskuje się jednoznaczną postać zapytania wykonywalną w SQL.

Oba warianty realizują tę samą ideę polegającą na przejściu od formalnego opisu nieostrości do algorytmów i struktur umożliwiających wykonanie zapytania w praktycznym środowisku obliczeniowym. Różnią się jednak miejscem realizacji obliczeń i wynikającymi z tego konsekwencjami. Dotyczą one zależności od konkretnego DBMS i dostępnych mechanizmów rozszerzeń oraz rozkładu kosztu obliczeniowego pomiędzy serwerem bazy danych a warstwą aplikacyjną. Obejmują także koszt komunikacji aplikacja − DBMS w scenariuszach wymagających dodatkowych przeliczeń oraz łatwość zastosowania rozwiązania w środowiskach heterogenicznych, w których możliwości rozszerzania DBMS są ograniczone.

3.1.5. Weryfikacja wykonalności i kosztu obliczeniowego

Weryfikację w obszarze I przeprowadzono na zestawie reprezentatywnych klas zapytań rozmytych obejmujących warunki lingwistyczne, warunki złożone oraz przypadki wykorzystujące grupowanie i agregację. Dla tych klas wykazano możliwość uzyskania postaci wykonywalnej w standardowym SQL, w której stopień spełnienia warunków jest wyznaczany jawnie i może być wykorzystywany zarówno do ograniczania zbioru wynikowego z zastosowaniem przekrojów α, jak i do porządkowania odpowiedzi według stopnia spełnienia.

Równolegle przeprowadzono ocenę kosztu obliczeniowego w funkcji rosnącej złożoności zapytań oraz rosnącej skali danych. Analiza obejmowała wpływ liczby i typu warunków rozmytych, w tym warunków opartych na rozkładach możliwości i relacjach podobieństwa, a także wpływ struktury warunków złożonych, w których stopnie spełnienia agregowano zgodnie z przyjętą parą operatorów $\mathrm{T}-\mathrm{normy}$ i $\mathrm{S}-\mathrm{normy}$. W eksperymentach przyjęto wariant odpowiadający operatorom $\mathrm{MIN}/\mathrm{MAX}$ dla koniunkcji i alternatywy. Uzyskane wyniki potwierdzają, że narzut związany z wyznaczaniem stopni spełnienia oraz ich agregacją jest mierzalny i możliwy do kontrolowania w ramach przyjętej metody transformacji, przy zachowaniu wykonalności w środowisku relacyjnym. Dla zapytań o rosnącej złożoności, uruchamianych na 1000 rekordów w środowisku Oracle 11g XE, czasy wykonania mieściły się w zakresie 8–54 ms, a analiza skalowalności przy zmianie liczby rekordów od 10 do 1000 potwierdziła przewidywalny wzrost kosztu wraz z rozmiarem danych.

Warto podkreślić, że celem eksperymentów była ocena narzutu i trendów kosztu wynikających z transformacji oraz z przyjętej semantyki stopniowej, a nie porównawcza ocena wydajności różnych systemów bazodanowych. Szczegółową konfigurację eksperymentów oraz zestawy zapytań testowych przedstawiono w publikacjach [1], [2].

3.1.6. Wyniki I obszaru badawczego

W obszarze I uzyskano spójne ujęcie, w którym nieostrość (rozmytość) warunków zapytań relacyjnych została formalnie zdefiniowana w postaci semantyki stopniowej, a następnie przełożona na algorytmiczne procedury pozwalające uzyskać postać wykonywalną w standardowym SQL. Przedstawiono odtwarzalną metodykę transformacji zapytań rozmytych do zapytań SQL obejmującą zasady wyznaczania stopni spełnienia dla warunków prostych i złożonych oraz ich propagacji w zapytaniu, z zachowaniem jawnej interpretacji wyniku.

Wynik zapytania może być traktowany jako relacja wzbogacona o miarę dopasowania, umożliwiając zarówno filtrowanie odpowiedzi przez przekroje α, jak i porządkowanie wyników według stopnia spełnienia. Ponadto wskazano możliwość objęcia transformacją wybranych konstrukcji grupowania i agregacji poprzez zdefiniowanie rozmytych odpowiedników agregatów (m.in. (1) FUZZY_COUNT i (2) FUZZY_AVG) oraz wskazanie ich jednoznacznej realizacji w standardowym SQL jako operacji wykorzystujących stopnie spełnienia.

Istotnym rezultatem jest wyróżnienie dwóch uzupełniających się wariantów realizacji transformacji. Pierwszy zakłada wykonanie obliczeń po stronie DBMS z wykorzystaniem programowalnych mechanizmów systemu, tj. zdefiniowanych w bazie funkcji i procedur składowanych. W drugim wariancie warunki zapytania są przetwarzane algorytmicznie w warstwie aplikacyjnej (rozkład na elementy składniowe oraz procedura oparta na stosie prowadząca do konstrukcji drzewa wyrażeń), co umożliwia uzyskanie jednoznacznej postaci zapytania wykonywalnej w SQL oraz spójne wyznaczanie stopnia spełnienia dla warunków złożonych.

Weryfikacja empiryczna potwierdziła, że transformacja zapytań rozmytych do postaci wykonywalnej w standardowym SQL może być realizowana w relacyjnym środowisku bazodanowym. W badaniach oszacowano narzut obliczeniowy wynikający z wyznaczania i agregowania stopni spełnienia, analizując go wraz ze wzrostem złożoności zapytań oraz skali danych. Wyniki pokazują, że przetwarzanie informacji nieprecyzyjnej w relacyjnych systemach bazodanowych może być realizowane bez modyfikacji rdzenia DBMS, przy zachowaniu interpretowalności odpowiedzi dzięki jawnie zdefiniowanej semantyce stopniowej.

Na tle prac dotyczących elastycznego wyszukiwania w relacyjnych bazach danych w obszarze I akcent położono na wykonanie zapytań rozmytych w standardowym SQL, tj. bez ingerencji w silnik DBMS, przy zachowaniu formalnie określonej semantyki. Zestawienie

porównawcze podejść oraz syntetyczne umiejscowienie wkładu obszaru I przedstawiono w Tabeli 3.1.1 oraz w Tabeli 3.1.2.

Wyniki obszaru I opublikowano w [1], [2], natomiast ograniczenia rozwiązań oraz kierunki dalszych badań omówiono w Rozdziale 5.

3.1.7. Pozycjonowanie wyników obszaru I na tle literatury

Zestawienie funkcjonalne (Tabela 3.1.1) propozycji z literatury dotyczących zapytań rozmytych w relacyjnych bazach danych pokazuje, że wiele elementów funkcjonalności, takich jak wyznaczanie stopnia spełnienia, rozkłady możliwości czy operatory porównań rozmytych, było obecnych od dziesięcioleci. Jednocześnie porównanie wskazuje częsty rozdźwięk między zakresem deklarowanej funkcjonalności a możliwością jej realizacji w praktycznym środowisku relacyjnym. W tym kontekście ostatnia kolumna tabeli stanowi syntetyczne ujęcie wkładu [1], [2] w obszar I oraz identyfikuje, które elementy opisywane na poziomie modeli zostały przełożone na postać wykonywalną w standardowym SQL.

Tabela 3.1.1. Porównanie funkcjonalne wybranych podejść do zapytań rozmytych w relacyjnych bazach danych

Model	**Praca / podejście**							
	[32]	[102]	[103]	[37]	[104]	[41]	[105]	**[1], [2]**
Obsługa danych skalarnych	✓	✓	✓	✓	✓	✓	✓	✓
Obsługa danych złożonych (nieskalarnych)	✓	✓	✓	✓	–	–	–	–
Relacja podobieństwa	✓	–	✓	✓	–	–	–	✓
Rozkłady możliwości	–	✓	✓	✓	–	✓	✓	✓
Stopień spełnienia na poziomie atrybutu	✓	✓	–	✓	✓	–	–	–
Stopień spełnienia na poziomie krotki (rekordu)	–	✓	✓	✓	✓	✓	✓	✓
Modyfikatory rozmyte	–	–	✓	–	–	✓	–	–
Kwantyfikatory rozmyte	–	–	–	✓	–	✓	✓	–
Operatory porównania rozmyte	✓	–	✓	✓	✓	✓	✓	✓
Grupowanie rozmyte (GROUP BY)	–	–	–	–	–	✓	✓	✓
Złączenia rozmyte (fuzzy joins)	–	✓	–	✓	✓	✓	–	–
Przechowywanie danych rozmytych	✓	✓	✓	✓	✓	–	–	–
Zapytania rozmyte	✓	✓	✓	✓	✓	✓	✓	✓
Rozszerzenia SQL	–	–	–	✓	✓	✓	✓	✓

Zestawienie potwierdza, że zaproponowane podejście nie polega na wprowadzeniu nowych pojęć do modelu rozmytego, lecz na domknięciu ścieżki: formalizacja → wykonanie w warunkach ograniczeń praktycznych. Kluczowe jest tu utrzymanie jawnej interpretacji stopnia spełnienia w wyniku oraz możliwość realizacji zapytania w standardowym środowisku SQL, bez modyfikacji rdzenia DBMS.

Tabela 3.1.2. Typologia podejść do realizacji zapytań rozmytych

Praca / podejście	Typ rozwiązania	Oparte na SQL	Zakres zapytań	Kluczowe zagadnienia
SEQUEL, język zapytań rozmytych [106]	język	–	zapytania proste	terminy lingwistyczne (wartości lingwistyczne)
FQUERY III+, prototypowy system zapytań rozmytych [107]	system/ prototyp	–	zapytania proste	kwantyfikatory lingwistyczne
Prototypowy procesor zapytań rozmytych [108]	procesor/ prototyp	–	zapytania proste	klasteryzacja rozmyta z progowaniem
Ujęcie oparte na rozmytej całce Sugeno [109]	koncepcja formalna	–	zapytania proste	spójniki logiczne, zdania rozmyte z kwantyfikacją, operator dzielenia relacji rozmytych
FQL, język zapytań rozmytych [110]	język	–	zapytania proste	konstrukcje językowe obejmujące modyfikację, kompozycję, kwantyfikację, kwalifikację
SQLf, język zapytań rozmytych [41]	język/ rozszerzenie SQL	✓	zapytania proste	predykaty rozmyte, kwantyfikatory, ograniczenia rozmyte, złączenia rozmyte, operatory rozmyte
FQUERY, język zapytań rozmytych [111]	język/ rozszerzenie SQL	✓	zapytania proste	terminy lingwistyczne, relacje rozmyte, kwantyfikatory lingwistyczne
FSQL, język zapytań rozmytych [112]	język/ rozszerzenie SQL	✓	zapytania proste	etykiety lingwistyczne, komparatory rozmyte, progi spełnienia, kwalifikatory, wartości rozmyte
Soft-SQL, język zapytań rozmytych [113]	język/ rozszerzenie SQL	✓	zapytania proste	predykaty lingwistyczne, kwantyfikatory lingwistyczne
Bipolarność w teorii możliwości [114]	koncepcja formalna	–	zapytania złożone	zapytania z preferencjami negatywnymi i pozytywnymi
Preferencje rozmyte [115]	koncepcja/ model	✓	zapytania złożone	podejścia ilościowe (oceny punktowe) i jakościowe (relacje preferencji)
Transformacja zapytań rozmytych do standardowego SQL oraz wykonanie w relacyjnym DBMS bez ingerencji w rdzeń [1], [2]	**metodyka, algorytmy wykonawcze**	✓	**zapytania proste oraz wybrane elementy zapytań złożonych**	**jawna semantyka stopniowa, parametry funkcji przynależności w tabelach, propagacja i akumulacja stopnia spełnienia, redefinicja agregatów, dobór operatorów łączenia (T-normy i S-normy), wykonalność i koszt w czystym SQL**

Uzupełnieniem porównania funkcjonalnego jest zestawienie typologiczne (Tabela 3.1.2), które porządkuje propozycje z literatury według strategii realizacji zapytań rozmytych. Pozwala ono odróżnić rozwiązania skoncentrowane na rozszerzeniach języka i prototypowych procesorach zapytań od tych akcentujących mechanizm ewaluacji, w tym progi akceptacji (przekroje α) oraz porządkowanie wyników.

W ostatnim wierszu tej tabeli ujęto syntetycznie wkład [1], [2] w obszar I, pokazując, że zapytania rozmyte zostały sprowadzone do postaci wykonywalnej w standardowym SQL w relacyjnym DBMS bez ingerencji w rdzeń, przy zachowaniu jawnej semantyki stopniowej oraz z uwzględnieniem mechanizmów takich jak propagacja stopnia spełnienia, redefinicja agregatów i dobór operatorów łączenia ($\mathrm{T}-\mathrm{normy}\ /\ \mathrm{S}-\mathrm{normy}$).

Wynik typologii wspiera tezę, że o zastosowalności zapytań rozmytych decyduje nie tylko ekspresyjność formalizmu, ale przede wszystkim możliwość jego realizacji w typowym środowisku relacyjnym oraz kontrola kosztu wykonania. Z tego względu w [1], [2] priorytetem jest transformacja do standardowego SQL oraz wykazanie wykonalności i narzutu obliczeniowego w eksperymencie.

3.2. Obszar II: Nieprecyzyjność na poziomie celu użytkownika – integracja heterogenicznych usług

Obszar badawczy II dotyczy klas problemów, w których źródłem nieprecyzyjności nie są same dane (jak w przypadku zapytań rozmytych), lecz niepełność specyfikacji procesu przetwarzania. Użytkownik potrafi określić, jaki rezultat chce uzyskać (cel obliczeniowy wyrażony w języku dziedziny), natomiast nie dysponuje wiedzą, w jaki sposób ustalić sekwencję kroków przetwarzania (przepływ pracy, *ang. workflow*) oraz dobrać narzędzia, usługi i źródła danych w środowisku heterogenicznym. W konsekwencji dobór komponentów oraz ich kompozycja nie są zadane jawnie na wejściu i muszą zostać wyznaczone w sposób systematyczny, z zachowaniem poprawności i wykonalności procesu.

Problem ten jest szczególnie widoczny w zastosowaniach interdyscyplinarnych realizowanych w ekosystemie World Data System oraz ośrodków World Data Center, gdzie współistnieją rozproszone aplikacje, usługi i zasoby danych o zróżnicowanych formatach, interfejsach i protokołach. W takich warunkach ręczne konstruowanie przepływu pracy staje się istotną barierą dla użytkownika-badacza, ograniczając możliwość efektywnego wykorzystania dostępnych zasobów i narzędzi w ramach jednego spójnego procesu obliczeniowego.

Aby temu zaradzić, system konstruuje plan wykonania w postaci drzewa wykonania (*ang. execution tree*), które w realizacji ma postać acyklicznego grafu skierowanego (DAG, *ang. directed acyclic graph*). W tej strukturze obok kroków sekwencyjnych mogą występować operacje rozdzielania i scalania danych, dzięki czemu możliwe jest wskazanie fragmentów, które mogą być wykonywane równolegle.

W obszarze badawczym II, rozwiniętym w publikacjach [3]–[5], przedstawiono metodykę przejścia od celu użytkownika do wykonalnego planu integracji usług (przepływu pracy) poprzez formalny opis usług w bazie wiedzy, wnioskowanie prowadzące do konstrukcji dowodu oraz rekonstrukcję planu wykonania.

3.2.1. Formalizacja celu użytkownika i usług, baza wiedzy oraz warunki wstępne i końcowe

Punktem wyjścia w przyjętej metodyce jest założenie, że brak jawnie określonego przebiegu procesu przetwarzania (przepływu pracy) można kompensować poprzez formalną reprezentację wiedzy o usługach oraz o celu użytkownika. Każda usługa jest opisywana jako metoda o określonych warunkach wstępnych i końcowych (*ang. preconditions/ postconditions*). Analogicznie formalizowany jest problem użytkownika, tj. oczekiwane przejście od stanu wejściowego do stanu docelowego.

Formalizm oparto na klauzulowej logice pierwszego rzędu wykorzystującej klauzule Horna (*ang. horn clauses*), co pozwala sprowadzić kompozycję usług do zadania wnioskowania w bazie wiedzy. W notacji stosowanej w [3]–[5] wykorzystuje się specyfikatory opisujące konstrukcje używane w formułach logicznych, metody (usługi) wraz z ich warunkami wstępnymi i końcowymi oraz specyfikacje celu użytkownika. Zależności między metodami i danymi są zapisywane w postaci klauzul, które stanowią podstawę dalszego dowodzenia wykonalności celu.

Baza wiedzy jest dodatkowo wspierana warstwą ontologiczną. Opisy usług i źródeł danych mogą być wyrażone w standardach OWL/RDF, co porządkuje semantykę pojęć dziedzinowych i ujednolica opisy w środowisku heterogenicznym. W tym ujęciu warstwa ta dostarcza spójnej terminologii, natomiast klauzule i reguły wnioskowania umożliwiają algorytmiczne konstruowanie planu wykonania.

3.2.2. Algorytmiczne przetwarzanie, wnioskowanie i konstrukcja dowodu jako kompozycji

System generuje odpowiedź na zapytanie użytkownika poprzez konstrukcję dowodu, który formalnie wskazuje, jak zestawić dostępne usługi, aby przejść od stanu początkowego, tj. zestawu spełnionych warunków wstępnych oraz dostępnych zasobów do oczekiwanego rezultatu opisanego jako warunek końcowy. Mechanizm wnioskowania opiera się na metodzie m – rezolucji (*ang. m-resolution*) oraz jej wariancie uporządkowanej liniowej m – rezolucji (*ang. ordered linear m-resolution*). Strategia ta operuje na m-klauzulach rozumianych jako multizbiory literałów (dopuszczających powtórzenia), w której dowód powstaje jako sekwencja kroków unifikacji i redukcji prowadząca od sformułowanego celu do warunków początkowych [3]–[5]. Zastosowanie wariantu uporządkowanego narzuca strategię doboru reguł w kolejnych krokach dowodzenia, ograniczając liczbę alternatywnych ścieżek wnioskowania i ułatwiając implementację mechanizmu planowania w systemie integracyjnym.

Uzyskany dowód reprezentuje kompozycję usług. Określa, które aksjomaty i metody należy zastosować oraz w jakiej zależności, aby uzyskać stan spełniający cel użytkownika. Dowód pełni w tym ujęciu rolę reprezentacji pośredniej, z której następnie rekonstruowany jest plan wykonania.

Kontrola kosztu obliczeniowego stanowi odrębne wyzwanie. Wraz ze wzrostem liczby dostępnych usług i możliwych podstawień rośnie liczba alternatywnych ścieżek dowodzenia, co może istotnie wydłużać czas znajdowania rozwiązania. Aby temu przeciwdziałać, zastosowano selektywne sterowanie przeszukiwaniem, w tym mechanizm oparty na abstrakcji typów (*ang. typification abstraction*). W pierwszym etapie wnioskowanie jest prowadzone na poziomie uogólnionych typów (klas pojęć), a dopiero potem zawężane do instancji szczegółowych. Pozwala to wcześniej odrzucać kierunki nieprowadzące do celu i ograniczać koszt wnioskowania w przestrzeni rozwiązań. Wynikiem wnioskowania nie jest jedynie odpowiedź „wykonalny / niewykonalny”, lecz dowód, który stanowi podstawę do algorytmicznej rekonstrukcji struktury rozwiązania w kolejnych etapach. W ten sposób weryfikacja obejmuje zarówno poprawność konstrukcji rozwiązania (dowód i rekonstrukcja planu), jak i kontrolę kosztu obliczeniowego przez mechanizmy ograniczania złożoności przeszukiwania.

3.2.3. Rekonstrukcja planu wykonania, od dowodu do rozwiązania

Sam dowód logiczny nie jest jeszcze planem wykonania w sensie systemowym. Kluczowym elementem jest zatem rekonstrukcja struktury rozwiązania (*ang. solution tree recovery*), czyli

algorytmiczne przekształcenie dowodu w plan wykonania w postaci drzewa. Drzewo to można interpretować jako acykliczny graf skierowany (DAG) opisujący wykonalną integrację usług. Węzły planu odpowiadają usługom (metodom) oraz operacjom przekształceń danych. Uwzględniane są również operacje rozdzielania i scalania strumieni danych, co pozwala modelować fragmenty wykonywane równolegle pomiędzy etapami rozgałęzienia i ponownego połączenia.

Oznacza to przejście od specyfikacji celu użytkownika, w której nie jest podany jawnie określony przebieg procesu przetwarzania (przepływ pracy), do planu wykonania realizowanego jako uporządkowany zestaw wywołań usług uzupełniony o pośrednie transformacje danych i potencjalną równoległość wybranych fragmentów.

3.2.4. Aspekt systemowy, architektura integracji i warstwa wykonawcza

Rozwiązanie osadzono w architekturze dwupoziomowej obejmującej warstwę usług bazowych oraz warstwę inteligentnych agentów pełniących rolę orkiestratorów. Agenci odpowiadają za interpretację celu użytkownika, konstrukcję dowodu, rekonstrukcję planu wykonania oraz uruchomienie przebiegu procesu przetwarzania w środowisku integracyjnym.

W warstwie integracyjnej przyjęto rozwiązania typowe dla architektury zorientowanej na usługi SOA (*ang. Service-Oriented Architecture*), w tym rejestrację i odkrywanie usług UDDI (*ang. Universal Description, Discovery and Integration*), opis interfejsów WSDL (*ang. Web Services Description Language*), komunikację SOAP (*ang. Simple Object Access Protocol*) oraz orkiestrację BPEL (*ang. Business Process Execution Language*). Do realizacji warstwy agentowej wykorzystano platformę JADE *(ang. Java Agent DEvelopment Framework).* Warstwa wykonawcza rozdziela poziom specyfikacji od poziomu realizacji i wspiera przejście od celu użytkownika, przez plan wykonania, do uruchomienia złożonego procesu opartego na usługach.

Takie ujęcie wiąże część formalną i algorytmiczną. Formalne opisy usług w bazie wiedzy (warunki wstępne i końcowe oraz aksjomaty) są bezpośrednio wykorzystywane przez algorytmy wnioskowania i rekonstrukcji planu, a wynik jest realizowany w systemie przeznaczonym do pracy w środowisku heterogenicznym.

3.2.5. Weryfikacja: demonstracja przejścia od celu do planu wykonania

Weryfikacja polega na wykazaniu, że użytkownik może opisać problem w języku dziedziny, bez jawnego określenia przebiegu procesu przetwarzania (przepływu pracy), a system potrafi automatycznie skonstruować plan wykonania jako kompozycję usług prowadzącą do uzyskania

wyniku. Studium przypadku przeprowadzono dla scenariusza analizy wskaźników zagrożeń oraz bezpieczeństwa życia w wybranych regionach Ukrainy. Mechanizm wnioskowania najpierw konstruuje dowód realizowalności celu, a następnie na jego podstawie rekonstruowany jest plan wykonania w postaci drzewa wykonania w mechanizmie rekonstrukcji struktury rozwiązania. Struktura ta ma postać acyklicznego grafu skierowanego (DAG). Wykonanie polega na wywoływaniu usług zgodnie z zależnościami danych i kolejnością wynikającą z planu, z uwzględnieniem fragmentów możliwych do uruchomienia równolegle pomiędzy etapami rozdzielania i scalania danych. W publikacji [5] zaprezentowano przykładowy rezultat rekonstrukcji w postaci wyznaczonej sekwencji wywołań usług $(\mathrm{S1}, \mathrm{S2}, \mathrm{S5}, \mathrm{S7}, \mathrm{S6}, \mathrm{S9})$, gdzie S_i oznacza usługę/metodę opisaną formalnie przez parę warunków wstępnych i końcowych w notacji użytej [5], przy czym gałęzie po rozdzieleniu danych mogą być wykonywane równolegle do momentu ich scalenia.

3.2.6. Wyniki II obszaru badawczego

Wyniki obszaru badawczego II stanowią spójne ujęcie, w którym nieprecyzyjność spowodowana niepełną specyfikacją przebiegu procesu przetwarzania (przepływu pracy) jest domknięta w schemacie: formalizacja → wnioskowanie → plan wykonania → wykonanie. Użytkownik określa cel w kategoriach dziedzinowych, natomiast system uzupełnia brakującą część specyfikacji poprzez automatyczną konstrukcję wykonalnego planu integracji usług w środowisku heterogenicznym.

Formalny rdzeń rozwiązania polega na jawnym opisie usług i problemu użytkownika w bazie wiedzy. Usługi są modelowane przez warunki wstępne i końcowe oraz aksjomaty w postaci klauzul Horna, dzięki czemu kompozycja usług zostaje sprowadzona do zadania dowodzenia twierdzeń. Warstwa algorytmiczna realizuje to zadanie poprzez wnioskowanie oparte na metodzie $\mathrm{m}-\mathrm{rezolucji}$ oraz jej wariancie − $\mathrm{uporządkowanej\ liniowej\ m}-\mathrm{rezolucji}$ − prowadząc tym samym do konstrukcji formalnego dowodu realizowalności postawionego celu.

Rezultatem metodycznym jest wykazanie, że dowód może zostać przekształcony w jednoznaczny artefakt wykonawczy. Zastosowany mechanizm rekonstrukcji struktury rozwiązania wyznacza plan wykonania w postaci drzewa wykonania, które ma postać skierowanego grafu acyklicznego (DAG). W strukturze tej wyróżnia się węzły odpowiadające metodom/usługom oraz operacjom przetwarzania danych, w tym rozdzielaniu i scalaniu strumieni. Umożliwia to identyfikację fragmentów możliwych do uruchomienia równolegle.

Wyniki obejmują również aspekt złożoności. Zastosowano selektywne sterowanie przeszukiwaniem z wykorzystaniem abstrakcji typów. Najpierw wnioskowanie odbywa się na poziomie uogólnionych typów pojęć powiązanych z ontologią, a dopiero w kolejnym kroku rozwiązanie jest uszczegóławiane do poziomu predykatów i instancji usług. Oznacza to zawężenie zbioru rozpatrywanych aksjomatów i usług do tych zgodnych typowo z wymaganym formatem danych, zanim zostaną rozważone szczegółowe podstawienia. Jeżeli w trakcie konstrukcji dowodu dla danego kroku istnieje kilka zgodnych aksjomatów, mechanizm buduje równoległe warianty dowodu generując tym samym zbiór alternatywnych kompozycji o różnej długości i złożoności. Następnie wybierany jest wariant o najmniejszej liczbie wierzchołków (trójek) oraz o najmniejszej liczbie użytych aksjomatów, natomiast warianty dłuższe, cykliczne lub prowadzące do ślepych zakończeń są odrzucane.

Na poziomie systemowym rozwiązanie osadzono w architekturze SOA oraz w warstwie agentowej (JADE), co pozwala realizować przejście od celu, przez plan, do uruchomienia usług zgodnie z zależnościami danych i kolejnością wynikającą z planu. W warstwie agentowej wyróżniono agenta nadrzędnego odpowiedzialnego za przetwarzanie zapytania użytkownika w oparciu o bazę wiedzy (aksjomaty i reguły wnioskowania) oraz agentów poziomu usługowego odpowiedzialnych za wyszukiwanie dostępnych usług i zasobów oraz uruchamianie planu. Przepływ realizacji obejmuje powiązanie opisów usług (np. z rejestru i opisu interfejsów) z odpowiadającymi im aksjomatami w bazie wiedzy, konstrukcję dowodu realizowalności celu, rekonstrukcję drzewa wykonania na podstawie dowodu, a następnie wywoływanie usług zgodnie z zależnościami danych, z uwzględnieniem rozdzielania i scalania strumieni.

Weryfikacja została przedstawiona na przykładzie studium przypadku analizy wskaźników zagrożeń i bezpieczeństwa życia w regionach Ukrainy; w publikacji [5] zaprezentowano przykładowy rezultat rekonstrukcji w postaci sekwencji wywołań usług $(\mathrm{S1}, \mathrm{S2}, \mathrm{S5}, \mathrm{S7}, \mathrm{S6}, \mathrm{S9})$, gdzie S_i oznacza usługę/metodę opisaną formalnie przez parę warunków wstępnych i końcowych, a kolejność realizacji wynika z zależności danych pomiędzy warunkami końcowymi poprzedników a warunkami wstępnymi następców, z możliwością równoległego wykonania gałęzi pomiędzy etapami rozdzielania i scalania danych.

Wyniki obszaru badawczego II opublikowano w pracach [3]–[5], natomiast ograniczenia rozwiązań oraz kierunki dalszych badań omówiono w Rozdziale 5.

3.3. Obszar III: Nieprecyzyjność specyfikacji na etapie projektowania systemów

Obszar badawczy III dotyczy nieprecyzyjności ujawniającej się przed implementacją, podczas definiowania wymagań oraz specyfikacji projektowej. W tworzeniu aplikacji webowych wymagania są często formułowane na wysokim poziomie ogólności, bywają niepełne i ewoluują w czasie, a ich interpretacja zależy od kontekstu dziedziny, przyjętego stylu architektonicznego oraz dostępnych komponentów. W rezultacie nieprecyzyjność ma w tym obszarze charakter semantyczno-projektowy. Oznacza to, że nie istnieje prosta, jednoznaczna reguła odwzorowania pozwalająca bezpośrednio przejść od opisu potrzeb użytkownika do doboru komponentów systemu oraz ich konfiguracji. Ręczne przełożenie wymagań na spójną strukturę aplikacji bywa kosztowne, czasochłonne i podatne na niespójności, zwłaszcza gdy specyfikacja jest iteracyjnie modyfikowana w trakcie projektu.

W publikacji [6] redukcja tej nieprecyzyjności realizowana jest przez formalizację opisu komponentów i wymagań w bazie wiedzy oraz wnioskowanie prowadzące do syntezy struktury aplikacji.

Wyniki obszaru III przedstawiono w publikacji [6].

3.3.1. Formalizacja projektowania jako przetwarzania informacji nieprecyzyjnej

W obszarze badawczym III przyjęto założenie, że redukcja nieprecyzyjności specyfikacji jest możliwa wtedy, gdy proces projektowania zostanie oparty na jawnych modelach formalnych oraz na bazie wiedzy opisującej komponenty systemu i relacje między nimi. Projektowanie aplikacji jest traktowane nie jako ręczne zestawianie modułów, lecz jako proces, w którym na podstawie wymagań oraz założeń architektonicznych konstruuje się formalny opis struktury rozwiązania. Wynikiem tego procesu jest struktura aplikacji opisana jako zestaw warstw, komponentów oraz ich powiązań.

Kluczową rolę pełni jawna reprezentacja wiedzy architektoniczno-komponentowej. Komponenty, ich własności oraz reguły łączenia są opisane w sposób umożliwiający wnioskowanie i kontrolowane przetwarzanie algorytmiczne. W rezultacie nieprecyzyjność wymagań nie musi być rozstrzygana wyłącznie na drodze ręcznych decyzji projektowych, lecz może zostać ujęta formalnie i wykorzystana do skonstruowania spójnej konfiguracji aplikacji. Rozwiązanie to przedstawiono w publikacji [6].

3.3.2. Baza formalna, logika klauzulowa i semantyczna baza wiedzy komponentów

W [6] przyjęto, że przejście od niepełnej specyfikacji do struktury rozwiązania wymaga formalnego opisu zarówno wymagań, jak i dostępnych komponentów oraz reguł ich łączenia.

Projektowanie sprowadzono więc do wnioskowania w oparciu o bazę wiedzy, której część formalna jest zapisana w postaci klauzulowej i dodatkowo wspierana ontologią pojęć dziedzinowych.

Formalną podstawę rozwiązania stanowi klauzulowa logika pierwszego rzędu. W jej ramach zdefiniowano język opisu elementów istotnych dla procesu projektowania obejmujący encje dziedzinowe i relacje między nimi, konstrukty i komponenty, metody oraz problem projektowy rozumiany jako specyfikacja oczekiwanego artefaktu wraz z wymaganymi własnościami.

Na tej podstawie zbudowano bazę wiedzy o komponentach, która opisuje ich znaczenie i zależności oraz umożliwia automatyzację projektowania.

Bazę wiedzy wsparto ontologicznie z wykorzystaniem OWL/RDF, co umożliwia ujednolicenie pojęć dziedzinowych, ich własności oraz zależności pomiędzy komponentami.

Ontologia komponentów pełni funkcję formalnej warstwy pojęciowej projektowania. Wspiera ocenę zgodności semantycznej oraz uzasadnianie doboru komponentów do określonych ról architektonicznych w tworzonym rozwiązaniu.

3.3.3. Algorytmizacja, reguły wnioskowania i dedukcja struktury aplikacji

Przetwarzanie informacji nieprecyzyjnej na etapie projektowania przyjmuje w tym przypadku postać wnioskowania opartego na bazie wiedzy o komponentach. Na podstawie wymagań oraz przyjętych założeń architektonicznych system dobiera i łączy komponenty zgodnie z regułami, zapewniając semantyczną spójność uzyskanej struktury aplikacji webowej. W publikacji [6] wnioskowanie przebiega w sposób semantycznie kontrolowany, zapewniając zgodność z wymaganiami i założeniami architektonicznymi.

W pierwszym kroku dokonywany jest wstępny dobór metod i komponentów powiązanych wspólnymi wejściami i wyjściami. Dopiero w drugim kroku weryfikowane są warunki wstępne i końcowe oraz domykana jest architektura rozwiązania. Taka dedukcja ogranicza przestrzeń przeszukiwania i porządkuje proces konstrukcji rozwiązania. W ten sposób kontrolowany jest koszt obliczeniowy syntezy rozumiany jako ograniczanie liczby rozważanych wariantów w procesie wnioskowania.

Rezultatem dedukcji nie jest dwuwartościowa odpowiedź o spełnieniu warunków, lecz konstrukcja opisująca projekt aplikacji jako konfigurację modułów, klas, metod i ich powiązań. Otrzymany wynik stanowi schemat współdziałania komponentów, który można bezpośrednio zinterpretować w kontekście projektu i wykonania

3.3.4. Wynik systemowy: od nieprecyzyjnej specyfikacji do schematu wykonania modułów

Projektowanie aplikacji webowej określonej klasy jest traktowane jako proces, w którym przyjmuje się referencyjny wariant architektury trójwarstwowej, a następnie formalnie konfiguruje się jej elementy poprzez dobór komponentów zgodnie z wymaganiami. W publikacji [6] architektura ta obejmuje warstwę prezentacji, warstwę logiki biznesowej oraz warstwę dostępu do danych, a dobór komponentów jest porządkowany przez zastosowanie typowych wzorców projektowych w tych warstwach. Dzięki temu nieprecyzyjność specyfikacji jest redukowana przez ograniczenie przestrzeni rozwiązań do konfiguracji zgodnych z przyjętą architekturą oraz przez formalne reguły dopuszczalnych połączeń między komponentami.

Istotnym uzupełnieniem tego podejścia jest przedstawienie go jako procesu tworzenia aplikacji. Użytkownik, korzystając z interfejsu webowego, definiuje model bazy danych, a system generuje bazowy szkielet aplikacji obejmujący operacje CRUD oraz interfejs w postaci REST API. Następnie możliwe jest włączanie gotowych szablonów na różnych poziomach architektury − od poziomu metod po poziom komponentów złożonych z klas. Przewidziano także możliwość tworzenia, przechowywania i ponownego użycia szablonów użytkownika, co wspiera szybkie wytwarzanie rozwiązań w warunkach zmieniających się wymagań.

W publikacji [6] zaproponowano również kierunek dalszego rozwoju, który zakłada trójwymiarową wizualizację procesu tworzenia aplikacji webowych. Wizualizacja ta miałaby wykorzystywać zaproponowany model teoretyczny oraz szablon procesu budowy systemu z reprezentacją webową, wspierając inspekcję kolejnych kroków syntezy oraz interakcję z systemem wspomagającym automatyzację.

3.3.5. Wyniki III obszaru badawczego

Wyniki obszaru badawczego III tworzą spójne ujęcie, w którym nieprecyzyjność specyfikacji projektowej wynikająca z niepełnych i ewoluujących wymagań jest redukowana przez formalizację oraz wnioskowanie prowadzące do syntezy struktury rozwiązania. W publikacji [6] formalizacja obejmuje opis komponentów, ich własności oraz reguły łączenia komponentów zapisane w bazie wiedzy. Dzięki temu dobór i konfiguracja elementów aplikacji nie opierają się wyłącznie na ręcznych decyzjach projektowych, lecz mogą być realizowane w sposób semantycznie kontrolowany.

Kluczowym rezultatem jest domknięcie ścieżki: formalizacja → dedukcja → struktura rozwiązania w obrębie ustalonej klasy architektury aplikacji webowych. Wnioskowanie prowadzi do wyznaczenia struktury aplikacji jako kompozycji modułów, klas i metod,

a jednocześnie jest organizowane w sposób ograniczający przestrzeń poszukiwań. Najpierw wykonywany jest wstępny dobór elementów powiązanych wspólnymi wejściami i wyjściami, a dopiero następnie weryfikowane są warunki wstępne i końcowe oraz domykana jest architektura rozwiązania.

W ujęciu systemowym podejście z [6] przedstawia proces generowania aplikacji na podstawie specyfikacji formalnej. Automatyzacja dotyczy przejścia od modelu danych i wymagań do wygenerowania szkieletu aplikacji oraz jego dalszej rozbudowy o komponenty i szablony zgodnie z regułami zapisanymi w bazie wiedzy. Użytkownik definiuje model bazy danych w interfejsie webowym, a system generuje szkielet aplikacji obejmujący operacje CRUD oraz interfejs REST API. Następnie system automatycznie integruje szablony i komponenty na różnych poziomach architektury. Elastyczność w tym procesie umożliwia łatwą modyfikację aplikacji, dzięki czemu użytkownicy mogą integrować komponenty w zależności od bieżących potrzeb.

W ten sposób wykazano, że formalne modele i metody mogą skracać drogę od niepełnej i niejednoznacznej specyfikacji do struktury rozwiązania możliwej do implementacji, przy zachowaniu kontroli semantycznej. Podejście to ogranicza ryzyko niespójności oraz błędów projektowych, ponieważ automatyzuje część decyzji konstrukcyjnych w oparciu o jawnie zdefiniowane reguły i założenia architektoniczne. W rezultacie generowana struktura pozostaje zgodna z przyjętym stylem architektury i wymaganiami opisanymi w modelu.

Wyniki obszaru badawczego III opublikowano w pracy [6], natomiast ograniczenia rozwiązań oraz kierunki dalszych badań omówiono w Rozdziale 5.

3.4. Obszar IV: Rozszerzenia – inne nośniki nieprecyzyjności

Obszar badawczy IV obejmuje publikacje [7]–[9], w których nieprecyzyjność nie występuje bezpośrednio ani jako nieostrość warunków zapytania, ani jako niepełność specyfikacji przebiegu procesu przetwarzania, lecz ujawnia się w innych formach. Dotyczy to (I) zewnętrznego sygnału kontekstowego o potencjalnym znaczeniu decyzyjnym opartego na informacjach z mediów społecznościowych [7], (II) braków struktury w tekście wymagających rekonstrukcji [8] oraz (III) pośredniej, empirycznej reprezentacji wiedzy w danych uczących związanej z kompresją i destylacją zbioru danych [9].

Wspólnym mianownikiem pozostaje schemat badawczy przyjęty w rozprawie, czyli przejście od formalizacji problemu, przez jego ujęcie algorytmiczne, do empirycznej weryfikacji skuteczności rozwiązania. W trzecim z wymienionych wątków realizowanym w publikacji [9] formalizacja ma charakter optymalizacyjny, wyrażony numerycznie poprzez

funkcję celu. Uzasadnia to uznanie tej publikacji jako granicznej w obrębie omawianego cyklu publikacji.

3.4.1. Nieprecyzyjna informacja zewnętrzna, zależność od sytuacji rynkowej i krótkotrwałość wpływu

W pracy [7] przedstawiono analizę wpływu postów publikowanych przez znane osoby w mediach społecznościowych na kształtowanie się kursu wymiany kryptowalut. Jako ekspertów wybrano osoby rozpoznawalne, które jednocześnie wykazują związek z rynkiem finansowym (w szczególności kryptowalut) lub są powiązane z określoną kryptowalutą. W analizie wykorzystano przypadek Dogecoin oraz wpisy Elona Muska, przedsiębiorcy i osoby publicznej, w serwisie X (Twitter). Punktem wyjścia była obserwacja, że taki sygnał zewnętrzny ma charakter niejednoznaczny i silnie zależy od aktualnej sytuacji rynkowej. Ton wypowiedzi może zostać oceniony jedynie w sposób uproszczony, a skala reakcji rynku bywa krótkotrwała i trudna do odseparowania od innych równoległych czynników.

W publikacji [7] nieprecyzyjność sygnału zewnętrznego została ujęta w formie formalizacji wykorzystywanej w algorytmie ATAPSN (*ang. algorithm for forecasting the cryptocurrency exchange rate taking into account posts on social networks*), który jest narzędziem do uwzględniania wpisów z mediów społecznościowych w prognozowaniu kursu kryptowaluty. Wpływ konkretnego wpisu został opisany przez współczynnik istotności c_j zdefiniowany jako iloczyn dwóch składników. Pierwszym jest dyskretna ocena tonacji wpisu ch_j, która przyjmuje wartość 1 dla wpisu pozytywnego, 0 dla neutralnego oraz -1 dla negatywnego. Drugim składnikiem jest wielkość odchylenia prognozy z poprzedniego kroku k_j liczona jako wartość bezwzględna różnicy między prognozą y_j a wartością rzeczywistą x_j. Tak wyznaczony współczynnik c_j jest następnie używany do modyfikacji prognozy w kolejnym kroku, co tworzy jednoznaczny łańcuch przetwarzania od sygnału zewnętrznego do zmodyfikowanej prognozy.

Weryfikację przeprowadzono empirycznie, porównując prognozy uzyskane metodą ATAPSN z prognozami metod klasycznych, w tym ARIMA (*ang. autoregressive integrated moving average*) oraz wygładzania wykładniczego (*ang. exponential smoothing*).

W publikacji [7] prognozy tych trzech algorytmów zestawiono dla tego samego okresu i tych samych momentów pomiaru, a następnie wykorzystano do zbudowania próbek statystycznych $Y_1(\text{ATAPSN})$, $Y_2(\text{ARIMA})$, $Y_3(\text{WYGŁADZANIE WYKŁADNICZE})$. Ocena jakości prognoz obejmowała miary błędu oraz analizę zależności między wartościami prognozowanymi i rzeczywistymi z wykorzystaniem współczynnika korelacji Spearmana i Pearsona. Ponadto

przeprowadzono test istotności statystycznej za pomocą testu $t-\text{Studenta}$ przy poziomie istotności $\alpha = 0.05$. Wyniki pokazały, że nawet uproszczona formalizacja bodźca zewnętrznego może być skutecznie włączona do procesu prognozowania w sposób algorytmiczny i oceniona ilościowo.

3.4.2. Informacja niepełna strukturalnie: rekonstrukcja interpunkcji i kapitalizacji

Drugi wątek rozszerzeń dotyczy sytuacji, w której tekst jest pozbawiony części informacji o strukturze, na przykład w wyniku automatycznego rozpoznawania mowy (ASR, *ang. automatic speech recognition*) lub specyfiki treści internetowych. W takich danych często brakuje znaków interpunkcyjnych oraz wielkich liter, co obniża czytelność tekstu i utrudnia jego dalszą analizę w zadaniach przetwarzania języka naturalnego (NLP). W tym ujęciu nieprecyzyjność ma postać brakującej struktury językowej, która nie zmienia treści leksykalnej, ale ogranicza poprawną interpretację granic zdań, typów wypowiedzi oraz nazw własnych.

W publikacji [8] zadanie sformalizowano jako klasyfikację tokenów z ustalonym zbiorem etykiet. Każdemu tokenowi przypisywana jest klasa kodująca jednocześnie informację o interpunkcji oraz o tym, czy następne słowo powinno zaczynać się wielką literą. Zdefiniowano 10 klas jako połączenie wariantu interpunkcji z wariantem kapitalizacji. Aby zapewnić, że predykcja dotyczy tylko pozycji mających sens językowy, wprowadzono dwie maski binarne. Pierwsza wskazuje tokeny odpowiadające pełnym słowom albo końcowym fragmentom słów w tokenizacji podwyrazowej (*ang. subword tokenization*). Druga odcina tokeny techniczne używane do wyrównywania długości sekwencji (*ang. padding)*, tak aby nie wpływały na uczenie i wnioskowanie.

Algorytmicznie zaproponowano architekturę hybrydową, w której wielojęzyczny enkoder transformerowy $\text{XLM} - \text{RoBERTa} - \text{base}$ dostarcza reprezentacji kontekstowych, a warstwa sekwencyjna BiLSTM wzmacnia modelowanie zależności lokalnych w ciągu tokenów. Predykcję klas realizuje warstwa liniowa. Uczenie przeprowadzono z użyciem funkcji straty CrossEntropyLoss oraz optymalizatora Adafactor, a proces treningu zorganizowano etapowo. W pierwszej epoce trenowano cały model, natomiast w kolejnych epokach zamrażano enkoder transformerowy i aktualizowano jedynie warstwy sekwencyjne BiLSTM oraz klasyfikator, co ogranicza koszty obliczeniowe i pamięciowe bez utraty jakości.

Weryfikację wykonano na zbiorze IWSLT 2012 (TED Talks) obejmującym ponad 140 tys. sekwencji. Jakość oceniano nie tylko miarami globalnymi, lecz także z zastosowaniem procedury raportowania uwzględniającej niezbalansowanie klas. Zaprezentowano metryki

zarówno z klasą dominującą 0 (brak interpunkcji i mała litera), jak i bez niej, ponieważ jej obecność może zawyżać oceny. Uzupełniająco przedstawiono analizę w podziale na klasy (*ang. per-class*) oraz macierz pomyłek, co pozwala ocenić, które znaki interpunkcyjne i przypadki kapitalizacji są dla modelu najtrudniejsze, a tym samym wprost przekładają się na użyteczność rozwiązania w zastosowaniach praktycznych.

3.4.3. Praca graniczna: kompresja wiedzy empirycznej w danych syntetycznych (formalność optymalizacyjna)

Trzeci wątek w obszarze IV jest graniczny względem wcześniejszych obszarów badawczych, ponieważ dotyczy sytuacji, w której źródłem nieprecyzyjności jest empiryczny charakter danych wykorzystywanych do uczenia modelu. Zbiór obserwacji jest ograniczony, może być zaszumiony i niezbalansowany, a zarazem stanowi przybliżenie rozkładu, do którego model ma się odnieść w procesie uczenia. W takim ujęciu wiedza o zjawisku jest dostępna pośrednio poprzez dane i musi zostać skondensowana do postaci bardziej zwartej, przy zachowaniu skuteczności generalizacji.

W publikacji [9] formalizacja przybiera charakter optymalizacyjny. Celem jest stworzenie niewielkiego syntetycznego zbioru danych, który umożliwia uczenie modeli o jakości porównywalnej z uczeniem na zbiorze pełnym. Zaproponowane w [9] rozwiązanie GDADD (*ang. generative data augmentation by dataset distillation*), oparte na DDMTT (*ang. dataset distillation by matching training trajectories*), realizuje tę ideę poprzez dopasowanie trajektorii procesu uczenia. W tym kontekście DDMTT pozwala na wyznaczenie trajektorii procesu uczenia na danych syntetycznych, które mają zbliżoną charakterystykę do trajektorii uzyskanych na danych rzeczywistych. Dane syntetyczne są optymalizowane w taki sposób, by proces uczenia na nich prowadził do wyników porównywalnych z wynikami uzyskanymi na pełnych zbiorach danych rzeczywistych. W tym sensie formalnie zdefiniowany obiekt optymalizacji, czyli trajektorie uczenia, staje się podstawą algorytmicznej kompresji informacji empirycznej w danych syntetycznych.

Weryfikację przeprowadzono na zbiorach $\mathrm{CIFAR}-10$, $\mathrm{CIFAR}-100$ oraz MedMNIST. Zbiory MedMNIST obejmują różne podzbiory danych medycznych, takie jak PathMNIST, DermaMNIST i RetinaMNIST, które są używane do zadań klasyfikacji obrazów medycznych. Analizowano wpływ liczby IPC (liczba obrazów na klasę*, ang. images per class*) oraz stopnia zbalansowania klas na skuteczność modeli uczonych na danych syntetycznych mierzoną na zbiorze testowym. Wyniki wskazują, że w wybranych konfiguracjach syntetyczny zbiór danych pozwala uzyskać jakość porównywalną z uczeniem na danych rzeczywistych, przy znacząco

mniejszej liczności danych treningowych. Obserwowany efekt zależy od charakterystyki danych oraz doboru IPC.

3.4.4. Wyniki IV obszaru badawczego

Wyniki obszaru badawczego IV potwierdzają, że schemat przyjęty w rozprawie, czyli formalizacja problemu, ujęcie algorytmiczne oraz weryfikacja empiryczna, pozostaje zasadny także wtedy, gdy nośnik nieprecyzyjności różni się od tego występującego w obszarach I–III. W pracach [7]–[9] nieprecyzyjność nie wynika bezpośrednio ani z nieostrości warunków zapytania, ani z niepełnej specyfikacji przebiegu procesu przetwarzania, lecz ujawnia się kolejno jako (I) sygnał zewnętrzny o krótkotrwałym wpływie, (II) brak struktury w tekście oraz pośrednia, (III) empiryczna postać wiedzy zawartej w zbiorach uczących.

W publikacji [7] wykazano, że kontekstowy sygnał zewnętrzny może zostać sformalizowany na tyle, aby dało się go włączyć do algorytmu prognozowania i ocenić jego znaczenie ilościowo. W algorytmie ATAPSN wpływ postu jest modelowany współczynnikiem zależnym od dyskretnej oceny tonacji oraz składnika związanego z historyczną trafnością prognoz, a następnie wykorzystywany do korekty wartości prognozy. Weryfikacja na danych Dogecoin (kursy z Binance) obejmowała porównanie z metodami klasycznymi (ARIMA i wygładzanie wykładnicze) z użyciem średniego błędu względnego oraz miar zależności Spearmana i Pearsona wraz z testem istotności współczynników korelacji przy poziomie $\alpha = 0.05$. Istotność statystyczną potwierdzono jedynie dla zależności między wartościami rzeczywistymi i prognozami uzyskanymi metodą ATAPSN, natomiast dla ARIMA i wygładzania wykładniczego nie potwierdzono istotności przy tym samym poziomie α. Jednocześnie średni błąd względny wyniósł ok. 0.23% dla ATAPSN, wobec ok. 3.28% dla ARIMA i ok. 3.14% dla wygładzania wykładniczego.

W publikacji [8] nośnikiem nieprecyzyjności jest niepełność struktury tekstu. Brak interpunkcji i kapitalizacji potraktowano jako brakującą informację strukturalną, którą należy odtworzyć algorytmicznie. Zadanie to zostało sformalizowane jako klasyfikacja sekwencji tokenów, gdzie uczenie i ocena obejmują tylko te tokeny, które odpowiadają rzeczywistym słowom. Aby to zapewnić, zastosowano maski eliminujące tokeny techniczne, takie jak tokeny wypełniające (padding) oraz znaczniki specjalne, dzięki czemu raportowane miary odnoszą się tylko do tych fragmentów tekstu, dla których rekonstrukcja ma sens semantyczny.

Proponowane rozwiązanie ma postać architektury hybrydowej $\mathrm{XLM-RoBERTa-LSTM}$, która łączy reprezentacje kontekstowe uzyskane przez enkoder transformerowy z modelowaniem zależności sekwencyjnych w warstwie rekurencyjnej. Skuteczność oceniono

w dwóch ujęciach. Pierwsze obejmuje metryki globalne, które zostały raportowane w dwóch wariantach − bez klasy dominującej „brak interpunkcji” oraz z jej uwzględnieniem. Drugie ujęcie przedstawia wyniki per – klasa dla znaków interpunkcyjnych, co jest szczególnie istotne w praktyce, ponieważ klasy rzadkie w największym stopniu wpływają na jakość rekonstrukcji, a zarazem są niekiedy maskowane przez wyniki globalne zdominowane przez klasę najliczniejszą.

Porównanie per – klasa dla pięciu znaków interpunkcyjnych przedstawiono w Tabeli 3.4.1. Zestawienie pokazuje, że model z [8] osiąga wysokie wartości F1 nie tylko dla klas częstych, lecz także dla klasy myślnika, dla której uzyskano F1 = 0.831. W części prac porównawczych wynik dla tej klasy jest znacząco niższy (np. 0.425) albo w ogóle nie jest raportowany, co utrudnia porównanie jakości rekonstrukcji dla rzadszych znaków.

Tabela 3.4.1. Porównanie F1 dla poszczególnych klas interpunkcji

Model	**F1-score**				
	_ (brak znaku)	**, (przecinek)**	**. (kropka)**	**? (znak zapytania)**	**– (myślnik)**
[116]	0.991	0.819	0.948	0.890	0.425
[117]	0.990	0.760	0.880	0.820	–
[84]	–	0.678	0.840	0.876	–
[118]	–	0.625	0.373	0.268	–
[119]	–	0.809	0.938	0.846	–
[8]	**0.987**	**0.896**	**0.946**	**0.863**	**0.831**

Uzupełnieniem analizy jest zestawienie metryk globalnych w wariancie bez klasy 0 oraz z klasą 0, gdzie klasa 0 odpowiada przypadkowi „brak interpunkcji”. Wyniki tego porównania przedstawiono w Tabeli 3.4.2. Dla modelu XLM – RoBERTa– LSTM uzyskano odpowiednio (bez klasy 0 / z klasą 0): Accuracy 0.923 / 0.959, Precision 0.906 / 0.927, Recall 0.899 / 0.919, F1 0.901 / 0.923. Zestawienie pokazuje, że uwzględnienie klasy dominującej podnosi miary globalne. W związku z tym wariant bez klasy 0 lepiej odzwierciedla jakość rekonstrukcji znaków interpunkcyjnych, ponieważ ogranicza wpływ dominującej klasy braku interpunkcji na miary globalne.

Tabela 3.4.2. Porównanie metryk globalnych bez klasy i z klasą 0 „brak interpunkcji”

Model	Accuracy (bez klasy 0 / z klasą 0)	Precision (bez klasy 0 / z klasą 0)	Recall (bez klasy 0 / z klasą 0)	F1-score (bez klasy 0 / z klasą 0)
[116]	–	–	–	– / 0.94
[117]	– / 0.97	0.813 / 0.858	0.83 / 0.87	0.82 / 0.863
[84]	–	0.758 / –	0.851 / –	0.798 / –
[118]	–	0.607 / 0.661	0.341 / 0.426	0.422 / 0.518
[119]	–	0.887 / 0.912	0.844 / 0.88	0.864 / 0.894
[8]	**0.923 / 0.959**	**0.906 / 0.927**	**0.899 / 0.919**	**0.901 / 0.923**

Publikacja [9] pełni rolę pracy granicznej, ponieważ formalizacja nie przyjmuje tu postaci reguł lub semantyki logicznej, lecz ma charakter optymalizacyjny i jest wyrażona numerycznie przez funkcję celu. W tym ujęciu nieprecyzyjność wynika z faktu, że wiedza o zjawisku jest dostępna pośrednio jako skończony zbiór danych uczących, który bywa zaszumiony i obciążony, a jednocześnie jest kosztowny w przechowywaniu i w użyciu obliczeniowym. W pracy pokazano, że „domknięcie” łańcucha: formalizacja → algorytm → weryfikacja jest możliwe również wtedy, gdy formalizacja oznacza precyzyjnie zdefiniowany cel optymalizacji prowadzący do destylacji zbioru danych i generatywnej augmentacji.

Weryfikację empiryczną w [9] przeprowadzono na zbiorach zbalansowanych ($\mathrm{CIFAR}-10$, $\mathrm{CIFAR}-100$) oraz na zbiorach o nierównowadze klas z kolekcji MedMNIST. W publikacji podano maksymalną dokładność walidacyjną (średnia ± odchylenie standardowe) uzyskaną w badanych ustawieniach i zestawiono ją z wartościami referencyjnymi odpowiadającymi uczeniu na pełnym oryginalnym zbiorze. Dla danych zbalansowanych destylacja połączona z generatywną augmentacją (GDADD), oparta na destylacji przez dopasowanie trajektorii uczenia (DDMTT), uzyskała wyniki wyższe niż uczenie na pełnym zbiorze, co przedstawiono w Tabeli 3.4.3. Dla $\mathrm{CIFAR}-10$ uzyskano 0.879 ± 0.001 wobec 0.848 ± 0.001, a dla $\mathrm{CIFAR}-100$ uzyskano 0.591 ± 0.003 wobec 0.562 ± 0.003, czyli odpowiednio ok. 3.1 p.p. i 2.9 p.p. przewagi na korzyść rozwiązania z [9] przy jednoczesnym użyciu mniejszej reprezentacji syntetycznej.

Tabela 3.4.3. Maksymalna dokładność walidacyjna dla zbalansowanych zbiorów CIFAR

Eksperyment	CIFAR-10	CIFAR-100
DD - pełny, oryginalny zbiór danych [93]	0.848 ± 0.001	0.562 ± 0.003
GDADD - mniejszy zbiór zdestylowany (syntetyczny) [9]	**0.879 ± 0.001**	**0.591 ± 0.003**

Uzupełnieniem tej oceny jest analiza zbiorów medycznych o różnym stopniu niezbalansowania klas, ujęta w Tabeli 3.4.4. Dla PathMNIST (lekka nierównowaga) metoda

z [9] uzyskała 0.917 ± 0.003, przewyższając wartość referencyjną w zestawieniu (ResNet-50: 0.911). Jednocześnie dla silnie niezbalansowanych zbiorów DermaMNIST i RetinaMNIST wyniki są wyraźnie niższe od metod referencyjnych, w tym od najwyższych wartości referencyjnych w tabeli. Ten kontrast wskazuje, że skuteczność GDADD jest wyraźnie zależna od struktury danych: przy lekkiej nierównowadze klas wyniki pozostają konkurencyjne, natomiast przy silnym niezbalansowaniu obserwuje się istotny spadek jakości. Ponieważ w obecnej konfiguracji metody niezbalansowanie klas nie jest uwzględniane jawnie w funkcji celu ani w procedurze optymalizacji, destylacja może w większym stopniu odtwarzać regularności klas dominujących kosztem klas rzadkich, co pogarsza wynik. W rezultacie [9] dopełnia schemat: formalizacja → algorytm → weryfikacja także dla formalizacji optymalizacyjnej, jednocześnie sugerując, że w warunkach silnej nierównowagi klas potrzebne są dodatkowe mechanizmy uwzględniające tę własność danych w funkcji celu lub procedurze optymalizacji.

Tabela 3.4.4. Maksymalna dokładność walidacyjna zbiorów zbalansowanych i niezbalansowanych (PathMNIST: lekka nierównowaga klas; DermaMNIST i RetinaMNIST: silna nierównowaga)

Model / podejście	Zbalansowane (PathMNIST)	Niezbalansowane (DermaMNIST)	Niezbalansowane (RetinaMNIST)
ResNet-18	0.907	0.735	0.524
ResNet-50	0.911	0.735	0.528
auto-sklearn	0.716	0.719	0.515
AutoKeras	0.834	0.749	0.503
Google AutoML Vision	0.728	0.768	0.531
GDADD [9]	**0.917 ± 0.003**	**0.655 ± 0.009**	**0.440 ± 0.020**

Znaczenie publikacji [9] dla obszaru IV polega na tym, że rozszerza on rozumienie nieprecyzyjności na warunki typowe dla informatyki XXI wieku. Duże wolumeny danych nie eliminują niepewności, lecz często ujawniają ją w postaci szumu, nierównowagi klas i obciążeń systematycznych, a jednocześnie generują koszt obliczeniowy uczenia. W takim kontekście *nieprecyzyjność* dotyczy nie tylko interpretacji informacji, lecz także tego, w jaki sposób informacja empiryczna jest kodowana w danych uczących i może być skondensowana do postaci syntetycznej. Włączenie formalizacji optymalizacyjnej pokazuje, że schemat przyjęty w rozprawie nie ogranicza się do podejść symbolicznych, lecz obejmuje również przypadki, w których formalność ma postać jednoznacznie zdefiniowanej funkcji celu, a skuteczność musi zostać potwierdzona eksperymentalnie.

Podsumowując, obszar IV pełni rolę weryfikacyjną dla tezy rozprawy. Potwierdza, że przejście od formalizacji, przez algorytm, do weryfikacji empirycznej pozostaje spójne również wtedy, gdy nieprecyzyjność ma postać sygnału kontekstowego, braków struktury tekstu lub pośredniej, empirycznej reprezentacji wiedzy w danych uczących.

Wyniki obszaru badawczego IV opublikowano w pracach [7]–[9], a ograniczenia przedstawionych rozwiązań oraz kierunki dalszych badań omówiono w Rozdziale 5.

3.5. Podsumowanie części syntetycznej

W Rozdziale 3 przedstawiono syntetyczny opis wyników badań zaprezentowanych w publikacjach [1]–[9], które składają się na cykl w układzie czterech obszarów badawczych odpowiadających kolejnym poziomom ujawniania się informacji nieprecyzyjnej w systemach informatycznych. Obszar I obejmuje nieostrość w danych i w zapytaniach relacyjnych oraz metody jej przetwarzania poprzez formalną semantykę stopniową i transformację zapytań rozmytych do postaci wykonywalnej w standardowym SQL. Obszar II dotyczy sytuacji, w której użytkownik określa cel obliczeniowy, natomiast system, na podstawie formalnych opisów usług, automatycznie konstruuje plan wykonania prowadzący do uzyskania oczekiwanego rezultatu. Obszar III przenosi ciężar rozważań na etap projektowania systemu, wskazując rolę formalnych modeli w ujednoznacznianiu niepełnych i ewoluujących wymagań oraz w skracaniu drogi od specyfikacji do spójnej struktury aplikacji możliwej do implementacji. Obszar IV pokazuje, że ten sam schemat badawczy można zastosować także do innych nośników nieprecyzyjności: sygnałów zewnętrznych o ograniczonej trwałości i kontekstowej interpretacji, braków struktury w tekście wymagających rekonstrukcji oraz pośredniej, empirycznej reprezentacji wiedzy w danych syntetycznych.

Przyjęty układ rozdziału pozwala odróżnić dwie perspektywy. Pierwsza ma charakter syntetyczny i pokazuje spójność formalno-algorytmicznego ujęcia informacji nieprecyzyjnej oraz jego realizację w kolejnych obszarach badawczych. Druga ma charakter źródłowy, ponieważ pełne wyniki, szczegóły metod oraz procedury eksperymentalne przedstawiono w publikacjach stanowiących podstawę rozprawy. Z tego też względu w Rozdziale czwartym zamieszczono pełne teksty publikacji wchodzących w skład cyklu.

4. Teksty publikacji stanowiących podstawę rozprawy

Rozdział 4 zawiera pełne teksty publikacji wchodzących w skład cyklu i stanowiących podstawę niniejszej rozprawy doktorskiej. Zamieszczenie publikacji w całości umożliwia weryfikację wyników syntetyzowanych przedstawionych w Rozdziale 3 oraz zapoznanie się ze szczegółami formalizacji, algorytmów i procedur eksperymentalnych w oryginalnym kontekście badawczym. Wykaz publikacji wraz z zakresem stron, na których znajdują się ich pełne treści w niniejszym dokumencie, przedstawiono poniżej.

[1] **Nowakowski G.**, *Fuzzy queries on relational databases*, s. 58–64.

[2] **Nowakowski G.**, *Methodology of transformation of fuzzy queries into queries in the SQL standard*, s. 65–70.

[3] Telenyk S., **Nowakowski G.**, Yefremov K., Khmeliuk V., *Logics based application integration for interdisciplinary scientific investigations*, s. 71–76.

[4] **Nowakowski G.**, Telenyk S., Yefremov K., Khmeliuk V., *The approach to applications integration for World Data Center interdisciplinary scientific investigations*, s. 77–83.

[5] **Nowakowski G.**, Telenyk S., Yefremov K., Khmeliuk V., *Simple and flexible way to integrate heterogeneous information systems and their services into the world data system*, s. 84–98.

[6] Telenyk S., **Nowakowski G.**, Zharikov E., Vovk J., *Conceptual foundations of the use of formal models and methods for the rapid creation of web applications*, s. 99–105.

[7] Telenyk S., **Nowakowski G.**, Gavrilenko O., Miahkyi M., Khalus O., *An analysis of the influence of famous people's posts on social networks on the cryptocurrency exchange rate*, s. 106–114.

[8] Shymkovych V., **Nowakowski G.**, Telenyk S., *Joint Punctuation Restoration and Text Capitalisation with a Hybrid XLM-RoBERTa–LSTM Model*, s. 115–122.

[9] Gordienko Y., **Nowakowski G.**, Kochura Y., Taran V., Stirenko S., *Generative Data Augmentation by Dataset Distillation*, s. 123–136.

Fuzzy queries on relational databases

Grzegorz Nowakowski
Cracow University of Technology
Department of Automatic Control and Information Technology
Faculty of Electrical and Computer Engineering
Cracow, Poland
gnowakowski@pk.edu.pl

***Abstract*—This article presents various forms of fuzzy queries, a detailed analysis of these queries and their conversion into standard SQL queries by means of Oracle 11g XE. A qualitative and quantitative study about the use of fuzzy queries on relational databases has been included in this article, as well. Given that fuzzy queries arrange results according to the degree in which they meet the conditions of the query, it is easier to analyze the results and the risk of obtaining an empty result is reduced thanks to an extended interpretation of the conditions of the query. The provided examples of conversion of fuzzy queries into standard SQL queries by means of Oracle 11g XE point to easy to implement methods of obtaining fuzzy information from the database, and thereby expand its functionality.**

Keywords—fuzzy logic, fuzzy queries, relational databases, SQL, Oracle 11g XE

I. INTRODUCTION

Classic query languages (with SQL being the most widely used) define the scope of data as well as conditions that must be met by data throughout the entire process of searching for information in databases. These conditions should be precisely defined, as precision is the basic requirement when defining conditions in query languages. At the same time, these conditions impose limitations. For example, a customer of a car dealership who is looking for a cheap car has to precisely define the range of the price he is willing to pay. Regardless of how the limits of the range are defined, a car priced slightly above the set limit will not meet the conditions of the query. This shows that the limits in question result from the necessity of precisely defining the conditions, which were initially expressed in a natural language and with imprecise terms. [1]

The problem can be solved by the use of linguistic terms modelled and processed with fuzzy logic in queries addressed to a database. These are so-called fuzzy queries. Linguistic terms are presented as fuzzy sets in an appropriate, usually numerical, space. Therefore, matching data to the query is no longer perceived as dichotomy: matched – unmatched. The notion of a degree of matching data to the query is thus introduced, and one assumes that the value of this degree of matching corresponds (in a somewhat simplified view) to the degree in which the data belongs to the fuzzy set representing the condition of the query. [1, 2, 3]

This article discusses various forms of fuzzy queries, and presents a detailed analysis of fuzzy queries as well as documents conversion of them into standard SQL queries by means of Oracle 11g XE A qualitative and quantitative study about the use of fuzzy queries on relational databases has been included in this article, as well. The provided examples of conversion of fuzzy queries into standard SQL queries by means of Oracle 11g XE point to easy to implement methods of obtaining fuzzy information from the database and thereby expand its functionality.

II. FUZZY QUERIES ON RELATIONAL DATABASES

A. The concept of a fuzzy query

A fuzzy query addressed to a relational database [1, 2, 3] is a query which includes overtly used expressions of the natural language, referred to as linguistic terms (modelled with fuzzy logic), defining the following: imprecise values, imprecise comparisons and non-standard methods of aggregation of degrees of meeting partial conditions of the query.

Particular rows meet the conditions of such a query to a certain degree expressed with a number in range <0,1> where 1 means that the query requirement is fully met, and 0 means that the query requirement is fully unmet. Therefore, the result of the query is a set of rows arranged according to the degree of meeting the conditions of the query. This way, it is easier to reflect the attempts undertaken by a given individual to match the data to the query. In case of complex queries, which include both fuzzy linguistic terms and non-standard aggregation schemes, meeting conditions of the query is gradual. In such a situation, a human being naturally evaluates the data and perceives it as data which meets his/her requirements to a greater or a lesser degree instead of distinguishing only between the data which meets or does not meet the set requirements. The notion of the degree of meeting conditions of the query allows us to formally present these complex queries and thereby naturally arrange the results in such a manner, as to make sure the data which meets the conditions of the query to the greatest degree is at the beginning of the results list. [1]

B. Conversion of fuzzy queries into standard SQL queries

Construction of fuzzy queries as well as their execution and the applied grammar of fuzzy queries are usually strongly connected with the query language of a given database. This item discusses the problem of conversion of fuzzy queries into standard SQL queries by means of an Oracle 11g XE relational database, in which SQL query language is used.

978-1-5386-6143-7/18/$31.00 ©2018 IEEE

There are numerous fuzzy elements in queries addressed to databases:

- atomic predicates based on linguistic terms: young, tall
- atomic predicates based on similarity of linguistic terms: information technology ≈ artificial intelligence
- complex predicates: fuzzy sum and product operators
- modified predicates: very, around, rather, antonyms
- fuzzy operators: approximately, a little more, etc.
- linguistic quantifiers
- fuzzy combination of relations
- fuzzy aggregate functions
- grouping by fuzzy values

In order to discuss the approach towards the representation of imprecision, examples of imprecise elements in queries addressed to database contained in the last two items of the above list were taken into account. The first three items are presented in [1].

Two approaches towards the representation of imprecision can be distinguished: based on distribution of possibilities, and based on similarity.

From the above reasoning, the whole interpretation of fuzzy queries comes down to the following general rule: the degree of meeting conditions of the query equates with the value of the membership function of the relevant fuzzy set. In case of complex queries, degrees of meeting partial conditions of the query calculated in the above-mentioned manner, corresponding to particular conditions contained in the query, are aggregated with selected operators. [1]

1) Exemplary data. Assuming there are already exemplary tables with data in Oracle 10g XE database, another column labelled *fuzzy_degree*, with the initially set value of 1.0 for each row, was added to the table labelled *tbl_cars* presented in Fig. 1. This denotes the initial state - total degree of membership of a given row in the table labelled *tbl_cars*.

```
SELECT * FROM tbl_cars;
+--------+-------------+---------+-----------+-----------------+---------+--------------+
| ID_SAM | MODEL       | TYPE    | PROD_YEAR | ENGINE_CAPACITY | PRICE   | FUZZY_DEGREE |
+--------+-------------+---------+-----------+-----------------+---------+--------------+
|      1 | VOLVO       | SEDAN   |      2010 |             1.6 |  35500  |  1.000000000 |
|      2 | ASTON MARTIN| COUPE   |      2015 |             6.0 | 900000  |  1.000000000 |
|      3 | MAZDA       | CABRIO  |      2012 |             2.0 |  43990  |  1.000000000 |
|      4 | OPEL        | COMBI   |      1993 |             1.7 |   1500  |  1.000000000 |
|      5 | VOLKSWAGEN  | SEDAN   |      2007 |             1.9 |  23900  |  1.000000000 |
|      6 | AUDI        | COUPE   |      2005 |             1.9 |  21000  |  1.000000000 |
+--------+-------------+---------+-----------+-----------------+---------+--------------+
```

Fig. 1. Exemplary data from *tbl_cars* table

Table *tbl_possibility* contains data which will be used at the stage of conversion of fuzzy queries into typical SQL queries.

The *tbl_possibility* table, as presented in Fig. 2, stores linguistic variables described by the membership function:

- column *linguistic_variable_value* stores the value of the linguistic variable,
- column *type* contains information on the type of the membership function describing the linguistic variable (in the table in question all variables are described by the trapezoidal membership function),
- columns *a, b, c* and *d* contain parameters of the function,
- columns *table_name* and *column_name* store the following information: on which column and from which table the linguistic variable was presumed.
- column *username* contains information on the name of the user to which a given definition applies; this way various users can describe linguistic variables with identical names, presumed for the same columns of the same tables, in a different way. [1]

```
SELECT * FROM tbl_possibility;
+----------+------------+-----------------+---------------------------+------+------+------+------+-----------+
| username | table_name | column_name     | linguistic_variable_value | a    | b    | c    | d    | type      |
+----------+------------+-----------------+---------------------------+------+------+------+------+-----------+
| user     | tbl_cars   | prod_year       | NEW                       | 2010 | 2015 | 2050 | 2050 | trapezoid |
| user     | tbl_cars   | prod_year       | AVERAGE                   | 1995 | 2000 | 2010 | 2015 | trapezoid |
| user     | tbl_cars   | prod_year       | OLD                       |    0 |    0 | 1995 | 2000 | trapezoid |
| user     | tbl_cars   | engine_capacity | SMALL                     |    0 | 0.33 | 0.66 |    1 | trapezoid |
| user     | tbl_cars   | engine_capacity | MIDDLE                    |    0 |    1 |  1.6 |  2.5 | trapezoid |
| user     | tbl_cars   | engine_capacity | HIGH                      |  1.6 |  2.5 |    5 |    8 | trapezoid |
+----------+------------+-----------------+---------------------------+------+------+------+------+-----------+
```

Fig. 2. Exemplary data from *tbl_possibility* table

2) Fuzzy queries – FUZZY COUNT. Fuzzy queries, similarly to classic SQL queries, can contain group functions and the GROUP BY clause. Group functions must be redefined. In a classic SQL query, the query condition can be either met or unmet. Each result row meets the conditions of the query to the same degree. In fuzzy relations, a row can appear partially, which can be interpreted as certainty or a degree of meeting conditions of the query.

In order to count the number of instances of a given expression in classic SQL, the *COUNT(expression)* group function is used. Considering that it has already been mentioned above, the *COUNT(expression)* has been defined as *SUM(FUZZY_DEGREE*) in fuzzy query.

TABLE I. EXAMPLE: DISPLAY *NUMBER* OF CARS THAT HAVE *SMALL*, *MIDDLE* AND *HIGH* ENGINE CAPACITY

Fuzzy query (conventional form)
SELECT engine_capacity, FUZZY_COUNT(engine_capacity) **FROM** tbl_cars **GROUP BY** engine_capacity;
An equivalent query in SQL - Oracle 11g XE
SELECT linguistic_variable_value, ROUND(**SUM(FP_TRAPEZOID(engine_capacity,a,b,c,d))**,2) FUZZY_COUNT **FROM** tbl_cars, tbl_possibility **WHERE** FP_TRAPEZOID(engine_capacity,a,b,c,d)>0 and (column_name='engine_capacity') **GROUP BY** linguistic_variable_value;

A fuzzy query converted to an equivalent standard SQL query by means of Oracle 11g XE, as presented in Table I, will display the *number* of cars that have *small*, *middle*, *high* engine capacity.

In the discussed example, all variables are defined by the trapezoidal membership function, the definition of which (presented in Fig. 3) was formulated in PL/SQL in Oracle 10g XE database.

```
create or replace FUNCTION
FP_TRAPEZOID(x IN NUMBER, a IN NUMBER, b IN NUMBER, c IN NUMBER, d IN NUMBER)
RETURN NUMBER IS
y number(6,5);
BEGIN
IF (x <= a) THEN y:= 0.0;
ELSIF (x > a AND x <= b) THEN y:= (x - a)/(b - a);
ELSIF (x > b AND x <= c) THEN y:= 1.0;
ELSIF (x > c AND x <= d) THEN y:= (d - x)/(d - c);
ELSE y:= 0.0;
END IF;
  RETURN y;
END FP_TRAPEZOID;
```

Fig. 3. Trapezoidal membership function defined in PL/SQL in Oracle 10g XE

The value of the column labelled FUZZY_COUNT and its results were presented in Fig. 4, were set as SUM(FUZZY_DEGREE). Its definition can be found in Table I.

```
SELECT linguistic_variable_value, ROUND(
       SUM(FP_TRAPEZOID(engine_capacity,a,b,c,d)),2) FUZZY_COUNT
FROM tbl_cars, tbl_possibility
WHERE FP_TRAPEZOID(engine_capacity,a,b,c,d)>0 and
(column_name='engine_capacity')
GROUP BY linguistic_variable_value;
```

```
+---------------------------+---------------+
| LINGUISTIC_VARIABLE_VALUE | FUZZY_COUNT   |
+---------------------------+---------------+
| MIDDLE                    | 3.78          |
| HIGH                      | 1.89          |
+---------------------------+---------------+
```

Fig. 4. Result of the fuzzy query from Table I - FUZZY COUNT

In Fig. 5, the column labelled *fuzzy_degree* contains the information on the degree to which a given row meets conditions of the query. In this case, it informs us as to what extent the engine capacity is *small, middle, high*.

```
SELECT c.*, linguistic_variable_value, FP_TRAPEZOID(engine_capacity,a,b,c,d) FUZZY_DEGREE
FROM tbl_cars c, tbl_possibility
WHERE FP_TRAPEZOID(engine_capacity,a,b,c,d)>0 and (column_name='engine_capacity');
```

```
+-------+--------------+---------+-----------+-----------------+--------+---------------------------+--------------+
| ID_SAM | MODEL        | TYPE    | PROD_YEAR | ENGINE_CAPACITY | PRICE  | LINGUISTIC_VARIABLE_VALUE | FUZZY_DEGREE |
+-------+--------------+---------+-----------+-----------------+--------+---------------------------+--------------+
|      1 | VOLVO        | SEDAN   |      2010 |             1.6 |  35500 | MIDDLE                    | 1.00000      |
|      2 | ASTON MARTIN | COUPE   |      2015 |             6.0 | 900000 | HIGH                      | 0.66667      |
|      3 | MAZDA        | CABRIO  |      2012 |             2.0 |  43990 | MIDDLE                    | 0.55556      |
|      3 | MAZDA        | CABRIO  |      2012 |             2.0 |  43990 | HIGH                      | 0.44444      |
|      4 | OPEL         | COMBI   |      1993 |             1.7 |   1500 | MIDDLE                    | 0.88889      |
|      4 | OPEL         | COMBI   |      1993 |             1.7 |   1500 | HIGH                      | 0.11111      |
|      5 | VOLKSWAGEN   | SEDAN   |      2007 |             1.9 |  23900 | MIDDLE                    | 0.66667      |
|      5 | VOLKSWAGEN   | SEDAN   |      2007 |             1.9 |  23900 | HIGH                      | 0.33333      |
|      6 | AUDI         | COUPE   |      2005 |             1.9 |  21000 | MIDDLE                    | 0.66667      |
|      6 | AUDI         | COUPE   |      2005 |             1.9 |  21000 | HIGH                      | 0.33333      |
+-------+--------------+---------+-----------+-----------------+--------+---------------------------+--------------+
```

Fig. 5. Result of the fuzzy query before grouping by fuzzy values

3) Fuzzy queries – FUZZY AVG. Group functions work on row sets, called groups. Rows belong to the same group if they have the same value for the grouping expression. The group function calculates a single value for each group based on the expression that is its parameter.

In order to count average value of a given expression in classic SQL, the *AVG(expression)* group function is used. In fuzzy query the *AVG(expression)* has been defined as *SUM(FUZZY_DEGREE * (VALUE from COLUMN_NAME)) / SUM(FUZZY_DEGREE).*

TABLE II. EXAMPLE: DISPLAY *AVERAGE* PRICE OF A CAR DEPENDING ON THE GROUP TO WHICH IT BELONGS: *NEW* CARS, *AVERAGE* CARS, *OLD* CARS

Fuzzy query (conventional form)
SELECT prod_year, FUZZY_AVG(prod_year) **FROM** tbl_cars **GROUP BY** prod_year;

An equivalent query in SQL - Oracle 11g XE
SELECT linguistic_variable_value, ROUND(**SUM(FP_TRAPEZOID(prod_year,a,b,c,d)*price) /** **SUM(FP_TRAPEZOID(prod_year,a,b,c,d)),**2) FUZZY_AVG **FROM** tbl_cars, tbl_possibility **WHERE** FP_TRAPEZOID(prod_year,a,b,c,d)>0 and (column_name='prod_year') **GROUP BY** linguistic_variable_value;

A fuzzy query converted to an equivalent standard SQL query by means of Oracle 11g XE, as presented in Table II, will display *average* price of a car depending on the group to which it belongs: *new* cars, *average* cars, *old* cars. All variables are defined by the trapezoidal membership function, the definition of which (presented in Fig. 3) was formulated in PL/SQL in Oracle 10g XE database. In the query in question (the result of which was presented in Fig. 6):

- a defined fuzzy set corresponding to the term *new* for the attribute *prod_year* was applied,
- a defined fuzzy set corresponding to the term *average* for the attribute *prod_year* was applied,
- a defined fuzzy set corresponding to the term *high* for the attribute *prod_year* was applied,
- group function FUZZY_AVG was defined as SUM(FP_TRAPEZOID(prod_year,a,b,c,d)*price) / SUM(FP_TRAPEZOID(prod_year,a,b,c,d))

```
SELECT linguistic_variable_value, ROUND(
       SUM(FP_TRAPEZOID(prod_year,a,b,c,d)*price ) /
       SUM(FP_TRAPEZOID(prod_year,a,b,c,d)),2)
       FUZZY_AVG
FROM tbl_cars, tbl_possibility
WHERE FP_TRAPEZOID(prod_year,a,b,c,d)>0 and (column_name='prod_year')
GROUP BY linguistic_variable_value;
```

```
+---------------------------+---------------+
| LINGUISTIC_VARIABLE_VALUE | FUZZY_AVG     |
+---------------------------+---------------+
| AVERAGE                   | 29665         |
| NEW                       | 655425.71     |
| OLD                       | 1500          |
+---------------------------+---------------+
```

Fig. 6. Result of the fuzzy query from Table II - FUZZY COUNT

4) Fuzzy queries – combining fuzzy conditions. Fuzzy queries, similarly to classic SQL queries, can contain complex conditions resulting from combining single conditions with logical operators. In SQL, the keyword AND corresponds to the conjunction operator, the keyword OR corresponds to the alternative operator and the keyword NOT to the negation operator. [1]

Identical keywords are applied in case of fuzzy queries, but they correspond, respectively, to: fuzzy conjunction, fuzzy alternative and fuzzy negation.

Nevertheless, it needs pointing out that in the fuzzy sets theory [7], there is a number of operators carrying out intersection operation (product operation which corresponds to the logical operation AND). These are applied interchangeably depending on the problem at hand. Most of these operators meet the criteria of the so-called triangular norm T (T-norm):

$$\mu_{A\cap B}(x) = T(\mu_A(x), \mu_B(x)) \quad (1)$$

Similarly to the execution of the intersection operation, numerous operators are used for the purpose of the operation of joining (logical sum which corresponds to the logical operation OR). The most commonly applied operators meet the criteria of the so-called triangular norm S (S-norm) also referred to as T-conorm:

$$\mu_{AUB}(x) = S(\mu_A(x), \mu_B(x)) \quad (2)$$

The criteria defining triangular norms consist of four fundamental conditions, and will not be presented here due to space constraints. These conditions were described in [2, 3, 8, 9], among others.

The most commonly used T-norms (mapping logical operator AND) are the minimum MIN (a, b) and the product (PROD) $a \cdot b$, whereas the most commonly used S-norm operators (mapping logical operator OR) are the maximum MAX (a, b) and the so-called algebraic (probabilistic) *sum* $a + b - a \cdot b$.

A relevant S-norm corresponds to each T-norm, provided that the following condition is met:

$$T(a,b) = 1\text{-}S(1\text{-}a,1\text{-}b) \text{ or } S(a,b) = 1\text{-}T(1\text{-}a,1\text{-}b) \quad (3)$$

Operators which meet the condition (3) form the so-called complementary (conjugate, dual) pairs. Numerous operators meeting the conditions of T-norms and S-norms have been developed and described. These operators are divided into non-adjustable, with a constant mode of operation, and adjustable (parametrised), also referred to as families of triangle norms, in the case of which the mode of operation changes depending on the accepted parameter (the degree of freedom) for the operator in question. In [1] selected triangle norms forming complementary pairs have been presented.

Operators of triangle norms execute operations only on two fuzzy (variable) sets. Operations on a greater number of sets can be executed gradually, by combining sets into pairs with the sequence of combining sets into pairs having no effect on the result (coherency quality).

From the above reasoning, the degree of meeting a complex condition is calculated based on the degrees of meeting partial conditions as well as selected T-norm and its complementary S-norm.

A fuzzy query converted to an equivalent standard SQL query by means of Oracle 11g XE, as presented in Table III (first version) or Table IV (second version), will display the offer of *new* cars with *high* engine cubic capacity. All variables are defined by the trapezoidal membership function, the definition of which (presented in Fig. 3) was formulated in PL/SQL in Oracle 10g XE database. In the query in question:

- a defined fuzzy set corresponding to the term *new* for the attribute *prod_year* was applied,
- a defined fuzzy set corresponding to the term *high* for the attribute *engine_capacity* was applied,
- a conjunction of both of the above-mentioned fuzzy conditions was carried out, which resulted in defining T-norm as the minimum operation *MIN (a, b)* [1]

TABLE III. EXAMPLE: DISPLAY *NEW* CARS FOR SALE WITH *HIGH* ENGINE CAPACITY (FIRST VERSION)

Fuzzy query (conventional form)
SELECT * **FROM** tbl_cars **WHERE** prod_year IS **'NEW'** **AND** engine_capacity IS **'HIGH'**
An equivalent query in SQL - Oracle 11g XE
SELECT * **FROM** (**SELECT** c.model, c.type, c.prod_year, c.engine_capacity, **MIN**(CASE column_name WHEN 'prod_year' THEN FP_TRAPEZOID(prod_year,a,b,c,d) WHEN 'engine_capacity' THEN FP_TRAPEZOID(engine_capacity,a,b,c,d) END) AS FUZZY_DEGREE **FROM** tbl_cars c, tbl_possibility **WHERE** (linguistic_variable_value=**'NEW'** or linguistic_variable_value=**'HIGH'**) AND (column_name='prod_year' or column_name='engine_capacity') group by c.model, c.type, c.prod_year, c.engine_capacity) **WHERE** FUZZY_DEGREE > 0;

TABLE IV. EXAMPLE: DISPLAY *NEW* CARS FOR SALE WITH *HIGH* ENGINE CAPACITY (SECOND VERSION)

Fuzzy query (conventional form)
SELECT * **FROM** tbl_cars **WHERE** prod_year IS **'NEW'** **AND** engine_capacity IS **'HIGH'**
An equivalent query in SQL - Oracle 11g XE
SELECT c.model, c.type, c.prod_year, c.engine_capacity, **MIN**(CASE column_name WHEN 'prod_year' THEN FP_TRAPEZOID(prod_year,a,b,c,d) WHEN 'engine_capacity' THEN FP_TRAPEZOID(engine_capacity,a,b,c,d) END) AS FUZZY_DEGREE **FROM** tbl_cars c, tbl_possibility **WHERE** (linguistic_variable_value=**'NEW'** or linguistic_variable_value=**'HIGH'**) AND (column_name='prod_year' or column_name='engine_capacity') group by c.model, c.type, c.prod_year, c.engine_capacity **having MIN**(CASE column_name WHEN 'prod_year' THEN FP_TRAPEZOID(prod_year,a,b,c,d) WHEN 'engine_capacity' THEN FP_TRAPEZOID(engine_capacity,a,b,c,d) END) >0;

The value of the column labelled *fuzzy_degree*, the result of which was presented in Fig. 7, was set as $min(\mu_{prod\_yearNEW}(prod\_year), \mu_{engine\_capacityHIGH}(engine\_capacity))$.

```
SELECT *
FROM (
  SELECT c.model, c.type, c.prod_year, c.engine_capacity,
         MIN(CASE column_name
             WHEN 'prod_year' THEN FP_TRAPEZOID(prod_year,a,b,c,d)
             WHEN 'engine_capacity' THEN FP_TRAPEZOID(engine_capacity,a,b,c,d)
             END) AS FUZZY_DEGREE

  FROM tbl_cars c, tbl_possibility
  WHERE (linguistic_variable_value='NEW' or linguistic_variable_value='HIGH') AND
        (column_name='prod_year' or column_name='engine_capacity')
  group by c.model, c.type, c.prod_year, c.engine_capacity
)
where FUZZY_DEGREE > 0;

+-------------+--------+-----------+-----------------+--------+--------------+
| MODEL       | TYPE   | PROD_YEAR | ENGINE_CAPACITY | PRICE  | fuzzy_degree |
+-------------+--------+-----------+-----------------+--------+--------------+
| ASTON MARTIN | COUPE  |      2015 |             6.0 | 900000 |  0.666666667 |
| MAZDA       | CABRIO |      2012 |             2.0 |  43990 |  0.400000000 |
+-------------+--------+-----------+-----------------+--------+--------------+
```

Fig. 7. Result of the fuzzy query from Table III - combining fuzzy condition

C. *Tests*

Two different tests on the Oracle 11g XE database have been performed to analyze the proposal presented in this article:

- varying complexity of the query (examples of queries presented in this article),
- varying the number of tuples computed on a same query, to analyze the system scalability.

Tests have been conducted on a PC equipped with an Intel Core i5 2.80 GHz CPU and 12 GB of RAM. Queries, varying in complexity, are presented in Table V together with descriptions.

TABLE V. SET OF FUZZY QUERIES [1]

Query ID	Fuzzy Query	Description	Rows	Time
Q1	cars with bodywork *similar* to *coupe* [1]	fuzzy condition for a similarity type variable	1000	8
Q2	*new* cars for sale with *high* engine capacity (first version)	combining fuzzy conditions	1000	13
Q3	*new* cars for sale with *high* engine capacity (second vesion)	combining fuzzy conditions	1000	22
Q4	*average* price of a car depending on the group to which it belongs: *new* cars, *average* cars, *old* cars	fuzzy aggregate functions, grouping by fuzzy values	1000	37
Q5	*number* of cars that have *small*, *middle*, *high* engine capacity	fuzzy aggregate functions, grouping by fuzzy values	1000	38
Q6	*new* cars for sale [1]	fuzzy condition for possibility type variable	1000	54

A set of different queries has been designed to measure the performance of the system. These queries, varying in complexity, are presented in Table V together with descriptions, the number of tuples and their execution time in milliseconds. In Fig. 8, we can observe how the execution time increases when the number of attributes or the complexity of the query increases. It is important to notice that the exception is query Q6, it is the simplest, and its execution time is the longest.

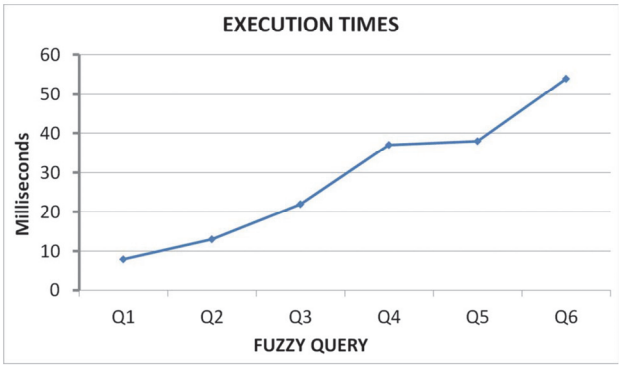


Fig. 8. Execution times in queries performed on Oracle 11g XE (a)

TABLE VI. EXECUTION TIMES (IN MILLISECONDS) VARYING NUMBER OF TUPLES

Number of tuples	Q1	Q2	Q3	Q4	Q5	Q6
10	4	3	13	3	5	3
50	4	3	13	5	5	8
100	4	4	13	9	7	12
200	4	6	14	12	11	21
300	3	6	14	16	15	22
400	3	9	18	19	19	36
500	5	9	19	24	21	46
1000	8	13	22	37	38	54

Scalability was measured by varying the number of tuples in the database from 10 to 1000. For this purpose, examples of queries presented in this article have been executed. Execution times are shown in Table VI and they have been illustrated in Fig. 9.

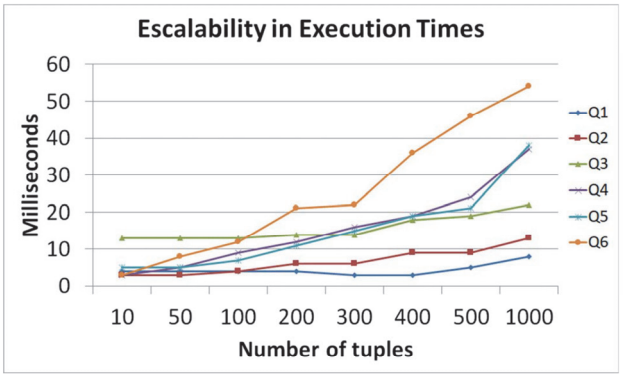


Fig. 9. Execution times in queries performed on Oracle 11g XE (b)

In these figures, we can notice how the scalability grows accordingly with the number of computed rows in the queries. The ones that have more computational needs and/or have more complex syntax (eg. this same example of a query (in Table V marked as Q2, Q3) has different execution times). However, it is noticeable that varying the number of rows is not enough to distinguish the delay provoked by the changes in the number of computed rows, especially in the most simplest queries (Q6).

III. Comparison of most relevant features in fuzzy query system

A comparison between the features of the main fuzzy relational databases in the literature and proposal presented in this article is shown in Tables VII-IX. First models were mainly theoretical proposals of fuzzy relational databases. Prade H. and Testemale C. [12] have presented the original code in MACLISP on DPS8 for fuzzy query processing.

TABLE VII. Comparison of most relevant features in fuzzy query systems (part I)

Model	Buckles B. P., Petry F. E. [11]	Prade H., Testemale C. [12	Zemankova M., Kandel A. [13]	Medina J. M., Pons O., Vila M. A. [14]
Manage scalar data	x	x	x	x
Manage non-scalar data	x	x	x	x
Similarity relationship	x		x	x
Possibility distributions		x	x	x
Degree in attributes level	x	x		x
Degree in tuple level		x	x	x
Fuzzy modifiers			x	
Fuzzy quantifiers				x
Fuzzy comparison operators	x	x	x	x
Fuzzy group by				
Fuzzy joins		x		x
Store fuzzy data	x	x	x	x
Fuzzy queries	x	x	x	x
Extension SQL language				x

Umano [15] and Fukami have presented FOOBD in SQL. The most complete implementations were provided by: Bosc P. and Pivert O. [16] called Sqlf and Kacprzyk J. and Zadrożny S. [17] called FQuery in Microsoft Access. Also, Medina et al. [14] proposed a conceptual framework for fuzzy representation called GEFRED (Generalized Model for Fuzzy Relational Databases) and a language called FSQL (Fuzzy SQL, SQL extension) in Oracle. An implementation presented in this article is based on a proposal by Kacprzyk J. and Zadrożny S. and Medina. It is worth noting that there are a few functionalities, that GEFRED has defined theoretically [18], that have been included neither in the implemented version, i.e. fuzzy joins and fuzzy quantifiers, nor in the implementation presented in this article. [1]

TABLE VIII. Comparison of most relevant features in fuzzy query systems (Part II)

Model	Umano M., Hatono I., Tamura H. [15]	Bosc P., Pivert O. [16]	Kacprzyk J., Zadrożny S. [17]	Martinez-Cruz C., Noguera J. M., Vila M. A. [18]
Manage scalar data	x	x	x	x
Manage non-scalar data				x
Similarity relationship				x
Possibility distributions		x	x	x
Degree in attributes level	x			x
Degree in tuple level	x	x	x	x
Fuzzy modifiers		x		
Fuzzy quantifiers		x	x	
Fuzzy comparison operators	x	x	x	x
Fuzzy group by		x	x	
Fuzzy joins	x	x		
Store fuzzy data	x			
Fuzzy queries	x	x	x	x
Extension SQL language	x	x	x	x

TABLE IX. Comparison of most relevant features in fuzzy query systems (Part III)

Model	Nowakowski G. [1]	Proposal presented in this article
Manage scalar data	x	x
Manage non-scalar data		
Similarity relationship	x	x
Possibility distributions	x	x
Degree in attributes level		
Degree in tuple level	x	x
Fuzzy modifiers		
Fuzzy quantifiers		

Fuzzy comparison operators	x	x
Fuzzy group by		x
Fuzzy joins		
Store fuzzy data		
Fuzzy queries	x	x
Extension SQL language	x	x

IV. CONCLUSIONS

Fuzzy queries allow using a natural language. Nevertheless, in order to maintain the fuzzy nature of these expressions, they are modelled with fuzzy sets. This way, fuzzy queries enable improved representation of the requirements of the user through direct expression of the same with linguistic terms and through the use of complex methods of aggregation of partial conditions.

Fuzzy queries can be applied even if the user precisely defined his/her requirements. Yet, their application is only justified if there is no data which meets these requirements. When a classic query with precisely defined conditions returns an empty data set, a fuzzy query with imprecisely defined conditions may return a data set which is not empty. Some of the obtained results (with the highest degree of meeting conditions of the query) may be accepted by the user. This way, the user will have a better chance of learning the content of the database and consequently will have the opportunity to modify the query. A modified query may take into account the contents of the database and may better reflect the actual requirements of the user. The following conclusion can be drawn: as fuzzy queries arrange results according to the degree of meeting conditions of the query, it is easier to analyse the results and the risk of obtaining an empty result is reduced thanks to an extended interpretation of the conditions of the query.

The provided examples of conversion of fuzzy queries into standard SQL queries by means of Oracle 11g XE point to easy to implement methods of obtaining fuzzy information from the database and thereby expand its functionality. Moreover, a qualitative comparison between the most relevant fuzzy query systems in the literature and proposal presented in this article has addressed the strengths and drawbacks of this contribution. [1]

ACKNOWLEDGMENT

Presented results of the research, which was carried out under the theme No. E-3/611/2017/DS, were funded by the subsidies on science granted by Polish Ministry of Science and Higher Education.

REFERENCES

[1] G. Nowakowski: “Conversion of fuzzy queries into standard SQL queries using Oracle 11G XE“, Technical Transactions. Electrical Engineering, Iss. 3-E, 2016, s. 197-213

[2] S. Zadrożny: “Zapytania nieprecyzyjne i lingwistyczne podsumowania baz danych“, Warszawa: Akademicka Oficyna Wydawnicza, 2006.

[3] K. Myszkorowski, S. Zadrożny, P. S. Szczepaniak: “Klasyczne i rozmyte bazy danych : modele, zapytania i podsumowania“, Warszawa : "Exit", 2008.

[4] S. Zadrożny, J. Kacprzyk: “Bipolar Queries Using Various Interpretations of Logical Connectives“, Foundations of Fuzzy Logic and Soft Computing, 2007.

[5] R. Mesiar, H. Thiele: “On T-Quantifiers and S-Quantifiers. In: Novak, V., Perfilieva“, I. (eds.) Discovering the World with Fuzzy Logic, pp. 310–326. Physica-Verlag, Heidelberg, 2000.

[6] V. Novak, I. Perfilieva: “Fuzzy Logic on the Basis of Classical Logic“. Kacprzyk J., Krawczak M., Zadrożny S. (red.): Issues in Information Technology, EXIT, 2002.

[7] D. Baczyński, S. Bielecki, M. Parol, P. Piotrowski, J. Wasilewski: “Sztuczna inteligencja w praktyce“, Warszawa, Oficyna Wydawnicza Politechniki Warszawskiej, 2008

[8] D. Rutkowska, M. Piliński, L. Rutkowski: “Sieci neuronowe, algorytmy genetyczne i systemy rozmyte“, Wydawnictwo Naukowe PWN, Warszawa 1997

[9] R. Yager, D. Filev: “Podstawy modelowania i sterowania rozmytego“, WNT, Warszawa 1995

[10] E. Czogała, W. Pedrycz: “Elementy i metody teorii zbiorów rozmytych“, Wydawnictwo Naukowe PWN, Warszawa 1985

[11] B. P. Buckles, F. E. Petry: “A fuzzy representation of data for relational databases“, Fuzzy Sets and Systems, Volume 7, Issue 3, May 1982, pp. 213-226

[12] H. Prade, C. Testemale: “Generalizing database relational algebra for the treatment of incomplete or uncertain information and vague queries“, Information Sciences, Volume 34, Issue 2, November 1984, pp. 115-143

[13] M. Zemankova, A. Kandel: “Implementing imprecision in information systems“, Information Sciences, Volume 37, Issue 1-3, Dec. 1985, pp. 107-141

[14] J. M. Medina, O. Pons, M. A. Vila: “GEFRED. A Generalized Model of Fuzzy Relational Databases“. Information Sciences, vol. 76(1-2), 1994, pp. 87-109

[15] M. Umano, I. Hatono, H. Tamura: “Fuzzy database systems“. In Proceedings of the FUZZ-IEEE/IFES’95 Workshop on Fuzzy Database Systems and Information Retrieval, Yokohama, Japan, 1995, pp. 53-36

[16] P. Bosc, O. Pivert: “SQLf: a relational database language for fuzzy querying“. IEEE transactions on Fuzzy Systems 3 (1), pp. 1-17, 1995

[17] J. Kacprzyk, S. Zadrożny: “SQLf and FQUERY for Access“. In Proceedings of the IFSA World Congress and 20th NAFIPS International Conference, vol. 4, 2001, pp. 2464–2469

[18] C. Martinez-Cruz, J. M. Noguera, M. A. Vila: “Flexible queries on relational databases using fuzzy logic and ontologies“, Information Sciences, Volume 366, 20 October 2016, pp. 150-164

The 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications
18-21 September, 2019, Metz, France

Methodology of Transformation of Fuzzy Queries into Queries in the SQL Standard

Grzegorz Nowakowski
Department of Automatic Control and Information Technology, Faculty of Electrical and Computer Engineering,
Cracow University of Technology, Cracow, Poland, gnowakowski@pk.edu.pl, pk.edu.pl

***Abstract*—The article presents two approaches to the method of implementation of transformations of fuzzy queries into queries in the SQL standard. This implementation can be carried out: (1) in a database (in the form of functions, procedures, triggers or other acceptable form), using the existing query processing system in the database, and thus it can expand the functionality of that database; (2) in an application (which communicates with the database), and thus a considerable portion of calculations will be transferred from the database to that application. An analysis of those approaches was conducted and it was demonstrated that the presented methods should enjoy extensive practical application.**

Keywords—fuzzy logic; fuzzy queries; conversion; transformation; methodology; SQL

I. INTRODUCTION

When creating information systems, which allow the possibility of formulation and implementation of standard SQL queries containing fuzzy components as well as use relational databases (imposing the necessity of a precise specification of conditions) at the same time, we can distinguish two approaches to the method of implementation of transformations of fuzzy queries into queries in the SQL standard. This implementation can be carried out as follows: (1) in a database (in the form of functions, procedures, triggers or other acceptable form), using the existing query processing system in the database, and thus can expand the functionality of that database; (2) in an application (which communicates with the database), and thus a considerable portion of calculations will be transferred from the database to that application.

The first approach was used, among others, in the following works: [1], [2], [15], [17], [20], [21]-[24], [26], [27] where after defining a number of operators, membership functions, sum functions, the product and fuzzy negations, aggregate functions and other components, the database was allowed to interpret queries containing fuzzy components.

The second approach was employed, among others, in the following works: [3], [4], [19], [25], [28]-[31] where it became necessary to prepare the handling of queries within a much wider scope.

A relational database management system is able to interpret only classical queries which in turn specify the scope of data and conditions that the data should fulfil during the entire process of retrieval of information from the database. Those conditions should be specified in a precise manner, since precision is the basic requirement when defining conditions in classical query languages. At the same time, those conditions impose limitations.

Therefore, employing the second approach involves the necessity to develop and implement a number of algorithms which are generally already integrated into every relational database. Thus, seemingly uncomplicated procedures (such as joining of conditions or grouping of results) takes on a whole new dimension once fuzzy logic has been used.

In the article, an analysis of those approaches was conducted and it was demonstrated that the presented methods should enjoy extensive practical application.

II. FUZZY QUERY

A fuzzy relational database [1]-[4] query is a query containing explicitly used expressions of a natural language called linguistic terms. They are modelled using fuzzy logic and they specify the following: imprecise values, imprecise comparisons and non-standard methods of aggregation of degrees of satisfaction of partial conditions of a query.

Individual rows satisfy such a query to a certain degree expressed by a number from the range <0,1> where 1 means total satisfaction and 0 total lack of satisfaction of a query. Therefore, the result of the query is a set of rows ordered by the degree of satisfaction of the query, which allows one to reflect better the attempts made by a given person to match the data to the query. In the case of complex queries, which contain both imprecise linguistic terms and non-standard aggregation patterns, the satisfaction is of a gradual nature. In such a situation, the human naturally assesses the data as those that meet their requirements more or less, and does not divide them only into those that fulfil or do not fulfil their requirements. The concept of the degree of satisfaction allows formal presentation of those complex queries and thus ensuring of natural ordering of query results so that the data that best satisfy the query are at the beginning of the result list [2].

978-1-7281-4069-8/19/$31.00 ©2019 IEEE

III. Methodology of Conversion of a Fuzzy Query to a Query in the SQL Standard

Both the construction of imprecise queries and their implementation as well as the adopted grammar of imprecise queries are usually strongly related to the language of queries of a specific database. A relational database (Oracle 11g XE), where the query language was SQL, was adopted for these considerations. At the same time, that database communicated with an application written in JAVA. Both the adopted database and the application will make it possible to present the highlighted approaches to the method of implementation of transformations of imprecise queries into queries in the SQL standard.

There are many imprecise components in database queries: (1) atomic predicates based on linguistic terms: young, tall; (2) complex predicates: fuzzy sum and product operators; (3) atomic predicates based on the similarity of linguistic terms: information technology ≈ artificial intelligence; (4) modified predicates: very, more or less, rather, antonyms; (5) fuzzy operators: approximately, a little more, etc.; (6) linguistic quantifiers; (7) fuzzy joining of relations; (8) fuzzy aggregate functions; (9) grouping by fuzzy values.

In this article, examples of imprecise components in database queries found in the first two items of the above list were taken into account..

The whole interpretation of fuzzy queries boils down to the following general rule: the degree of satisfaction of a query is identified with the membership function value of the appropriate fuzzy set. In contrast, in the case of complex queries, partial degrees of satisfaction calculated in such a manner, corresponding to individual conditions occurring in a query, are aggregated with the use of selected operators [1].

A number of membership functions [9]-[11] were developed. In particular, the following can be mentioned here: (1) polygonal functions – consisting of straight sections (triangular, trapezoidal and rectangular functions are used most commonly); (2) Gaussian functions.

The trapezoidal function (Figure 1) stands out amongst the above-mentioned membership functions with its good performance and simple implementation, and it will be employed in further considerations [2]. This function is described by four parameters (a, b, c, d). Knowing these parameters allows one to calculate the degree of membership of a given component in a set in line with the formula from Figure 1.

Fuzzy queries, similarly to queries of the classical SQL language, can contain complex conditions created as a result of joining of individual conditions by means of logical operators. In the SQL language, keyword AND corresponds to the conjunction operator and OR – to the disjunction operator, whereas NOT is the negation operator [1], [2].

$$\mu_A(x;a,b,c,d) = \begin{cases} 0 & x \le a \\ \frac{x-a}{b-a} & a < x \le b \\ 1 & b < x \le c \\ \frac{d-x}{d-c} & c < x \le d \\ 0 & x > d \end{cases}$$

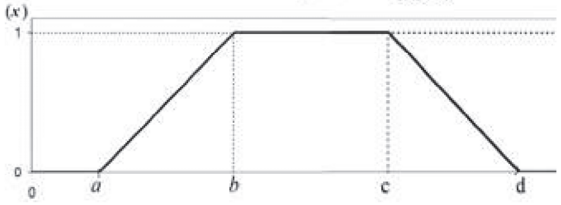


Figure 1. Trapezoidal membership function

The same keywords are used in fuzzy queries, however they correspond to fuzzy conjunction, fuzzy disjunction and fuzzy negation respectively.

Nevertheless, it should be noted that many operators completing the cut operation (of a product that corresponds to logical operation AND) are used in the theory of fuzzy sets [8]. They are used interchangeably depending on the problem being solved. Most of those operators meet the criteria of the so-called triangular norm (T-norm):

$$\mu_{A\cap B}(x) = T(\mu_A(x), \mu_B(x)) \quad (1)$$

Similarly to the completion of the cut operation, many operators are used for the joining operation (of a logical sum that corresponds to logical operation OR). The most commonly used operators satisfy the conditions of the so-called triangular conorm (S-norm or T-conorm):

$$\mu_{A\cup B}(x) = S(\mu_A(x), \mu_B(x)) \quad (2)$$

Criteria defining triangular norms consist of four basic conditions and they will not be presented due to this article's space constraints. Those conditions were described in [3], [4], [5], [9], [10], among others.

An S-norm corresponds to each T-norm if the following condition is satisfied:

$$T(a,b) = 1\text{-}S(1\text{-}a,1\text{-}b) \text{ or } S(a,b) = 1\text{-}T(1\text{-}a,1\text{-}b) \quad (3)$$

Operators fulfilling condition (3) make up the so-called complementary pairs (coupled, dual).

Many operators satisfying the conditions of T-norms and S-norms have been developed and described. Those operators are divided into non-parametrized, which have a fixed manner of operation, and parametrized (also called triangular norm families) whose manner of operation changes depending on the adopted parameter (degree of freedom) for such an operator. In [1] selected tringle norms forming complementary pairs have been presented.

Triangular norm operators perform operations on two fuzzy sets only (variables). Operations on a larger number of sets can be performed gradually, combining sets into pairs, the order of combination of sets having no impact on the result (characteristic of conjunctiveness).

From the above reasoning, it follows that the degree of satisfaction of a complex condition is calculated on the basis of degrees of satisfaction of constituent conditions as well as the selected T-norm and S-norm complementary to it.

The following operators are taken into account in further considerations: (1) T-norm (imitating logical operator AND) – MIN(a, b); (2) S-norm (imitating logical operator OR) – MAX(a, b)

Let consider the following imprecise query (formulated in a conventional manner) displaying a sale offer of *new* cars with *high* engine capacity:

```
SELECT *
FROM tbl_cars
WHERE prod_year IS 'NEW'
AND engine_capacity IS 'HIGH'
```

This query contains fuzzy conditions in phrase WHERE. The first one compares numeric column *prod_year* with a value expressed by means of linguistics value 'NEW', whereas the second one compares numeric column *engine_capacity* with a value expressed by means of linguistic value 'HIGH'. The conditions are joined by means of conjunction AND.

We will use this query to show the methodology of conversion of such a query into a query in the SQL standard, the conversion taking place in: (1) an Oracle 11g XE database; (2) an application written in JAVA that communicates with the Oracle 11g XE database

A. Implementation in the database

A fuzzy query was transformed into a query in the SQL standard using Oracle 11g XE:

```
SELECT *
FROM (
SELECT c.model, c.type, c.prod_year,
c.engine_capacity,
MIN(CASE column_name
WHEN 'prod_year' THEN
FP_TRAPEZOID(prod_year,a,b,c,d)
WHEN 'engine_capacity' THEN
FP_TRAPEZOID(engine_capacity,a,b,c,d) END)
AS FUZZY_DEGREE

FROM tbl_cars c, tbl_possibility

WHERE (linguistic_variable_value='NEW' or
linguistic_variable_value='HIGH') AND
(column_name='prod_year' or
column_name='engine_capacity')
group by c.model, c.type, c.prod_year,
c.engine_capacity
  ) WHERE FUZZY_DEGREE > 0;
```

All variables are described by a trapezoidal membership function whose definition was written in PL/SQL, in the Oracle 10g XE database. In that query:

- a defined fuzzy set corresponding to term *new* for attribute *prod_year* was used;
- a defined fuzzy set corresponding to term *high* for attribute *engine_capacity* was used;
- a conjunction of the two above fuzzy conditions was performed, defining T-norm as the minimum operation.

The value of column *fuzzy_degree*, whose result was shown in Figure 2, had been determined as $\min(\mu_{prod\_yearNEW}(\text{prod\_year}), \mu_{engine\_capacityHIGH}(\text{engine\_capacity}))$.

MODEL	TYPE	PROD_YEAR	ENGINE_CAPACITY	PRICE	fuzzy_degree
ASTON MARTIN	COUPE	2015	6.0	900000	0.666666667
MAZDA	CABRIO	2012	2.0	43990	0.400000000

Figure 2. Final result of the fuzzy query

The most important thing for the user are results with as high as possible degree of satisfaction. As a rule, results with a low degree of satisfaction are of little use. They can be obtained by introducing an additional expression placed at the end of phrase WHERE or HAVING, where the said minimum acceptable degree of satisfaction is specified (e.g. WHERE FUZZY_DEGREE > 0.75). We will then get only those steps whose degree of satisfaction is equal to or higher than the set threshold of satisfaction.

TABLE I. CALCULATION OF THE DEGREE OF SATISFACTION OF AN IMPRECISE QUERY (PREPARED ON THE BASIS OF [1],[[3], [4]).

Step	Database system
1	retrieves a row from the table
2	calculates partial degrees of satisfaction of all simple conditions of a query (or only selected ones, depending on the structure of the query and the optimisation employed), substituting in place of attributes their values from the current row
3	aggregates partial degrees of satisfaction calculated in the previous step to the total degree of satisfaction of a query
4	if the total degree of satisfaction is sufficiently high (exceeds the threshold value specified by the user or adopted as default), the row is attached to the query response along with the degree of satisfaction
5	proceeds to step 1
6	if there are no more steps, then STOP will occur

Considering the issue of calculation of the degree of satisfaction of a fuzzy query, a certain significantly simplified vision of implementation of such a query is to be adopted. The assumption is that when implementing a query, the whole table is viewed sequentially and that the

degree of satisfaction is calculated for each row. The database system performs steps presented in Table 1.

B. Implementation in the application

The presence of complex conditions in a fuzzy query sets the developer a task involving preparation of the appropriate procedure of dissection and analysis of such a condition. In the proposed solution, a tree of conditions is built for that purpose and then it is processed using a stack

The mechanism of creation of a tree of conditions in the considered imprecise query:

```
WHERE <condition1> AND <condition2>
```

In line with the algorithm, the following get placed on the stack sequentially: `condition1`, operator `AND` and `condition2`. The stack has the following form at this point:

```
condition2
AND
condition1
```

Since at this stage the conditions contained in phrase WHERE have been exhausted, a complete reduction of the stack takes place (its three top components), and we get the following form:

```
condition2 AND condition1
```

In practice, there is only a node describing conjunction `AND` (the root of the whole tree of agreement of conditions) at the top of the stack, whereas components `condition2` and `condition1` will be attached to it later. However, for transparency of the notation, conditions have been added which that operator joins.

The above considerations assume that the following operations should be performed: (1) dissection of a fuzzy complex condition in order to isolate individual simple conditions and logical operators joining them; (2) recording of information on the complex condition within the structure which will then be able to support the processing of results of a classical query obtained from the database into results of a fuzzy query and (3) processing of elementary fuzzy conditions into elementary classical conditions and building of a classical condition out of them.

As a result of the action of the above operations, the following result will be generated (Figure 3):

MODEL	TYPE	PROD_YEAR	ENGINE_CAPACITY	PRICE	PROD_YEAR_ FUZZY_DEGREE	ENGINE_CAPACITY_ FUZZY_DEGREE	FUZZY_DEGREE
ASTON MARTIN	COUPE	2015	6.0	900000	0.0	0.0	0.0
MAZDA	CABRIO	2012	2.0	43990	0.0	0.0	0.0
OPEL	COMBI	1993	1.7	1500	0.0	0.0	0.0
VOLKSWAGEN	SEDAN	2007	1.9	23900	0.0	0.0	0.0
AUDI	COUPE	2005	1.9	21000	0.0	0.0	0.0

Figure 3. Interjacent result of the fuzzy query

Analysing that result, one can see the following property: columns in the following form have been isolated: *column_name_fuzzy_degree*. Those are columns in which after the query has been carried out, the initial zeroes will be replaced with the appropriate degrees of satisfaction of rows with query conditions. In the event that it is necessary to create more than one columns of the degree of satisfaction for a given column, they get successive names (the first name corresponds to the first fuzzy condition when looking from the left, the second one corresponds to the second fuzzy condition, etc.). The last one on that list is column *fuzzy_degree*, whose values will be determined as the value of the adopted t-norm (imitating logical operator AND) – MIN(a, b).

Furthermore, it should be noted that the fuzzy conditions were processed into classical conditions. Parameters of a trapezoidal function for value 'NEW' had been determined and then the condition was processed into a classical one. The second fuzzy condition was treated similarly. Parameters of trapezoidal function for value 'HIGH' had been determined and then the condition was processed into a classical one.

The final step of the algorithm was the calculation of actual degrees of satisfaction of individual elementary values in successive rows with elementary conditions of a fuzzy query and the calculation of the final, total degree of satisfaction for each row. The result of that operation (and at the same time the entire processing) was shown in Figure 4.

MODEL	TYPE	PROD_YEAR	ENGINE_CAPACITY	PRICE	PROD_YEAR_ FUZZY_DEGREE	ENGINE_CAPACITY_ FUZZY_DEGREE	FUZZY_DEGREE
ASTON MARTIN	COUPE	2015	6.0	900000	1.00000	0.66667	0.66667
MAZDA	CABRIO	2012	2.0	43990	0.40000	0.44444	0.40000
OPEL	COMBI	1993	1.7	1500	0.0	0.11111	0.0
VOLKSWAGEN	SEDAN	2007	1.9	23900	0.0	0.33333	0.0
AUDI	COUPE	2005	1.9	21000	0.0	0.33333	0.0

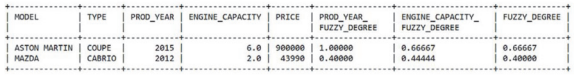
Figure 4. Final result of the fuzzy query

It is worth noting here that in the case of a complex fuzzy condition, the determination of the degree of agreement of individual tuples with the conditions of an imprecise query is done using a recursive algorithm. It envisages determination of the degree of agreement of a tuple with individual elementary fuzzy conditions and then, using t-norm operators or (and) s-norm operators, obtaining of increasingly more complex sub-conditions, and finally an entire complex fuzzy condition. It should be mentioned that the previously preserved (during the analysis of the fuzzy condition) structure of queries is employed here as are the employed logical operators joining complex sub-conditions and their priorities. That is why, due to the algorithm's recursion, particularly good results are obtained using a dynamic tree structure as the structure storing the data on a fuzzy condition while an algorithm based on a stack is used to construct it.

As in the case of implementation in a database, results with the highest possible degree of satisfaction are most important for the user. As a rule, results with a low degree of satisfaction are of little use. The level of α-cut

can be determined in the same way that was reported when discussing databases.

Final result of the fuzzy query by applying α-cut > 0 was shown in Figure 5.



MODEL	TYPE	PROD_YEAR	ENGINE_CAPACITY	PRICE	PROD_YEAR_ FUZZY_DEGREE	ENGINE_CAPACITY_ FUZZY_DEGREE	FUZZY_DEGREE
ASTON MARTIN	COUPE	2015	6.0	900000	1.00000	0.66667	0.66667
MAZDA	CABRIO	2012	2.0	43990	0.40000	0.44444	0.40000

Figure 5. Final result of the fuzzy query by applying α-cut > 0

IV. COMPARISON OF MOST RELEVANT FEATURES IN FUZZY QUERY SYSTEMS

A comparison between the features of the main fuzzy relational databases in the literature and proposal presented in this article is shown in Tables II-III. First models were mainly theoretical proposals of fuzzy relational databases. Prade H. and Testemale C. [13] have presented the original code in MACLISP on DPS8 for fuzzy query processing.

TABLE II. COMPARISON OF MOST RELEVANT FEATURES IN FUZZY QUERY SYSTEMS (PART I)

Model	Buckles B. P., Petry F. E. [12]	Prade H., Testemale C. [13]	Zemankova M., Kandel A. [14]	Medina J. M., Pons O., Vila M. A. [15]
Manage scalar data	x	x	x	x
Manage non-scalar data	x	x	x	x
Similarity relationship	x		x	x
Possibility distributions		x	x	x
Degree in attributes level	x	x		x
Degree in tuple level		x	x	x
Fuzzy modifiers			x	
Fuzzy quantifiers				x
Fuzzy comparison operators	x	x	x	x
Fuzzy group by				
Fuzzy joins		x		x
Store fuzzy data	x	x	x	x
Fuzzy queries	x	x	x	x
Extension SQL language				x

Umano [16] and Fukami have presented FOOBD in SQL. The most complete implementations were provided by: Bosc P. and Pivert O. [17] called Sqlf and Kacprzyk J. and Zadrożny S. [18] called FQuery in Microsoft Access. Also, Medina et al. [15] proposed a conceptual framework for fuzzy representation called GEFRED (Generalized Model for Fuzzy Relational Databases) and a language called FSQL (Fuzzy SQL, SQL extension) in Oracle. An implementation presented in this article is based on a proposal by Kacprzyk J. and Zadrożny S. and Medina. It is worth noting that there are a few functionalities, that GEFRED has defined theoretically [19], that have been included neither in the implemented version, i.e. fuzzy joins and fuzzy quantifiers, nor in the implementation presented in this article [1], [2].

TABLE III. COMPARISON OF MOST RELEVANT FEATURES IN FUZZY QUERY SYSTEMS (PART II)

Model	Umano M., Hatono I., Tamura H. [16]	Bosc P., Pivert O. [17]	Kacprzyk J., Zadrożny S. [18]	Nowakowski G. [2]
Manage scalar data	x	x	x	x
Manage non-scalar data				
Similarity relationship				x
Possibility distributions		x	x	x
Degree in attributes level	x			
Degree in tuple level	x	x	x	x
Fuzzy modifiers		x		
Fuzzy quantifiers		x	x	
Fuzzy comparison operators	x	x	x	x
Fuzzy group by		x	x	x
Fuzzy joins	x	x		
Store fuzzy data	x			
Fuzzy queries	x	x	x	x
Extension SQL language	x	x	x	x

V. CONCLUSION

The article describes the methodology of conversion of a fuzzy query into a query in the SQL standard, taking place in: (1) an Oracle 11g XE database where the needed functions, procedures, etc. were implemented and where the existing query processing system in the database was used, thus expanding the functionality of that database; (2) in an application written in JAVA which communicates with the Oracle 11g XE database where the needed algorithms were implemented (related to the correct interpretation of the syntax of instruction SELECT and the interpretation of filtering conditions in phrases HERE and HAVING) that proved useful in the processing of fuzzy queries and interpretations of obtained results and thus contributed to a large portion of

calculations being moved over from the database to that application.

This leads to the conclusion that it is possible to process fuzzy queries and their results regardless of the database used. There are still a number of important algorithms to be developed without which the proposed methodology of conversion of fuzzy queries in the application will definitely have limited practical application. One can mention here, say, the issue of fuzzy data aggregation or implementation of nested queries. So, further optimisation and development are required to put forward a solution.

Nonetheless, the success of conversion of fuzzy queries into classical queries in the presented methodology as well as ever richer theoretical basis for this type of conversion bodes well for more and more efficient use of this methodology in the future.

ACKNOWLEDGMENT

Presented results of the research, which was carried out under the theme No. E-3/586/2018/DS, were funded by the subsidies on science granted by Polish Ministry of Science and Higher Education.

REFERENCES

[1] G. Nowakowski, *Conversion of fuzzy queries into standard SQL queries using Oracle 11G XE*, Technical Transactions. Electrical Engineering, Iss. 3-E, 2016, pp. 197-213, Y. 113, Iss. 13, doi: 10.4467/2353737XCT.16.277.6076

[2] G. Nowakowski, *Fuzzy queries on relational databases*, 2018 International Interdisciplinary PhD Workshop (IIPhDW), IEEE,. doi: 10.1109/IIPHDW.2018.8388376, 2018, pp. 293-299 (online)

[3] S. Zadrożny, *Flexible queries and linguistic summaries of databases*. EXIT, Warszawa 2006, 247 pages (in Polish).

[4] K. Myszkorowski, S. Zadrożny, P. S. Szczepaniak, *Classical and fuzzy databases: models, queries and summaries*, EXIT, Warszawa 2008, 282 pages (in Polish)

[5] D. Dubois, H. Prade, *Bipolarity in flexible querying*. In: Andreasen, T., Motro, A., Christiansen, H., Larsen, H.L. (eds.) FQAS 2002. LNCS (LNAI), vol. 2522, pp. 174–182. Springer, Heidelberg (2002)

[6] R. Mesiar, H. Thiele, *On T-Quantifiers and S-Quantifiers. In: Novak, V., Perfilieva*, I. (eds.) Discovering the World with Fuzzy Logic, pp. 310–326. Physica-Verlag, Heidelberg, 2000

[7] V. Novak, I. Perfilieva, *Fuzzy Logic on the Basis of Classical Logic*. In book: Selected Topics on Information Technology, Publisher: EXIT, Editors: Kacprzyk J. and Krawczak M. and Zadrożny S., pp.83-130

[8] L.A Zadeh, *Fuzzy sets as a basis for a theory of possibility*, Fuzzy Sets and Systems, Volume 1, Issue 1, 1978, pp. 3-28, ISSN 0165-0114, https://doi.org/10.1016/0165-0114(78)90029-5.

[9] D. Rutkowska, M. Piliński, L. Rutkowski, *Neural network, genetic algorithms and fuzzy systems*, PWN, Warszawa-Lodz, 1997, 410 pages (in Polish)

[10] R. Yager, D. Filev, *Essential of Fuzzy Modeling and Control*, Wiley-Interscience New York, 1994, 408 pages

[11] J. Dombi, *Membership function as an evaluation,* Fuzzy Sets and Systems, Volume 35, Issue 1, 1990, pp. 1-21, ISSN 0165-0114, https://doi.org/10.1016/0165-0114(90)90014-W.

[12] B. P. Buckles, F. E. Petry, *A fuzzy representation of data for relational databases*, Fuzzy Sets and Systems, Volume 7, Issue 3, May 1982, pp. 213-226

[13] H. Prade, C. Testemale, *Generalizing database relational algebra for the treatment of incomplete or uncertain information and vague queries*, Information Sciences, Volume 34, Issue 2, November 1984, pp. 115-143

[14] M. Zemankova, A. Kandel, *Implementing imprecision in information systems*, Information Sciences, Volume 37, Issue 1-3, Dec. 1985, pp. 107-141

[15] J. M. Medina, O. Pons, M. A. Vila, *GEFRED. A Generalized Model of Fuzzy Relational Databases*. Information Sciences, vol. 76(1-2), 1994, pp. 87-109

[16] M. Umano, I. Hatono, H. Tamura, *Fuzzy database systems*. In Proceedings of the FUZZ-IEEE/IFES'95 Workshop on Fuzzy Database Systems and Information Retrieval, Yokohama, Japan, 1995, pp. 53-36

[17] P. Bosc, O. Pivert, *SQLf: a relational database language for fuzzy querying*. IEEE transactions on Fuzzy Systems 3 (1), pp. 1-17, 1995

[18] J. Kacprzyk, S. Zadrożny, *SQLf and FQUERY for Access*. In Proceedings of the IFSA World Congress and 20th NAFIPS International Conference, vol. 4, 2001, pp. 2464–2469

[19] C. Martinez-Cruz, J. M. Noguera, M. A. Vila, *Flexible queries on relational databases using fuzzy logic and ontologies*, Information Sciences, Volume 366, 2016, pp. 150-164

[20] D. Dubois, H. Prade, *Using Fuzzy Sets in Flexible Querying: Why and How?*, Flexible Query Answering Systems, 1996, pp. 45-60

[21] J. Galindo, J. M. Medina, O. Pons, J. C. Cubero, *A server for Fuzzy SQL queries*, Flexible Query Answering Systems, 1998, pp. 164-174

[22] H. Larsen, *An approach to flexible information access systems using soft computing*, Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences, doi: 10.1109/HICSS.1999.772640,1999, pp. 231

[23] D. Rasmussen, R. Yager, *Summary SQL - A Fuzzy Tool For Data Mining*, Intelligent Data Analysis, Volume 1 Issue 1, January 1997, pp. 49-58

[24] A. Yazici, G. Roy, *Fuzzy Database Modeling*, Studies in Fuzziness and Soft Computing, 1999, 233 pages

[25] R. Mama and M. Machkour, *A study on fuzzy interrogation systems of database*, 2019 International Conference of Computer Science and Renewable Energies (ICCSRE), Agadir, Morocco, 2019, pp. 1-6. Doi: 10.1109/ICCSRE.2019.8807564

[26] N. Hoa, *A Type-2 Fuzzy Relational Database Model*, Research and Development on Information and Communication Technology, 2017, pp. 19-26, doi: 10.32913/rd-ict.vol3.no14.352

[27] M. Hudec, M. Vucetic, *Some issues of fuzzy querying in relational databases*, Kybernetika, Volume 51, Number 6, 20`5, pp. 994 - 1022

[28] S. Skrbic, M. Racković, A. Takaci, *Prioritized fuzzy logic based information processing in relational databases*, Knowledge-Based Systems, 2013, pp. 62-73, doi: 10.1016/j.knosys.2012.01.017

[29] R. Rishi, *Fuzzy Querying Based on Relational Database*, IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727 Volume 16, Issue 1, Ver. I (Jan. 2014), pp. 53-59

[30] P. Bosc, O. Pivert, *Fuzzy Preference Queries to Relational Databases,* Imperial College Press, London, UK, 2012, 330 pages

[31] E. Doumard, O. Pivert, G. Smits, V. Thion, *Processing Fuzzy Relational Queries Using Fuzzy Views*, In Proc. of the 28th IEEE International Conference on Fuzzy Systems (Fuzz-IEEE'19), New Orleans, LA, USA, 2019

The 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications
21-23 September, 2017, Bucharest, Romania

Logics Based Application Integration for Interdisciplinary Scientific Investigations

Sergii Telenyk [1,3], Grzegorz Nowakowski [1], Kostiantyn Yefremov [2], Volodymyr Khmeliuk [3]
[1] Department of Automatic Control and Information Technology, Faculty of Electrical and Computer Engineering, Cracow University of Technology, Cracow, Poland, stelenyk@pk.edu.pl, gnowakowski@pk.edu.pl, pk.edu.pl
[2] World Data Center for Geoinformatics and Sustainable Development, Kyiv, Ukraine, k.yefremov@wdc.org.ua,
[3] Department of Automation and Control in Technical Systems, National Technical University of Ukraine "Igor Sikorsky Kyiv, Polytechnic Institute", Kyiv, Ukraine, hmelyuk@gmail.com, kpi.ua/en

***Abstract*—The approach to the integration of applications based on mathematical logic and artificial intelligence for World Data Center (WDC) interdisciplinary scientific investigations has been developed in the article. Key elements of the approach are a multilevel system architecture, a formal logical system, implementation based on the interaction of intelligent agents. The formal logical system has been proposed. The inference method and solution tree recovery mechanism have been elaborated. The implementation of application integration for interdisciplinary scientific research has been based on a stack of modern protocols, enabling communication of business processes over the transport layer of the OSI model. Application integration is also based on coordinated models of business processes, for which an integrated set of business applications have been designed and realized.**

Keywords—interdisciplinary research, application integration; business processes; mathematical logic

I. INTRODUCTION

At present, National Science is obliged to integrate into world and European organizations promoting the consolidation of research [1]. This is a very important process, given the urgent need for interdisciplinary research. It is necessary to create appropriate conditions for exchanging information in solving scientific problems in order to effectively implement interdisciplinary research. Information exchange can be aided by modern information and communication technologies (ICT). But for this to happen, ICT has to be effectively used by specialists from other fields of activity, and not just ICT experts [2].

Consolidation and virtualization have allowed to integrate computing power, data storage and telecommunication systems into a global network to provide access to accumulated information resources. The service approach has spread to infrastructure (Infrastructure as a Service – IaaS), software development tools (Platform as a Service – PaaS), applications (Software as a Service – SaaS). The emergence of a wide range of new types of services, in particular content, has led to the convergence of services and the formation of a generalized concept of information and communication services (ICS). Extensive functionality, a high level of quality and the moderate price of new services have allowed companies to abandon the development of their own IT infrastructure and use a wide variety of services available to design component-based information systems [3].

Unified access to services becomes a condition for distributed systems- and service-oriented technologies. The formation of a new democratic IT-environment, in which even small businesses can supply services, have been accompanied by the use of various tools and access technologies [3]. As historically the IT-environment is heterogeneous, user access to its resources is to some extent complicated or at least uncomfortable. As a result, there is primarily a need for uniform access to services. Otherwise, the advantages of distributed systems and service-oriented technologies will remain unsuitable.

The WDC preserves enormous amounts of astronomical, geophysical, and other scientific data. These data are processed by different programs. This is one of the most powerful information resources used by hundreds of thousands scientists. The demand for it increases in proportion to the need for interdisciplinary research related to sustainable development, the environment protection and others. However, the lack of unified access to data and applications prevents the use of WDC resources effectively. For this reason, in 2008, the World Data System (WDS) was created. The main purpose of the new organization is to ensure a uniform right of access to scientific data for researchers, educators and decision-makers. There is a number of problems facing the WDC, and the main one seems to be the challenge of data and applications integration.

II. RELATED WORK

Integration is generally understood as providing a unified, compatible interface for accessing heterogeneous independent data sources [4]. There are a few successfully realized and sufficiently widespread approaches to data integration in the real systems, for example data spaces [5]. Research in application

978-1-5386-0697-1/17/$31.00 ©2017 IEEE

integration was accompanied by the expansion of business solutions CRM, ERP, SSM and others. The technologies like Component Object Model (COM), Common Object Request Broker Architecture (CORBA), Distributed COM (DCOM), Electronic Access Interchange (EAI) i CORBA have provided integration of different program systems and contributed to solving the urgent problems of users. Service-oriented Architecture (SOA) promoted the formation of the web-applications integration concept [6] [13].

However, the real solution to the problem of integration became possible only after the 2008 crisis, with the idea of cloud computing. IT leaders offered solutions based on integrating platforms that provide full functionality. For most enterprises this solutions were overpriced and they used the business-application of different developers through program-mediators.

Existing application integration decisions are based on different technologies, but their general principles and theoretical approach remain the same. Heterogeneous services must be registered, described and made available to customers and to each other. Network co-operation is based on the protocols of all levels of OSI model, primarily taking into account international, states and branches, and corporate documents [7]. In general, the integration framework consists of the basic and applied parts. The basic part covers four traditional levels of standards of stack protocols of TCP/IP (NIUS, network co-operation, transport and applied), that embrace 7 levels of model of OSI. The applied part covers the user applications in accordance with the class of information system, level of ontology, level of services intellectualization operations based on the XML protocols system, Universal Description, Discovery of and of Integration (UDDI), Web Service Description Language (WSDL), SOAP [9]-[11].

It is possible to clarify the key technology of applied applications integration on the basis of review: applications co-operation on the base of stack UDDI, SOAP (or REST), BPEL, WSDL, XML above a transport level; WCF; the association of all services into one multifunctional business-service using the service-orchestrator. The role of service-orchestrators can be performed by embedded agents.

The scientific community is already using systems based on the decomposition of the overall problem (flow) into a sequence of individual subtasks (sub flow) - the so-called workflow system, for example Triana workflows, Kepler, Taverna Workflow Management System [12]. Unfortunately, the description of streams and components of these systems are incompatible. They do not have the developed intellectual constituent and any working sequence of services needs to be collected by hand or to edit. On their basis a few decisions are worked out, for example GEOSS, for the creation of the application of accumulation, treatment and exchange by scientific data in different areas. But there is a string of defects, above all complication of the association of heterogeneous types of data from different sources in one query.

There is a requirement in the creation of mathematical models and methods as bases of integral decision for providing centers of storage and data processing that will give flexible possibilities at the level of integration of information and availability of services for users. The indicated models and methods must not lack an algorithmic approach, algorithms of processing of data related to the necessity of rescheduling at the appearance of new types of data and change of realized or appearance of new algorithms. Such a decision must provide flexibility and adaptability, the use of remote services, the ability to use the end-user without special knowledge and skills; quick and easy integration with existing information systems.

III. System Architecture

The WDC system organization method of services co-operation is an orchestrator. It allows to group the base services in the hierarchy, subordinating to their administrators – service-orchestrators. The WDC services can be united by common attributes (science area, functionality, regional location and others). The service-orchestrators receive powers in accordance with policy of international exchange by data. Thus, it is possible to organise a simple and effective system, such that the search of necessary service or construction of their composition for difficult queries inherent interdisciplinary research will not require large time, as every Service Orchestrator has information on the functionality of inferior services and in addition, they can coordinate the possibilities in the process of planning and realization of users' queries. Plugging new services in the system will not present specific problems either, as for this purpose it will be necessary only to describe a description of service, and introduce it to a register and fasten after certain Service Orchestrator.

Such an approach will work, when a user exactly knows his information necessities and has knowledge and skills to compile the chain requests to a known orchestrator. But the approach also allows to work with the system, only formulating needs in terms of a particular subject domain. But then for the association of services and the organization of their co-operation an intellectual constituent is needed. For this purpose intellectual agents are used. They realize system logic to promote the formation of queries, planning of their execution, and organization of basic services interaction. The system is two-tier: (1) the bottom level is formed by services that decide base tasks; (2) the overhead level consists of intellectual agents that orchestrate base services. Intelligent agents, using registers of basic subordinate services and their functionality, through interaction implement methods of inference. The inference result is the composition of based agents operations. This composition is transferred to the lower

level, which performs operations forming solution of the user task. At this point, control is transferred to the lower level agents, which give only the final result to be sent to the user.

The implementation of user query comes true after an inference. Actually user request is performed after the formation of proof. The proof contains pointers to data sources and IDs of methods which system services should perform on the data. Some requests do not require pre-processing of data and methods of service are performed immediately after receiving information from the data source. Others require a certain sequence of previously completed operation on the data, and as a result, execution of other service methods. The necessary connections are described in the axioms that are entered into the knowledge base in registration of service in the system. These operations yield the final product of the system - data that are solutions to the user problem. Since the operation of the system is determined by the user's request, the system interface should help him make a request in familiar terms. In the solution that is proposed, UDDI is used to create the service registry. For unified description of services WSDL is used. Messaging is implemented using SOAP, and orchestration of services - BPEL. Agents are described by means of JADE.

The description of the service interface is entered in the knowledge base of the agent-controller, to which the new service will be sub-ordered. The agent-controller is selected in accordance with the policies adopted in the WDC system, for example, by the criteria of the geographical location of the deployed agent. At agent level, the agent-orchestrator (or several), having received the user's request, interrogates the agents in order to search for the necessary services and free resources. The user's request is processed taking into account the services of the system, their functionality (described by the axioms of the system), and the inference rules defined by logical formalism. To realize this interaction of services in the WDC system, the most appropriate is a logical approach. The implementation of approach requires: (1) to describe the existing services in the formal language; (2) to formulate the rules of inference; (3) to determine the inference method; (4) to develop an algorithm of the execution tree recovery; (5) to implement these methods and mechanisms in the agents of the system.

IV. The Formal Logical System

The program system is built on the basis of the formal logic system described in [2]. Consider only the problem modifications necessary for integration. The constructive (construct), procedural (method), and object types (problems, entities, relation) of constants are important for inference. Further, the formulas are converted to the traditional for the clausal logics form (we are talking about the Horn clause).

Specifier - it is construction of kind τ_1, where τ_1 is a therm for an individual object. The specifiers of construct:

- if $e_1^e ... e_i^e$, - individual object of entity kind, $e_1^r ... e_i^r$, - individual object of relation kind, $A_1^j (a_1^1 ... a_j^1) ... A_i^j (a_1^k ... a_j^k)$ - atomic formulas for the individuals of primary types, τ – individual therm of kind construct, then τ: ($e_1^e ... e_i^e$, $e_1^r ... e_i^r$, $A_1^j (a_1^1 ... a_j^1) ... A_i^j (a_1^k ... a_j^k)$) - specifier of construct;
- other specifiers of construct are not.

Precondition:

- if c_1 – specifier of construct, Π - sequence of atomic formulas, then $<\tau_1:(c_1,\Pi)>$ - elementary precondition;
- elementary precondition - precondition;
- if $<\tau_1>$ - precondition, τ_2 - elementary precondition, then $<\tau_1, \tau_2>$ - precondition;
- other preconditions are not.

Post-condition:

- if τ_1 – specifier of construct, 1I - sequence of atomic formulas, then $<\tau_1:(c_1,\Pi)>$ - elementary post-condition;
- elementary post-condition - post-condition;
- if $<\tau_1>$ - post-condition, τ_2 - elementary post-condition, then $<\tau_1, \tau_2>$ - post-condition;
- other post-conditions are not.

Specifier of methods:

- if τ - individual therm of method kind, $<\tau_1>$ - precondition, $<\tau_2>$ - post-condition, then τ: ($<\tau_1>$, $<\tau_2>$) – specifier of method;
- other specifiers of methods are not.

Specifier of problems:

- if $<\tau_1>$ - precondition, a $<\tau_2>$ - post-condition, τ – therm for individual object of kind Problem, then τ: $<<\tau_1>,<\tau_2>>$ - specifier of problems;
- other specifiers of problems are not.

Clause - an expression of kind $\Pi \rightarrow \Lambda$, where Π - sequence of atomic formulas; Λ - only atomic formula.

Atomic formulas will provide these clauses depending on the type: (1) individual (contains only atomic formulas for individuals); (2) typified (contains only atomic formulas for types); (3) general type (contains atomic formulas for individuals and types).

The knowledge base of a system consists of axiom ontology, that depicts methods of system services, and data source ontologies, described in OWL language based on RDF. Also, the knowledge base includes rules of inference necessary to obtain a desired result when it is necessary to combine methods both one and different services.

Inference rules:

- **If** d1: (<τ1>,< τ2>) **and** d2: (<τ3>,< τ1>) **then** d1: (<d2>,< τ2>)
- **If** d1: (<τ1>,< τ2>) **and** d2: (<τ3 ∧ τ4>,< τ1>) **then** d1: (<d2>,< τ2>)
- **If** d1: (<τ1 ∧ τ2>,< τ3>) **and** d2: (<τ4>,< τ1>) **and** d3: (<τ5>,< τ2>) **then** d1: (<d2 ∧ d3>,< τ3 >)
- **If** d1: (<τ1>,< τ2>) **and** d2: (<τ3 ∨ τ4 >,< τ1>) **then** d1: (<d2>,< τ2>)
- **If** d1: (<τ2 ∧ τ2>,< τ1>) **then** d1: (<τ2>,< τ1>)
- **If** d1: (<τ1, τ2>,< τ3>) **and** d2: (<τ4>,< τ1>) **and** d3: (<τ5>,< τ2>) **then** d1: (<d2, d3>,< τ2>)
- **If** d1: (<τ1>,< τ2,τ3>) **and** d2: (<τ2>,< τ4>) **and** d3: (<τ3>,< τ5>) **then** d2: (<d1>,< τ4>) **and** d3: (<d1>,< τ5>)

V. THE FORMAL LOGICAL SYSTEM

We use the proposed method in [2], based on analogy and types of assertions. The described idea is to build proof in an abstract space with its subsequent use for process control in the initial solutions search space. This will increase the effectiveness of the deduction [8] by cutting off most of the 'conclusion unpromising' branches in the original space. Further multiset literals (atomic formulas) called multi clause (*m*-clause) and literal *L* will be recorded in *m*-clause as many times as it is repeated. An ordinary clause will be assumed as *m*-clause, the multiplicity of every literal in which is 1. The operations ∪ (union), ∩ (intersection), — (difference), · (concatenation) and relation ⊆ (occurrence) for multisets are naturally performed.

Let $A_1 \in C_1$, $A_2 \in C_2$ and α_1, α_2 be such substitutions that some literal *L* on the right side A_1 and on the left side A_2, $A_1\alpha_1=\{L\}$, $A_2\alpha_2=\{L\}$, and if $|A_i|>1$, i = 1, 2, the corresponding substitution α_i is the most common unification of literals from A_1. Then the clause, obtained from the union of clauses $C_1\alpha_1$ and $C_2\alpha_2$ by removing *L* of the left and right parts, is called *m*-resolvent of clauses C_1 and C_2. Let all of *m*-clauses in the original set *S* be ordered and C_1, C_2 are the two of them. Let $A_1 \in C_1$, $A_2 \in C_2$ and α_1, α_2 are such substitutions that for some literal *L* on the right side A_1 and on the left side A_2, $A_1\alpha_1=\{L\}$, $A_2\alpha_2=\{L\}$, and if $|A_i|>1$, i = 1, 2, the corresponding substitution α_i is the most common unification of the literals from A_1. Then the clause, obtained from the union of clauses $C_1\alpha_1$ and $C_2\alpha_2$ by removing *L* of the right part and last *L* of the left part and by eliminating the any underlined literal, after which there are no more literals, is called the ordered linear *m*-resolvent of ordered clauses C_1 and C_2. In the event that the last literal on the left side of the ordered *m*-clause is unified with underlined literal on the right side of the same clause, the ordered linear *m*-resolvent can be obtained by removing the last literal of the ordered *m*-clause in left part (or by means of *m*-clauses reduction).

Further pair, whose first component - the set of deduction vertices, and the second - a set of triples of vertices (for the selection of components a couple of the function-selectors s-N (Tm) and s-N (TT) will be respectively used), is called *m*-resolution proof and denoted *Tm*. Each vertex of *m*-resolution proof is characterized by the tag and the depth in the proof so, that if $n \in$ *s-N(Tm)*, then *s-L*(n) is the tag of vertex *n* and *s-D*(n) is its depth in the proof. If $<n_1, n_2, n_3> \in$ *s-M(Tm),* then *s-L*(n_3) is *m*-resolvent of *s-L*(n_1) and *s-L*(n_2). Each triple of the form is called *m*-resolution. If $n \in$ *s-N(Tm)* and vertex *n is* not the third component of any triples from *s-M(Tm),* then *n* is called the initial vertex of deduction *Tm.* The mark of such vertex is an initial *m*-clause. If $n \in$ *s-N(Tm)* and vertex *n* is neither the first nor the second component of any triples from *s-M (Tm),* then *n* is called the terminal vertex of proof and its mark is called the terminal *m*-clause. In general *Tm* must satisfy the following restrictions: if $< n_1, n_2, n_3> \in$ *s-M(Tm),* then $< n_2, n_1, n_3> \in$ *s-M(Tm).* The depth in the proof for each vertex *n*, $n \in$ *s-N(Tm),* by definition:1) *s-D*(n) = 0 for any initial vertex *n*; 2) *s-D*(n) = 1 + *min*{*max*{*s-D*(n_1), *s-D*(n_2)}|$<n_1, n_2, n> \in$ *s-M(Tm)*} for any no initial vertex *n.*

If initial vertex marks in *Tm* belong to the set of *m*-clauses *S*, then *m*-resolution proof *Tm* is called the proof from *S. C* is inferred from *S,* if *Tm* is proof from *S,* while *C is* mark of one of the *Tm* vertices. Let the value of function Result() be defined to *Tm* and be equal to *C*, if terminal clause *C* of proof *Tm* is the only one, that is Result(*Tm*) = *C.* Let Tm_1 and Tm_2 are *m*-resolution proofs. Proof Tm_1 is called the sub proof of Tm_2 (and denoted $Tm_1 \subseteq Tm_2$), if the following conditions are fulfilled: *s-N*(Tm_1) $\subseteq$ *s-N*(Tm_2) and *s-M*(Tm_1) $\subseteq$ *s-M*(Tm_2). Moreover, if all the initial vertices of Tm_1 are the initial vertices of Tm_2, then Tm_1 is called *the* initial sub proof of Tm_2.

Let *Tm* is the *m*-resolution proof from *S* and all its *m*-clauses are ordered. If for any triple $< n_1, n_2, n_3>$ from *s-M(Tm) s-L*(n_3) is the ordered linear *m*-resolvent of *s-L*(n_1) and *s-L*(n_2), then *Tm* is called the ordered linear proof from *S.* Then, by definition, for *m*-resolution proof must be satisfied by the condition: *max*{*s-D*(n_1), *s-D*(n_2)} = *s-D*(n_3) - 1.

In addition to the general properties of *m*-resolution proofs, the ordered linear proofs have the own features. So, for the ordered linear proof *m*-resolution proof *Tm* are satisfied that conditions: 1) if $< n_1, n_2, n_3> \subseteq$ *s-M(Tm),* then $< n_2, n_1, n_3> \notin$ *s-M(Tm);* 2) for any no initial vertex n_3 *with* the depth in the proof *r* exists the triple $< n_1, n_2, n_3>$, in which the vertex n_1 has the depth in the proof *r*-1, n_2 corresponds to the initial vertex or is absent; 3) *Tm* contains a single terminal clause.

To manage an ordered linear proof, the proposed abstraction of typing in [4] has been used. Let *f* be the mapping from the set of *m*-clauses into the set of *m*-clauses subset such, that: 1) if *m*-clause C_3 is the *m*-resolvent *m*-clauses C_1 and C_2, while $D_3 \in f(C_3)$, then exist $D_1 \in f(C_1)$ and $D_2 \in f(C_2)$ such, that the result of the substitution of some D_1 m-resolvent and i D_2

belongs to D_3; 2) $f(\varnothing) = \{\varnothing\}$; 3) if the result of the some substitution of *m*-clause C_1 belongs to *m*-clause C_2, then for any abstraction D_2 for C_2 exists abstraction D_1 for C_1 such, that the result of the D_1 substitution belongs to D_2. Such mapping is called *f m*-abstraction mapping, while any *D* from *f(C)* is called *m*-abstraction. The mapping of typing is some mapping ϕ from the set of literals into the set of literals, which reflects each atomic formula into the formula, the terms of which have the type that is closest in the hierarchy to the based types. In [4] it has been proved that the mapping of typing φ is a *m*-abstraction mapping.

A relationship between the ordered linear proofs *Tm* and *Um* is called the relationship of improvement (and noted $Tm \to_f Um$), where *f is m*-abstraction mapping, if the vertices of proofs *Tm* and *Um* are in the A relationship of accordance *R*, that: 1) simultaneously following conditions are fulfilled:
$\forall n\,(n \in s - N(Tm))\ \exists n'\,(n' \in s - N(Um))\ (nR\,n')$
$\forall n\,(n \in s - N(Um))\ \exists n'\,(n' \in s - N(Tm)\,(nR\,n')$; 2) if $nR\,n'$, where $n \in s - N(Tm)$, $n' \in s - N(Um)$ then *n* and n' may be initial or terminal vertices only simultaneously; 3) for the terminal vertices of proofs *Tm* and *Um* relationship *R* is the one to one relationship; 4) if $< n_1, n_2, n_3> \in$ *s-M(Tm),* $< n'_1, n'_2, n'_3 > \in$ *s-M (Um)* and $n_3 R n'_3$, then $n_1 R n'_1$ and $n_2 R n'_2$; 5) if *n* and n' are the initial vertices of proofs *Tm* and *Um* respectively and nRn', то *s-L*(n') ∈ *f(s-L(n))*; 6) if *n* and n' are the noninitial vertices of proofs *Tm* and *Um* respectively and nRn', then example *s-L*(n') belongs to *f(s-L(n)).*

So the ordered linear proofs have the property of minimality. In other words, there is exactly one terminal vertex and if a triple $<n_1, n_2, n_3>$ belongs to the minimal ordered linear deduction, then no other triple cannot contain vertex n_3 as the third component, except for triple $<n_1, n_2, n_3>$.

On this property of the ordered linear proofs the inference mechanism has been built, which is realized by agents. Once a user has formulated the problem, the system builds a proof - in fact, checking the possibility of the request satisfaction within existing resources. This task is performed by the Agents-Orchestrators as an important part of the system intellectual kernel. With access to the list of services, their descriptions and axioms that specifies and services application capabilities, these agents trigger the inference process. During the inference process the chain of services is being built that will provide the desired result for a user, starting from the obtaining of data from their relevant sources.

The ordered linear *m*-resolution proofs are specified by a pair of sets: the vertices $V = \{k_1, k_2, k_3 .. k_n\}$ and the vertices triple $T\{< k_1, k_2, k_3 >, < k_3, k_4, k_5 > ... < k_{n-2}, k_{n-1}, k_n >\}$, where k_i, i=1,2,…,*n*-1, is the vertex of the tree, k_n is the terminal vertex, result of working. Forming a proof, we take as its initial vertices the postcondition of the formulated problem and the compatible axiom of knowledge base as the lateral vertex. The axiom is compatible, if the sequence of atomic formulas, matching the proof vertex, or any part of it is a postcondition of axiom. The third vertex of the triple is the result of the inference rule, used to the first two vertices. On the next step the third vertex of the first triple becomes the first vertex of the second triple. This process is repeated until the terminal vertex, which corresponds the formulated problem precondition, will be reached. There are the instances when a formula, corresponding to third vertex of the triple, can be reduced by using the inference rule. Then the next triple will have only two vertices - the first, which corresponds to the formula to reduce, and the second, which corresponds the reduced formula.

To search the compatible axioms the constructs of the currently processed postcondition are compared with the constructs of the ontology axioms. Thus, from the ontology the set of axioms is selected. Moreover, these axioms describe services that process the objects of the desired type and format. The deduction mechanism in the process selects only the axioms in this set. If in continuing the proof from another vertex, there are several compatible axioms in the knowledge base, each of these axioms is used for further construction of the "parallel" version of proof. As a result, the inference mechanism in the process can form a set of different lengths and complexity proofs.

The number of proofs depends on the number of axioms in the knowledge base and the combinations of their application to each vertex. After the inference mechanism is over we have the set of obtained proofs. The proof with the least number of vertices triples and the fewest number of applied axioms is selected from this set. The longer, cyclic and blind proofs are discarded.

This set of formulas and clauses, which corresponds to the proof triples vertices, reflects the application of axioms and rules during the proof construction. This set is the result of the inference mechanism.

VI. The Mechanism of Solution Tree Recovery

To obtain functional sequence of actions from the proof it is necessary to restore the solution scheme. This scheme constitutes a connected directed graph without oriented cycles with parallel directed paths from root to vertices, which has three types of vertices, and is given by this three: $G =< V, E, \Theta >$, where $V = V_1 \cup V_2 \cup V_3$, V_1 - set of vertex-methods, V_2 - set of vertex-forewords and vertex-postconditions, V_3 - set of data vertexes, in which there are combination, or separation of data; E – set of edges; Θ – subset of the Cartesian product E×V×V, which determines the ratio of edges and pairs of vertices. The scheme determines the

sequence and ratio according to the data of actions that will be performed by the data processing system's executive mechanisms to obtain the result desired by the user.

Algorithm for vertex scheme reconstruction. Condition: proof Result = <V, T>. To find: $G =< V, E, \Theta >$.

Step 1. Identification of terminal triple – triple that contains a vertex *k*, that is neither the first nor the second component of any of the triples of the set *T*.

Step 2. In selected triple the initial vertex is post-condition vertex, second – method vertex, terminal – precondition vertex. The method that turns precondition into post-condition is determined by the unambiguous correspondence to the vertex formulas of the post-condition and precondition of the method's axiom. Vertices are connected with edges: precondition to method and method to post-condition. If a precondition of the method axiom consists of combining or intersecting several elements, a vertex of data is inserted between the vertex of precondition and the method, which groups the precondition according to the relevant rules of inference. If there are two vertices in the terminal tripe, the inference rules was used to convert the vertex clause, and the vertex of the transformation in this branch of the tree occupies the vertex of method in this branch, corresponding to the inference rule used, with the corresponding vertices of post-condition.

Step 3. The vertices, edges and their ratios are introduced into *G*.

Step 4. Having extracted the processed triple from the set of triple vertices T, we execute step 2. If the set is empty, the tree reconstruction is completed.

The restored decision tree is fed to the actuator input. Implementation of a particular method specified in the solution tree is represented by a construct that describes the input data (entities, connections, relations between them), the preconditions of the method, the post-conditions of the method, and the output data. The branches of the solution tree that are after the data splitting vertex can be executed in parallel until they reach the data merging vertex, where, after completing all the parallel branches involved in merging, continue to be executed consistently.

In the report, the use of a logical approach on the real scientific task of calculating the component of life safety and indicating critical values of threat indicators for analyzing the constant development of the regions of the Ukraine has been demonstrated.

VII. Conclusions

Based on the analysis of existing data centers, a qualitative solution that provides a simple and flexible way of integrating heterogeneous information systems and their services into the World Data System has been offered. One of the key features of the proposed solution is the automation of the construction of the actions chain, which executes the user's request. A logical formalism to describe this solution has been created, on its basis, the methods of inference and the restoring of the decision tree have been developed.

An analysis of the available technologies for distributed systems implementation made it possible to use this set of software solutions for the practical implementation of this solution: UDDI for creating a register of services entered into the system; WSDL for a unified description of services; SOAP for exchanging notifications between services; BPEL to implement the overall coordination of services. Intelligent agents can be implemented using JADE.

The implementation of the proposed solution has provided an opportunity to use all integrated computing capacities and data storage systems of the World Data System in a complex manner. In this way, users can easily access all the necessary resources and services that are available to the system.

Further research is aimed at developing more efficient inference methods including ontology, developing intelligent data analysis methods.

Acknowledgment

Presented results of the research, which was carried out under the theme No. E-3/627/2016/DS, were funded by the subsidies on science granted by Polish Ministry of Science and Higher Education.

References

[1] M. Z. Zgurovsky, A. D. Gvishiani, K. V. Yefremov and A. M. Pasichny, "Integration of the Ukrainian science into the world data system," *Cybernetics and Systems Analysis*, vol. 46, issue 2, pp. 211-219, 2010.

[2] O. Pavlov, S. Telenyk, *Algorithmization and IT in Management*, Kyiv: Technics, 2002, 320 p.

[3] M. Kavis, *Architecturing the Cloud*, Wiley, 2014, 199 p.

[4] A. Y. Levy, "Logic-based techniques in data integration," in: *Logic Based Artificial Intelligence*, edited by J. Minker, Kluwer Publishers, 2000.

[5] N. Shakhovska, et al., "Data space architecture for big data managering," in *Proceedings of the Xth International Scientific and Technical Conference "Computer Sciences and Information Technologies" (CSIT)*, 2015, pp. 184-187.

[6] D. Wage, "Cloud computing: will open services be useful?," *Open Systems*, no. 1, pp. 68-75.

[7] *OMG Systems Modeling Language*, [Online]. Available: http://www.omg.org/docs/formal/07-09-01.pdf

[8] R. Queiroz, et all, *Functional Interpretation of Logical Deduction, Advances in Logic*, v. 5. Singapore: World Scientific, 2012.

[9] J. Greer, *Web Services Description Language: 55 Most Asked Questions: What You Need to Know*, Emereo Publishing, 2014.

[10] S. Graham, et all, *Building Web Services with Java: Making Sense of XML, SOAP, WSDL, and UDDI (2nd Edition)*, Sams Publishing; 2 edition, 2004.

[11] J. Laznik, Y. Mannari, R. Dhruv, *BPEL and Java Cookbook: Over 100 Recipes to Help You Enhance Your SOA Composite Applications with Java and BPEL*, Birmingham, 2013.

[12] A. T. Guler, C. Waaijer, et all, "Automating bibliometric analyses using Taverna scientific workflows: A tutorial on integrating Web Services," in *Journal of Informetrics*, 2016.

[13] S. El-Seoud, H. El-Sofany, M. Abdelfattah, M. Reham, "Big data and cloud computing: trends and challenges," *International Journal of Interactive Mobile Technologies*, vol. 11, issue 2, pp. 34-52, 2017.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 539–545 DOI: 10.15439/2019F71 ISSN 2300-5963 ACSIS, Vol. 18

The Approach to Applications Integration for World Data Center Interdisciplinary Scientific Investigations

Grzegorz Nowakowski
Department of Automatic Control and Information Technology, Faculty of Electrical and Computer Engineering, Cracow University of Technology
Cracow, Poland
gnowakowski@pk.edu.pl

Sergii Telenyk
Department of Theoretical Electrical Engineering and Computer Science, Faculty of Electrical and Computer Engineering, Cracow University of Technology Cracow, Poland
stelenyk@pk.edu.pl

Kostiantyn Yefremov
World Data Center for Geoinformatics and Sustainable Development, Kyiv, Ukraine
k.yefremov@wdc.org.ua

Volodymyr Khmeliuk
Department of Automation and Control in Technical Systems National Technical University of Ukraine "Igor Sikorsky Kyiv Politechnic Institute" Kyiv, Ukraine
hmelyuk@gmail.com

***Abstract*—The approach to applications integration for World Data Center (WDC) interdisciplinary scientific investigations is developed in the article. The integration is based on mathematical logic and artificial intelligence. Key elements of the approach - a multilevel system architecture, formal logical system, implementation – are based on intelligent agents interaction. The formal logical system is proposed. The inference method and mechanism of solution tree recovery are elaborated. The implementation of application integration for interdisciplinary scientific research is based on a stack of modern protocols, enabling communication of business processes over the transport layer of the OSI model. Application integration is also based on coordinated models of business processes, for which an integrated set of business applications are designed and realized.**

***Index Terms*—research, application integration, business processes, mathematical logic, formal logic, inference mechanisms, multi-agent systems, protocols, software agents**

I. Introduction: Particularities of WDC Application Integration

IN VIEW of the globalization of the economy and social life, National Science must integrate into the world and European organizations that promote the consolidation of research and consequently the development of scientific activity [2]. This is a very important process since there is an urgent need for interdisciplinary research, primarily for the assurance of sustainable development globally and regionally [3]. However, effective implementation of interdisciplinary research requires the creation of appropriate conditions for information exchange in the process of solving scientific problems. The scientific and technical progress that in its time facilitated the creation of information and communication technologies (ICT) nowadays benefits greatly from them. They are developing rapidly, covering new spheres of human activity and enhancing performance. Yet only field specialists can use ICT in a rational way, whereas the need for efficient information exchange within interdisciplinary research can only be met through rational ICT [19] application by specialists with deep knowledge in their areas of expertise [4].

The development of the information technologies (IT) domain that is experiencing qualitative changes related to ICT strengthening has created conditions for distributed computation and the efficient use of information and other resources. Consolidation of resources and the introduction of virtualization technologies are contributing to the process of substituting local solutions with distributed ones that allow the comprehensive use of all computing powers and data storage systems linked into a global network, thus granting access to accumulated information resources. The service approach formed on the basis of communication services has spread to the infrastructure, software development tools, and applications. The emergence of a wide range of new types of services, especially content-based ones, has led to the convergence of services and the formation of a generalized concept of information and communication services (ICS). The number of ICS providers has grown rapidly and convergent providers have emerged. The wide functionality, high quality, and moderate price of new services rendered by providers allow businesses to abandon the development of in- house or IT infrastructure and to use a wide variety of available ICS for component-based design of their information and telecommunication systems (ITS).

However, unified access to services is becoming a condition for the efficient use of the advantages of distributed systems and the possibilities of service-oriented technologies. At the same time, the formation of a new democratic IT environment, in which even small businesses can render services, has naturally been accompanied by the use of various tools and access technologies [5]. Therefore, historically the IT environment is heterogeneous, hence user access to its resources is to some extent complicated or at least inconvenient.

The same situation is characteristic of scientific activity in particular. For example, the World Data Centers (WDC) system created in 1956 under the aegis of the International Council for Science (ICSU) ensures collection, storage, circulation, and analysis of data obtained in various science areas [2]. During its existence, the WDC system has accumulated a lot of data and applications that may be used to solve the challenging problems of social development. They are one of the most powerful information resources used by hundreds of thousands of scientists, and the demand for it is increasing in proportion to the need for interdisciplinary research related to sustainable development, the solution of

Presented results of the research, which was carried out under the theme No. E-3/586/2018/DS, were funded by the subsidies on science granted by Polish Ministry of Science and Higher Education.

urgent environmental protection issues, etc. However, problems related to the incompatibility of legacy applications caused by architecture differences, the variety of data presentation formats, and other factors prevent the effective use of such WDC resources.

Promising architectural solutions are being developed and gradually implemented in the ICT field, such as Next Generation Network (NGN) and Next Generation Service Overlay Network (NGSON) [6], that ensure the interaction of various transport layer technologies. Yet there is a need for a comprehensive integration of resources from various sources, and ICT developers' efforts should seek to enable scientists working in various domains to use the accumulated resources based on their areas of expertise, and not on the IT particularities. The sources, comprehensive access to which it is reasonable to ensure, include databases, websites and portals, various legacy file management systems, and data repositories structured according to various models.

Nowadays there exist more than 50 WDCs that for more than 50 years have created a system for data accumulation, analysis, processing, and international exchange. WDCs' powerful data storage systems retain huge volumes of astronomical, geophysical and other scientific data. WDCs' servers process this data using numerous and various applications powered by different technologies.

Certainly, from the point of view of scientists working on various resource-intensive problems, it is important to have access not to a large disordered system of possibilities, but to an integral complex of data and application sources intended to meet their specific needs, with a user-friendly interface that does not require any special IT knowledge. Nevertheless, the system for WDC data accumulation, analysis, processing, and international exchange does not provide such access. Furthermore, it was not designed for the growing level of scientific society's requirements and is not versatile enough to be used in interdisciplinary research. Therefore, a new interdisciplinary structure was created in 2008 — the World Data System (WDS) — to develop and implement a new coordinated global approach to scientific data, which guarantees omni-purpose equal access to quality data for research, education, and decision making. The new structure will have to solve the accumulated tasks, primarily the unification of formats and data transfer protocols, assurance of convenient access to data, and the organization of scientific data quality control [2].

Integrating formerly independent systems for the accumulation, storage, and processing of WDC data on a new advanced integration basis will allow a considerable enhancement of their overall efficiency. The creation of such a system will provide scientists with convenient centralized access to formerly separate resources, facilitating and quickening scientific and research activity around the world.

II. Related Work

An overview of existing integration solutions in all the mentioned aspects, starting with the integration technology aspect has been done. The existing solutions are powered by technologies for the creation, functioning, and development of distributed systems. Notwithstanding the differences of various technologies for the creation of distributed service-oriented systems, the overall principles and theoretical and methodological approaches are always similar. Independent services should be registered, described, and provided with the possibility to communicate transparently with clients and with each other. Furthermore, networking interaction requires us to determine protocols for all levels of the OSI model. To do so, the corporate ITS standard structure with services intellectualization operations may be used, as suggested in [8], taking into account international, state, and branch standards, and corporate documents, first of all [9]. The structure consists of two parts, where the first covers the traditional four levels of standards of the TCP/IP protocols stack, and the second covers the user applications in accordance with the ITS class, destination, and services intellectualization operations. Figure 1 presents an example of the corporate standard structure.

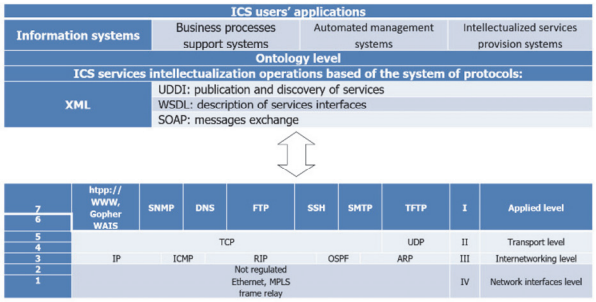


Fig. 1 Stack of services interaction protocols

For convenience, the figure presents the levels of the international standard for interaction of open OSI systems and their correlation with the corresponding levels of the TCP/IP protocols stack. The ITS classification used is proposed in [8]. It allows the systematization of various ICS by the level of users' requirements and queries, practical needs, and professional training, with no limitations imposed on the ITSs' functionality, attributes, or operations. The changes introduced apply to the first part of the structure and to the detailing of protocols that power the intellectualized interaction of applications in the course of solving users' more complex problems.

The services registry is maintained by the Universal Description, Discovery and Integration (UDDI) technology [10] that allows both people and client-programs to publish information on services and search for a required service.

The Web Service Description Language (WSDL) that, according to the W3C definition, constitutes an XML format for the description of networking services as a set of operations working with document- or procedure-oriented information through messages [11], is used for the unified description of services, which allows them to be used independently from the programming language. WSDL documents, by virtually creating a unified layer that allows the use of services created on the basis of various platforms, describes the service interface, URL, communication mechanisms 'understood' by the service, methods provided by it with corresponding parameters (type, name, location of the service Listener), and the service messages structure.

To implement the key part of the interaction—messages exchange—one of the widespread technologies may be used, and the selection should be determined by compliance with the requirements of the system for data accumulation, processing, and exchange:

SOAP: a user-friendly technology that is easy to use with the Business Process Execution Language (BPEL), which ensures interaction of distributed systems irrespective of the object model, operating system, or programming language. Data is transferred as special-format XML documents [12];
CORBA: a mechanism created to support the development and deployment of complex object-oriented applied systems, which is used for the integration of isolated systems, allowing programs that are developed in different programming languages and working in different network nodes to interact with the same ease as if they were in the address space of a single process. Such interaction is ensured through unified construction of their interfaces using a special-purpose declarative Interface Definition Language (IDL). At the same time, the interface and the description thereof do not depend on the operating systems or the processor architecture either [13];
REST (Representational State Transfer) is a style of software architecture for distributed systems, which is used to create web services. A global URL identifier unambiguously identifies each unit of information in an unvarying format. Data is transferred without any layers [14];
WCF (Windows Communication Foundation) is Microsoft's platform designed to create applications that exchange data through the network and are independent from style and protocol [15].

Although nowadays the most acceptable choice seems to be SOAP, based on which the interaction presented in Fig. 1 is performed, all the technologies allow the integration of services of various origin [16] – [17].

To take into consideration service-based, process-functional, and component-based approaches to the design and maintenance of users' solutions, the following well-known architectural means of organizing shared functionality are used:

Web Service Choreography is the approach that determines the protocols of web services' interaction to perform a single global task by performing parts thereof. The role assumed by the service determines its model of exchanging messages with other services. This method shows high efficiency for small tasks, but with increased complexity of tasks the number of services involved grows rapidly, solutions become too massive, and the efficiency drops quickly [18].

Based on the abovementioned technologies, a few solutions have been worked out, which may be used to create applied systems for the accumulation, processing, and exchange of scientific data in different areas. The best known among them are the ESIMO and GEOSS systems. They are quite widespread, although they have a number of drawbacks in terms of processing heterogeneous information.

One of the most crucial issues is that of developing mathematical models, methods, and means for the integration of various applications, both legacy ones based on traditional technologies and client-server and web-oriented technologies. The article proposes an approach to applications integration using the mathematical logic instrument and artificial intelligence theory for interdisciplinary research through the example of the WDC system functioning. The specificity of the applications integration for interdisciplinary research is that the coordination of business processes models is not required because it is, in fact, substituted with schemes for performing users' tasks. In general, the applications integration is performed on the basis of coordinated business processes models, and the integrated complex of business applications is intended to support them.

III. Definition Of The Application Integration Problem For Interdisciprinar Reasarch

Mathematical models and methods as the basis for a holistic solution should be created to power data storage and processing centres , which will provide users with versatile possibilities at the level of information integration and servers availability. The mentioned models and methods should be devoid of the drawbacks characteristic of the algorithmic approach, related to the need to reprogram data processing algorithms upon the emergence of new data types and changes in the implemented algorithms or the emergence of new ones. At the data processing level, such a solution should ensure:

- computations distribution and the use of remote hardware resources;
- versatility and adaptation to the system load;
- system usage space;
- data supply by remote client or service;
- availability of intellectual data processing means directly to the end user with no special knowledge or skills;
- fast and simple integration of data and applications of various global information systems

IV. The Appication Integration System Architecture

The principal tenet for the creation of a distributed system is the method of organizing services interaction. Of the two known main ways of services interaction - orchestration and choreography - the more efficient for the WDC systems is the former. Indeed, orchestration that aggregates basic services into hierarchically integrated systems, subordinating them to administrators – 'orchestrator' services - allows services to be unified by attributes that are convenient to WDC (science area, functionality, regional location, etc.), and to provide orchestrator services with powers in accordance with the international data exchange policies. Thus, it is possible to line up a simple and effective system, in which the search for the necessary services or the construction of their composition for the complex queries inherent in interdisciplinary research will not require a lot of time, as every Orchestrator Service has information on the functionality of inferior services, and in addition they can coordinate their possibilities in the process of the planning and execution of users' queries. Plugging new services into the system will not present a particular problem either, as doing so will only require the service description to be laid down, entered into the registry, and assigned to a certain orchestrator service.

Such an approach will work when users know exactly their information needs and have corresponding knowledge and skills to compile chains of queries to the known orchestrators. What is more important is that the approach will provide the opportunity to work with the system for users who cannot initiate services, by determining the sequence of their work and specifying execution and interaction parameters. Users only have to know how to formulate their needs in terms of a particular subject domain. In this case, the association of services and the organization of their cooperation require an intellectual constituent. For this purpose, intellectual agents are used, which constitute the system core, implementing its functioning logic, which promotes the formation of queries, plans their execution, and organizes basic services interaction. The introduction of intellectual agents as orchestrator services into the hierarchical structure forms a two-tier system; the bottom level consists of services performing basic tasks, the top level consists of intellectual agents that orchestrate basic services. By using the registries of subordinated basic services and their functionality and by interacting with each other, interconnected intellectual agents implement methods of logical inference. The inference result is the composition of basic agents' operations that allow user-defined tasks to be performed. This operations composition, or proof, is transferred to the lower level, where the operations necessary to solve the user's task are performed. At this point, control is handed over to the lower level agents that only return the final result to be sent to the user.

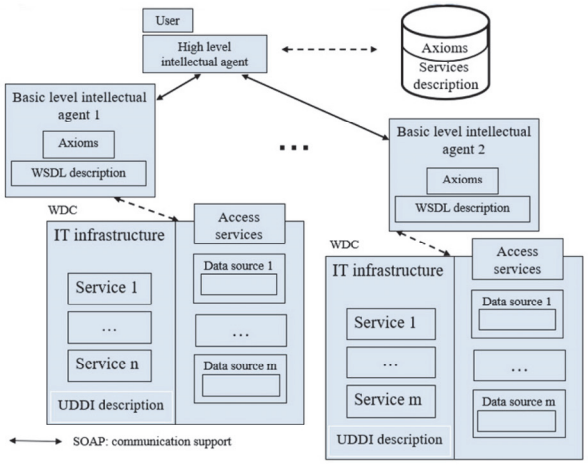


Fig. 2 System components interaction scheme

The user's query itself is performed upon formation of the proof containing references to one or several pointers to data sources and IDs of methods that the system services should apply to the data. Some queries do not require pre-processing of data, since the execution of services methods is sufficient upon receipt of information from the data source. Others require a certain sequence of prior data operations, and as a result, the execution of other services methods. The necessary connections are described in the axioms entered into the knowledge base upon registration of respective services in the system. These operations yield the final product of the system - data that are the solution to the user's problem. Since the system operation is determined by the user's query, the system interface should help them make a query in familiar terms.

The proposed solution uses UDDI to create the service registries and WSDL for the unified description thereof. The key part—messaging—is implemented using SOAP, and services orchestration—using BPEL. Agents are described by means of JADE. The system components interaction scheme in presented in Fig. 2.

The description of the interface (WSDL) of every low-level service registered in the system and of related axioms is input into the system knowledge base and directly into the knowledge base of the Controller Agent to which the new service will be subordinated. The Controller Agent is selected in accordance with the policies adopted in the WDC system, for example, by the criteria of the geographical location of the deployed agent, in order to minimize data exchange.

At the agent level, the orchestrator agent (or several), having received the user's query, interrogates the agents in order to search for the necessary services and available resources. The user's query is processed taking into account the services of the system, their functionality (described by the axioms of the system), and the inference rules defined by logical formalism. The resulting proof is transferred to the lower level to be implemented.

At the services level, each agent invokes the required services from its set and sends processed data according to the action chains. This process is carried out in accordance with the task solution tree reconstructed by the solution tree reconstruction mechanism based on the proof.

To realize this interaction of services in the WDC system, the most appropriate is the logical approach, which allows both the level of the abovementioned requirements to be reached and the drawbacks of a traditional algorithmic approach to be eliminated. Indeed, it is only required to develop the inference method and the solution tree reconstruction mechanism that will implement the query formation processes, plan their execution, reconstruct and implement the task solution scheme. The logical approach is the most appropriate one to create and describe these constituents. The logical approach implementation requires one to:

- describe the existing applications and their functional capacities in the formal language;
- formulate the inference rules;
- determine the inference method;
- develop an algorithm of the user's query execution tree reconstruction based on the proof;
- implement these methods and mechanisms in the agents of the system.

V. The Formal Logical System

Let us describe the formalism upon which the program system that will ensure solution of the formulated problem will be built. We will take the first order clausal logic for a base and describe the formal system language in accordance with the structural elements determined in [3].

Symbols: service: (,), [,], {, }, :, <, >, ,;

constant:

1) *individual, of primary types* (int, real, char, bool) - a_1^1, a_2^1, …, a_1^2, a_2^2, ... where each constant a_i^k pertains to type (primary type) *k; structural type* (construct) - c_1, c_2, ...; procedural type (method) - d_1, d_2, ...; objective type (problem, entity, relation) — e_1, e_2, ...;
2) *functional i-place,* for individuals of type *k* - h_1^1, h_2^1, …, h_1^2, h_2^2, ...;
3) *predicate i-place,* for individuals of type *k* - A_1^1, A_2^1, …, A_1^2, A_2^2, ... (this class includes taxonomic, relational and other predicates, as well as traditional relations, at least equality = and order ≥);

variable: for individuals of type *k* - x_1^1, x_2^1, …, x_1^2, x_2^2, …, where every variable x_i^k pertains to type *k;*

logical: $\neg, \wedge, \vee, \leftarrow, \exists, \forall, \Leftrightarrow$

Individual terms of type *k:*

1) each individual constant a_i^k of type *k* is an individual term of type *k;*
2) each free variable x_i^k for individuals of type *k* is an individual term of type *k*;
3) if h_i^j is a certain functional constant for individuals of type *k* and τ_1, …, τ_j are terms for individuals of type *k,* then h_i^j (τ_1, …, τ_j) is an individual term of type *k;*
4) there are no other individual terms of type *k*.

The terms obtained by applying construction rules 1 or 2 of the definition will be called primary, and all the others—complex.

Formulas for individuals:

1) if A_i^j is a predicate constant for individuals and τ_1, …, τ_j are terms for them, then A_i^j (τ_1, …, τ_j) is the atomic formula for individuals;
2) the atomic formula for individuals is the formula for them;
3) there are no other formulas for individuals.

Hereinafter we consider that the system contains an omni-purpose transformer of formulas into the traditional for the clausal form (we are talking about the Horn clauses) view with a single → symbol, atomic formulas to its left and right, and an implicit quantifier $\forall$.

Specifiers, preconditions, post-conditions, specifiers of methods, specifiers of problems, clause, the system's knowledge and inference rules are presented in detail in [1].

VI. The Inference Method

The method proposed in [2] based on analogy and types of assertions has been used. A detailed description of this method can be found in [1], [4], [7].

The inference mechanism work algorithm is presented in Fig. 3.

To improve the inference mechanism three known elements were integrated:

1) a multiset of literals (atomic formulas) (which will be called a multi clause (m-clause));
2) the ordered linear;
3) typification abstraction (to manage an ordered linear proof).

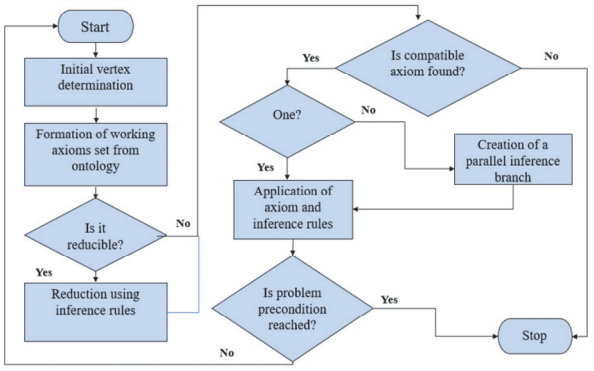


Fig. 3 Inference mechanism work algorithm

VII. The Mechanism Of Solution Tree Reconstruction

To obtain from the proof generated by the inference mechanism a functional sequence of actions to be used by our system, taking account of services features and nature, the solution scheme reconstruction mechanism should be initiated [1].

Condition: proof Result = <V, T>. **To find:** $G =< V, E, \Theta >$	
Step	
1	Identification of the terminal triple—the triple that contains vertex k that is neither the first nor the second component of any of the triples of set T.
2	In the selected triple, the initial vertex is the post-condition vertex, the second is the method vertex, and the terminal is the precondition vertex. The method that turns the precondition into the post-condition is determined by the unambiguous correspondence to the vertex formulas of the post-condition and precondition of the method axiom. The vertices are connected with the edges: precondition to method and method to post-condition. If the precondition of the method axiom consists of a combination or intersection of several elements, a data vertex is inserted between the precondition and the method vertices, which groups the preconditions according to the relevant inference rule. If the terminal triple only contains two vertices, it means that the inference rule was used to convert the vertex clauses, and the place of the method vertex in this tree branch is occupied by the transformation data vertex that corresponds to the inference rule used, with the corresponding post-condition vertices.
3	The vertices, edges, and their correspondence are introduced into G.
4	Having extracted the processed triple from the set of triple vertices T, we repeat step 2. If the set is empty, the tree reconstruction is completed.

Fig. 4 Algorithm for vertex scheme reconstruction

The solution scheme constitutes a connected directed graph with no oriented cycles with parallel directed paths from the root to the vertices; it has three types of vertices, and is specified by triple $G =< V, E, \Theta >$, where $V = V_1 \cup V_2 \cup V_3$, V_1 is a set of method vertices, V_2 is a

set of precondition vertices and post-condition vertices, V_3 is a set of data vertices in which data is merged or split; E is a set of edges; Θ is a subset of Cartesian product E×V×V, which determines the correspondence of edges to pairs of vertices. The scheme determines the sequence and the correspondence according to the data of actions to be performed by the data processing system's executive mechanisms to obtain the result desired by the user [1].

Algorithm for vertex scheme reconstruction was presented on Fig. 4.

The reconstructed solution tree is fed to the actuator input. The implementation of a particular method specified in the solution tree is represented by a construct that describes the input data (entities, connections, relations between them), method preconditions, method post-conditions, and output data. The solution tree branches downstream of the data-splitting vertex can be executed in parallel until they reach the data-merging vertex, where, after completing all the parallel branches involved in merging, they continue to be executed consecutively. The work algorithm of the solution tree reconstruction mechanism is presented in Fig. 5.

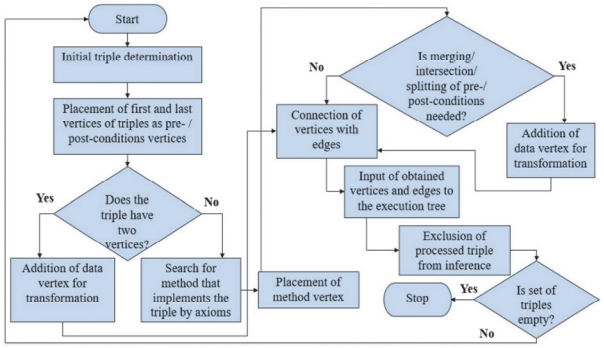


Fig. 5 Work algorithm of solution tree reconstruction mechanism

VIII. Application Of Logical Approach To Problem Solution

The approach efficiency on the real scientific task of calculating the component of life safety and indicating critical values of threat indicators for analyzing the sustainable development of the regions of Ukraine has been demonstrated.

The life safety component formula: $Csl = \sqrt[3]{\sum_0^n Threat^3}$, where the threat value is normalized by formulas (2) or (3). The purpose of indicating critical values of threat indicators is to determine the priority in consideration thereof in the decision-making process on the level of a single region and the entire country in order to mitigate the impact of threats on sustainable development.

Suppose that for every administrative unit $i = \overline{1, n}$ there is a set of values $\langle x_{i,1}, x_{i,2}, \ldots, x_{i,m} \rangle$ of indicators X_j, $j = \overline{1, m}$, which characterize the negative impact of certain phenomena on the sustainable development processes in the economic, social, and ecological spheres. Such indicators, whose essence and composition are determined by experts, will be called threat indicators.

Given the content of the critical values indication problem, the following characteristic function must be determined:

$$\Psi(x_{i,j}) = \begin{cases} 0, if\ value\ of\ x\ is\ not\ critical \\ 1, if\ otherwise \end{cases} \quad (1)$$

where $i = \overline{1, n}$, $j = \overline{1, m}$.

It is clear that the determination of function $\Psi(x_{i,j})$ must be based upon certain criteria that take into consideration exceeding by $x_{i,j}$ a hazardous limit, the relative position of region i in the indicators rating X_j compiled for the comparison groups and for the entire country, and the degree of "hazard" of value $x_{i,j}$ in comparison to values of other indicators for region i.

To account for the relative position of the region in the entire country's ratings, the following criterion is used:

$$R_{i,j} = \left(1 + e^{\frac{a - x_{i,j}}{b}}\right)^{-1} \quad (2)$$

if higher values of indicator X_j correspond to a higher impact of the respective threat on the sustainable development, and:

$$R_{i,j} = 1 - \left(1 + e^{\frac{a - x_{i,j}}{b}}\right)^{-1} \quad (3)$$

if lower values of indicator X_j correspond to a higher impact. In formulas (2)-(3), parameters a and b are calculated by the following formulas:

$$a = \overline{X_j} = \frac{1}{n}\sum_{i=1}^{n} x_{i,j},\ b = \sigma(X_j) = \sqrt{\frac{\sum_{i=1}^{n}(x_{i,j} - \overline{X_j})^2}{n}} \quad (4)$$

Criterion $R_{i,j}$ is a dimensionless number that assumes values within $[0, 1]$. Values of around 0.5 correspond to average values of X_j in the selection, and values higher than 0.75 correspond to values that exceed the average ones by more than a standard deviation. In this case, the characteristic function (1) taking account of one criterion $R_{i,j}$ may be expressed as follows:

$$\Psi_R(x_{i,j}) = \begin{cases} 0, R_{i,j} < 0,75; \\ 1, R_{i,j} \geq 0,75. \end{cases}$$

Criterion $P_{i,j}$ that takes into account a region's relative position in the comparison group may be calculated by formulas (2)-(3) taking account of the fact that parameters a

and b are calculated by formula (4) independently for each comparison group.

Criteria $R_{i,j}$ and $P_{i,j}$ are dimensionless numbers and are of similar nature and can therefore be aggregated through a weighted sum:

$$K_{i,j} = w_R R_{i,j} + w_P P_{i,j};\ w_R + w_P = 1,$$

where weighting factors w_R and w_P are determined by experts.

Thus, for region $i = \overline{1,n}$ we have a set $\langle K_{i,1}, K_{i,2}, \ldots, K_{i,m} \rangle$ of values of the aggregated criterion that accounts for a region's relative position in ratings compiled for comparison groups and for the entire country. Now, among values $K_{i,j}$, $j = \overline{1,m}$, the worst must be determined, which may also be performed by formula (2):

$$I_{i,j} = \left(1 + e^{\frac{a - K_{i,j}}{b}}\right)^{-1}$$

where parameters a and b are calculated in selection $K_{i,j}$, $j = \overline{1,m}$.

For values of criterion $I_{i,j}$ the same remarks apply as for criterion $R_{i,j}$. Therefore, characteristic function (1) may be expressed as follows:

$$\Psi_I(x_{i,j}) = \begin{cases} 0, I_{i,j} < 0,75; \\ 1, I_{i,j} \geq 0,75. \end{cases}$$

Thus, values $\Psi_I(x_{i,j}) = 1$ correspond to the highest priority of attention to be paid to the value of indicator X_j in the administrative decision-making process on the level of single region *i*.

IX. Conclusions

Based on the analysis of existing data centres, their equipment and software, a high quality solution is offered that provides a simple and flexible way to integrate heterogeneous information systems and their services into the World Data System. One of the key features of the proposed solution is the automation of thealgorit construction of the actions sequence that executes users' queries. A logical formalism has been created to describe this solution, and on its basis an inference method and a solution tree reconstruction mechanism have been developed.

The analysis of available technologies used for the implementation of distributed systems allowed the use of such a set of software solutions for practical implementation of this solution: UDDI for creating a registry of services entered into the system; WSDL for a unified description of services; SOAP for exchanging notifications between services; BPEL for the overall coordination of services. Intellectual agents can be implemented using JADE.

The implementation of the proposed solution will provide an opportunity to use all the integrated computing capacities and data storage systems of the World Data System in a comprehensive manner. Thus, users will be able to easily gain access to all the necessary resources and services available to the system.

References

[1] S. Telenyk, G. Nowakowski, K. Yefremov and V. Khmeliuk, "Logics based application integration for interdisciplinary scientific investigations", 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), pp. 1026-1031, Bucharest, 2017. DOI: 10.1109/IDAACS.2017.8095241

[2] M. Z. Zgurovsky, A. D. Gvishiani, K. V. Yefremov and A. M. Pasichny, "Integration of the Ukrainian science into the world data system", Cybernetics and Systems Analysis: Volume 46, Issue 2 (2010), pp. 211-219. DOI: 10.1007/s10559-010-9199-9

[3] M. Z. Zgurovsky, A. O. Boldak, K. V. Yefremov and others, "Analysis of Sustainable Development – Global and Regional Contexts", International Council for Science (ICSU) and others. – K.: NTUU «KPI». – Part 2. Ukraine in Sustainable Development Indicators (2011-2012). – 232 p, 2012.

[4] O. Pavlov, S. Telenyk. *Algorithmization and IT in management*, Kyiv: Technics, 2002 – 320 p.

[5] Data Integration Information, quick view on world of data, (online) homepage at: https://www.dataintegration.info/

[6] M. Ulema et al., "Next generation service overlay networks (NGSON)", JIEEE Communications Magazine 50(1):52-53, 2012, DOI: 10.1109/MCOM.2012.6122532.

[7] A.Y. Levy, "Logic-Based Techniques in Data Integration", In: Logic Based Artificial Intelligence, Edited by J. Minker. Kluwer Publishers, 2000.

[8] P. P. Maslianko, "Fundamentals of the methodology of system design of information and communication systems", Naukovi Visti NTUU "KPI," No. 6, 54–60 (2007).

[9] OMG Systems Modeling Language, (online) homepage at: https://www.omg.org/spec/SysML/About-SysML/

[10] UDDI Version 3.0.2 Specification, (online) homepage at: http://uddi.org/pubs/uddi_v3.htm

[11] J. Greer, "Web Services Description Language: 55 Most Asked Questions: What You Need to Know", Emereo Publishing, 2014

[12] SOAP Version 1.2 Part 1: Messaging Framework (Second Edition), (online) homepage at: https://www.w3.org/TR/soap12-part1/

[13] About the common object request broker architecture specification version 3.3, (online) homepage: https://www.omg.org/spec/CORBA

[14] G. Nowakowski, "Rest Api safety assurance by means of HMAC mechanism", Information Systems in Management, Vol. 5, No. 3, pp. 358-369, 2016.

[15] Windows Communication Foundation Architecture Overview, (online) homepage at: http://msdn.microsoft.com/en-us/library/ aa480210.aspx

[16] S. Graham, et all, "Building Web Services with Java: Making Sense of XML, SOAP, WSDL, and UDDI (2nd Edition)", Sams Publishing; 2 edition, 2004

[17] S. El-Seoud, H. El-Sofany, M. Abdelfattah, M. Reham, "Big Data and Cloud Computing: Trends and Challenges", International Journal of Interactive Mobile Technologies, 2017, Vol. 11 Issue 2, p34-52, 19p, 3 Diagrams; DOI: 10.3991/ijim.v11i2.6561

[18] J. Laznik, Y. Mannari, R. Dhruv, BPEL and Java Cookbook: Over 100 Recipes to Help You Enhance Your SOA Composite Applications with Java and BPEL, Birmingham, 2013

[19] E. Ziemba, "The ICT Adoption in Government Units in the Context of the Sustainable Information Society", 2018 Federated Conference on Computer Science and Information Systems, pp.725–733. DOI: 10.15439/2018F116

Simple and Flexible Way to Integrate Heterogeneous Information Systems and Their Services into the World Data System

Submitted: 8th October 2021; accepted: 17th March 2022

Grzegorz Nowakowski, Sergii Telenyk, Kostiantyn Yefremov, Volodymyr Khmeliuk

DOI: 10.14313/JAMRIS/4-2021/29

Abstract:
The approach to applications integration for World Data Center (WDC) interdisciplinary scientific investigations is developed in the article. The integration is based on mathematical logic and artificial intelligence. Key elements of the approach – a multilevel system architecture, formal logical system, implementation – are based on intelligent agents interaction. The formal logical system is proposed. The inference method and mechanism of solution tree recovery are elaborated. The implementation of application integration for interdisciplinary scientific research is based on a stack of modern protocols, enabling communication of business processes over the transport layer of the OSI model. Application integration is also based on coordinated models of business processes, for which an integrated set of business applications are designed and realized.

Keywords: *research; application integration; business processes; mathematical logic; formal logic; inference mechanisms; multi-agent systems; protocols; software agents*

1. Introduction

In view of the globalization of the economy and social life, National Science must integrate into the world and European organizations that promote the consolidation of research and consequently the development of scientific activity [27]. This is a very important process since there is an urgent need for interdisciplinary research, primarily for the assurance of sustainable development globally and regionally [26]. However, effective implementation of interdisciplinary research requires the creation of appropriate conditions for information exchange in the process of solving scientific problems. The scientific and technical progress that in its time facilitated the creation of information and communication technologies (ICT) nowadays benefits greatly from them. They are developing rapidly, covering new spheres of human activity and enhancing performance. Yet only field specialists can use ICT in a rational way, whereas the need for efficient information exchange within interdisciplinary research can only be met through rational ICT application by specialists with deep knowledge in their areas of expertise [19].

The development of the information technologies (IT) domain that is experiencing qualitative changes related to ICT strengthening has created conditions for distributed computation and the efficient use of information and other resources. Consolidation of resources and the introduction of virtualization technologies are contributing to the process of substituting local solutions with distributed ones that allow the comprehensive use of all computing powers and data storage systems linked into a global network, thus granting access to accumulated information resources. The service approach formed on the basis of communication services has spread to the infrastructure, software development tools, and applications. The emergence of a wide range of new types of services, especially content-based ones, has led to the convergence of services and the formation of a generalized concept of information and communication services (ICS). The number of ICS providers has grown rapidly and convergent providers have emerged. The wide functionality, high quality, and moderate price of new services rendered by providers allow businesses to abandon the development of in-house or IT infrastructure and to use a wide variety of available ICS for component-based design of their information and telecommunication systems (S).

However, unified access to services is becoming a condition for the efficient use of the advantages of distributed systems and the possibilities of service-oriented technologies. At the same time, the formation of a new democratic IT environment, in which even small businesses can render services, has naturally been accompanied by the use of various tools and access technologies [19]. Therefore, historically the IT environment is heterogeneous, hence user access to its resources is to some extent complicated or at least inconvenient.

The same situation is characteristic of scientific activity in particular. For example, the World Data Centers (WDC) system created in 1956 under the aegis of the International Council for Science (ICSU) ensures collection, storage, circulation, and analysis of data obtained in various science areas [22]. During its existence, the WDC system has accumulated a lot of data and applications that may be used to solve the challenging problems of social development. They are one of the most powerful information resources used by hundreds of thousands of scientists, and the demand for it is increasing in proportion to the need for interdisciplinary research related to sustainable development, the solution of urgent environmental protection issues, etc. However, problems related to

the incompatibility of legacy applications caused by architecture differences, the variety of data presentation formats, and other factors prevent the effective use of such WDC resources.

Promising architectural solutions are being developed and gradually implemented in the ICT field, such as Next Generation Network (NGN) and Next Generation Service Overlay Network (NGSON) [24], that ensure the interaction of various transport layer technologies. Yet there is a need for a comprehensive integration of resources from various sources, and ICT developers' efforts should seek to enable scientists working in various domains to use the accumulated resources based on their areas of expertise, and not on the IT particularities. The sources, comprehensive access to which it is reasonable to ensure, include databases [16], websites and portals, various legacy file management systems, and data repositories structured according to various models.

Nowadays there exist more than 50 WDCs that for more than 50 years have created a system for data accumulation, analysis, processing, and international exchange. WDCs' powerful data storage systems retain huge volumes of astronomical, geophysical and other scientific data. Certainly, from the point of view of scientists working on various resource-intensive problems, it is important to have access not to a large disordered system of possibilities, but to an integral complex of data and application sources intended to meet their specific needs, with a user-friendly interface that does not require any special IT knowledge. Nevertheless, the system for WDC data accumulation, analysis, processing, and international exchange does not provide such access. Furthermore, it was not designed for the growing level of scientific society's requirements and is not versatile enough to be used in interdisciplinary research. Therefore, a new interdisciplinary structure was created in 2008 – the World Data System (WDS) – to develop and implement a new coordinated global approach to scientific data, which guarantees omni-purpose equal access to quality data for research, education, and decision making. The new structure will have to solve the accumulated tasks, primarily the unification of formats and data transfer protocols, assurance of convenient access to data, and the organization of scientific data quality control [27].

The related difficulties are emerging because the existing WDCs are accumulating and providing heterogeneous data determined by the specificity of a corresponding science area and country. The possibilities of telecommunication systems are constantly growing, as well as the possibilities of modern computers and data storage, thus making way for new distributed computation technologies – grid systems and cloud-based computation. Integrating formerly independent systems for the accumulation, storage, and processing of WDC data on a new advanced integration basis will allow a considerable enhancement of their overall efficiency. The creation of such a system will provide scientists with convenient centralized access to formerly separate resources, facilitating and quickening scientific and research activity around the world.

To implement such integration solutions that support the continuous lifecycle of research data, it is proposed to apply the well-proven Continuous Integration (CI) and Delivery (CD) software development practices.) [28], [29], [30]. Continuous Integration and Continuous Delivery are often quoted as the essential ingredients of successful DevOps. DevOps is a software development approach which bridges the gap between development and operations teams by automating build, test and deployment of applications. It is implemented using the CI/CD pipeline. DevOps lifecycle contains six phases: Plan, Build, Continuous Integration, Deploy, Operate and Continuous Feedback. Throughout each phase, teams collaborate and communicate to maintain alignment, velocity, and quality.

CI/CD are required for software development using an Agile methodology, which recommends using automated testing to fast debug working software. Automated testing gives stakeholders access to newly created features and provides fast feedback.

The use of CI/CD practices requires the fulfillment of a set of basic requirements for a development project.

In particular, the source code and everything needed to build and test the project should be stored in a repository of the version control system, and operations for copying from the repository, building and testing the whole project should be automated and easily called from external programs. The GitLab environment allows to manage project repositories, document functionality and results of improvements and tests, as well as track errors and work with the CI/CD pipeline [31], [32].

2. Related Work

In this paper, the authors would primarily like to focus on the consideration of urgent among the WDC problems requiring efficient solutions is that of integration. Presently, numerous WDCs use legacy technologies based on program systems with rigid structures, whereas newer solutions have for a long time been based on client- server and web-oriented technologies. The service approach to data processing and analysis has not spread to the required extent. Sometimes connection to geoinformatics systems and even to basic public services such as Google Maps is unavailable. Combined with the lack of a unified standard, this fact shows that the need for data sources integration, including applications integration and databases [16] integration, may be considered sufficiently substantiated. Therewith, the formation of a modern IT environment envisages the following integration aspects:

- technology: integrated data and services should be based on a single stack of interaction protocols, using common data transfer formats in order to render the integration
- possible;
- content: a single data description format in the semantic structure of data sources should be supported;

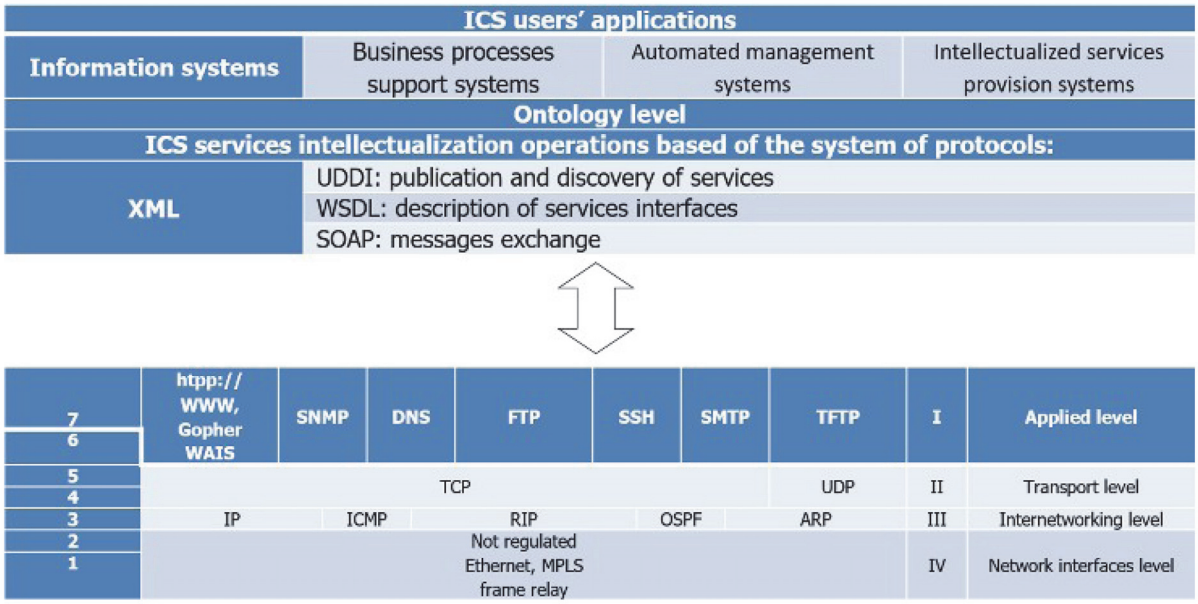


Fig. 1. Stack of services interaction protocols

- functionality: the use of different services in the single structure to solve user-defined problems.

An overview of existing integration solutions in all the mentioned aspects, starting with the integration technology aspect has been done. The existing solutions are powered by technologies for the creation, functioning, and development of distributed systems. Notwithstanding the differences of various technologies for the creation of distributed service-oriented systems, the overall principles and theoretical and methodological approaches are always similar. Independent services should be registered, described, and provided with the possibility to communicate transparently with clients and with each other. Furthermore, networking interaction requires us to determine protocols for all levels of the OSI model. To do so, the corporate ITS standard structure with services intellectualization operations may be used, as suggested in [15], taking into account international, state, and branch standards, and corporate documents, first of all [4]. The structure consists of two parts, where the first covers the traditional four levels of standards of the TCP/IP protocols stack, and the second covers the user applications in accordance with the ITS class, destination, and services intellectualization operations. Fig. 1 presents an example of the corporate standard structure.

For convenience, the figure presents the levels of the international standard for interaction of open OSI systems and their correlation with the corresponding levels of the TCP/IP protocols stack. The ITS classification used is proposed in [15]. It allows the systematization of various ICS by the level of users' requirements and queries, practical needs, and professional training, with no limitations imposed on the ITSs' functionality, attributes, or operations. The changes introduced apply to the first part of the structure and to the detailing of protocols that power the intellectualized interaction of applications in the course of solving users' more complex problems.

The services registry is maintained by the Universal Description, Discovery and Integration (UDDI) technology [6] that allows both people and client-programs to publish information on services and search for a required service.

The Web Service Description Language (WSDL) that, according to the W3C definition, constitutes an XML format for the description of networking services as a set of operations working with document- or procedure-oriented information through messages [11], is used for the unified description of services, which allows them to be used independently from the programming language. WSDL documents, by virtually creating a unified layer that allows the use of services created on the basis of various platforms, describes the service interface, URL, communication mechanisms "understood" by the service, methods provided by it with corresponding parameters (type, name, location of the service Listener), and the service messages structure. Let us mention that the majority of platforms support the formation of WSDL descriptions of services, enabling developers to work with services as simple classes.

To implement the key part of the interaction -messages exchange – one of the widespread technologies may be used, and the selection should be determined by compliance with the requirements of the system for data accumulation, processing, and exchange:

- SOAP: a user-friendly technology that is easy to use with the Business Process Execution Language (BPEL), which ensures interaction of distributed systems irrespective of the object model, operating system, or programming language. Data is transferred as special-format XML documents [5];
- CORBA: a mechanism created to support the development and deployment of complex object-oriented applied systems, which is used for the integration of isolated systems, allowing programs that are developed in different programming lan-

guages and working in different network nodes to interact with the same ease as if they were in the address space of a single process. Such interaction is ensured through unified construction of their interfaces using a special-purpose declarative Interface Definition Language (IDL). At the same time, the interface and the description thereof do not depend on the operating systems or the processor architecture either [1];

- REST (Representational State Transfer) is a style of software architecture for distributed systems, which is used to create web services. A global URL identifier unambiguously identifies each unit of information in an unvarying format. Data is transferred without any layers [17];
- WCF (Windows Communication Foundation) is Microsoft's platform designed to create applications that exchange data through the network and are independent from style and protocol [7].

Although nowadays the most acceptable choice seems to be SOAP, based on which the interaction presented in Figure 1 is performed, all the technologies allow the integration of services of various origin [9,10].

To take into consideration service-based, process-functional, and component-based approaches to the design and maintenance of users' solutions, the following well-known architectural means of organizing shared functionality are used:

Web Service Choreography is the approach that determines the protocols of web services' interaction to perform a single global task by performing parts thereof. The role assumed by the service determines its model of exchanging messages with other services. This method shows high efficiency for small tasks, but with increased complexity of tasks the number of services involved grows rapidly, solutions become too massive, and the efficiency drops quickly [13].

Service orchestration is the approach that uses an orchestrator service to associate services into a single multifunctional business service. The role of orchestrator services may be performed by agents introduced into the system, which allows its hierarchical structure to be maintained, thus combining the advantages of orchestration and multi-agent systems (MAS). To perform orchestration, a special-purpose language was developed to describe the business processes and the protocols of their interaction – BPEL. This language, sometimes called the orchestrating language, provides an orchestrator process (service) to coordinate the work of services subordinated to it. The coordinated interaction is enabled through messages exchange between web services. The problem lies in the fact that, in order to carry out long-term business processes with complex structure, not only messages themselves are important, but also the exchange sequence, account of state, results, time and other aspects of business processes. If the work of basic web services is organized by a program agent as a higher-level orchestrator service, we get the management mechanism at the web services level, whereas organizational aspects (messaging sequence, account of load, etc.) will be dealt with by higher-level agents. As a rule, SOAP is used along with BPEL to implement messages exchange, although there are alternatives [2].

The scientific community is already using systems based on the decomposition of the overall problem (flow) into a sequence of individual subtasks (subflows), the so-called workflow system. The effective interdisciplinary workflow environments worth mentioning are Triana Workflows and Kepler. A researcher user may edit the task-performing scripts, using system automation means, which renders it versatile and widens its application sphere. Those systems' architecture consists of the following levels:

1) client (allows a user to quickly develop new workflows using the graphic script editor);
2) middleware (its flow server ensures workflows performance, authorization and coordination with the heterogeneous resources involved);
3) resource (provides resources to perform separate subflows).

A workflow constitutes a separate ready-to-use component. Components may be parts of other components. Due to rigid typification (internal formats determined for flows and components) and the formation of metadata for each component, which describe the types of their input and output data, flows and components are verified to make sure their input and output data are compatible. These systems were widespread due to their easy-to-use graphic user interface, extensive library of standard components developed for various disciplines, effective script performance manager, and means of monitoring.

Taverna Workflow Management System, a system for the development, design, and performance of complex workflows, which allows the use of different services such as WSDL, BioMoby, SoapLab and other types to be coordinated, has also been widely used lately [8].

Alas, such systems are incompatible at the levels of both the performance manager and the description of flows and components. Furthermore, they lack an intellectual constituent, or such a constituent is not sufficiently developed, and therefore any working sequence of services needs to be built up manually or edited, which substantially limits these systems' versatility and application sphere.

Based on the abovementioned technologies, a few solutions have been worked out, which may be used to create applied systems for the accumulation, processing, and exchange of scientific data in different areas. The best known among them are the ESIMO and GEOSS systems. They are quite widespread, although they have a number of draw- backs in terms of processing heterogeneous information:

- difficulties in compiling the association of heterogeneous types of data from different sources in one query;
- difficulties in enhancing the functionality and degree of automation, caused by drawbacks in these systems' architecture, which does not envisage the possibility of forming a single network;

- no unified metadata model for all the metadata objects;
- no unified standard for data presentation and processing;
- users need special expertise, knowledge, and skills to form a query.

Based on this overview, we can draw a substantiated conclusion regarding the possibility of using the developed approaches, models, technologies, and means for integration of applications, primarily:

- schemes for the implementation of applications interaction based on the stack of protocols UDDI, SOAP (or REST), BPEL, WSDL, XML above the transport level;
- technology of WCF type;
- approach to the integration of services through an orchestrator service into a single multifunctional business service.

Nonetheless, for user-formulated tasks, the level of automation of services interaction schemes should be enhanced. Ultimately, this is a question of developing efficient models and methods of applications integration, based on which a new state-of-the-art applications integration system may be created, capable of providing users with the abovementioned advantages.

Secondly, let us dwell briefly on the content aspect. As a rule, data integration in information systems is viewed as the provision of a single unified interface to access a certain number of generally heterogeneous independent data sources [3,14]. For a user, information re- sources from all the integrated sources should be presented as a new single source, and the system providing such possibilities is called the data integration system [12]. There are several data integration approaches, some of which are successfully implemented and quite widespread in real systems. The analytical review thereof is presented in [12,20]. The data space concept viewed as a new abstraction of data management has been steadily changing its positions lately. In Ukraine and Poland, it was developed in works by N. Shakhovska, for instance [21].

Finally, let us move on to the third aspect of integration. The view on integration as the interaction of applications is not as widespread, yet is extremely important in the con- text of resolving the comprehensive integration issue in the new IT environment. It was mostly formed within the community of specialists in programming automation, notably in relation to the implementation of promising technologies for industrial software production, such as "component-based programming". In terms of counteracting program systems complexity, the focus was transferred from developing versatile, adjustable, and scalable structures, such as application packages, to developing means for the interaction of independent program systems.

The specific systematic research of the application integration began at the turn of the millennium, when the issue of the integration of numerous business applications, primarily of CRM, ERP, SSM classes, etc., abruptly emerged. Under the new economic conditions, business charges IT with new tasks, the fulfillment of which through the purchase of new and the modernization of available business applications does not ensure a return on investment any more. At that very moment, the need arose to build enterprises' information systems based on the integrated complex of business applications, and the first attempts were made at the formalized description of business applications and their interaction with the aim of reaching the business divisions' objectives. Thus emerged the Component Object Model (COM) and Common Object Request Broker Architecture (CORBA) technologies that ensured programs interaction on a single computer, and subsequently such possibilities were applied to the interaction of program systems located on different computers. The abovementioned possibilities implemented in the Distributed COM (DCOM), Electronic Access Interchange (EAI), and CORBA technologies ensured independent program systems integration to solve users' problems that are more complex.

Another solution that is important for applications integration was the Service-oriented Architecture (SOA), whose emergence gradually, through the integration of multi-bases and federated databases, large data storages, and various repositories, promoted the transition of the integration concept onto web applications. Nonetheless, a total renunciation by IT departments of solving efficacy issues through the integration of business applications into the existing IT infrastructure with minimal losses only became possible after the 2008 crisis, with the emergence and implementation of cloud-based computation technologies. At that very moment, there arose the obvious necessity to transparently see the IT infrastructure of an enterprise's entire information system and the means of its architecting, operation, and development, with an account of the new possibilities of the IT environment that was being intensely formed on the basis of consolidation, virtualization, and multi-services.

The industry leaders proposed a solution based on integration platforms ensuring complete functionality through advanced means for scaling and adjustment. However, such a centralized approach requires a respective IT infrastructure and not every enterprise can afford this. Furthermore, it entails a certain dependence of the enterprise on companies producing such integration solutions.

Therefore, the majority of enterprises are more interested in the integration approach related to the wide use of business applications by various developers through proxies, such as Message Oriented Middleware (MOM) programs, combined with modern approaches to cost-cutting in IT-infrastructure ownership, primarily the means of using legacy applications. Another obstacle in the way of implementing the decentralized approach is differences in the architectural components of business applications by different producers, notably in the models of data and business processes, technologies of designing the program system and its basic elements (DBMS serv-

ers and applications), component cooperation means, users' access, etc.

Cloud-based computation allows the provision of software services of a determined quality for a fair price, thus justifying the practicability of software (SaaS), infrastructure (IaaS) and platform (PaaS) services [25]. Nowadays, this activity may be viewed as a theoretical and methodological basis for the migration of the service approach to information systems creation and the wide introduction of component-based development thereof.

In order for legacy applications to become part of a service-based system, their data models and business processes must be in a certain way coordinated with those supported by the new systems, and interaction with the program system's basic elements must be ensured. Therefore, data integration in information systems and integration as the coordination of applications have to be combined. The necessity arises to develop a technical solution that will allow:

- creating mechanisms for efficient information resources management;
- widening areas of protocols standardization and data exchange formats based on the interaction of commodity systems;
- creating cross-platform applications, integrating data and applications;
- facilitating systems deployment and development by concealing software and hardware internal details.

To create solutions that provide the above possibilities, it is required to:

1) develop formal models and methods to supply metadata in order to describe sources, their functionality, and particularities of access;
2) develop models and methods for the automated design and implementation of unified-architecture applications that are capable of efficient interaction in order to reach the objectives of interdisciplinary research, and on the basis thereof to create platform solutions according to the PaaS model;
3) develop models, methods, and means to implement applications interaction at the semantic level;
4) develop models and methods to facilitate users' queries for required data using the terms from topical area(s) of research that are convenient for them;
5) bring the IT infrastructure of the WDC system into compliance with the needs through consolidation, virtualization, and multi-services powered by cloud-based computation models and technologies;
6) develop models, methods, and means for efficient administration of the IT infrastructure of the WDC system, capable of ensuring the determined level of information services quality.

One of the most crucial issues is that of developing mathematical models, methods, and means for the integration of various applications, both legacy ones based on traditional technologies and client-server and web-oriented technologies. The article proposes an approach to applications integration using the mathematical logic instrument and artificial intelligence theory for interdisciplinary research through the example of the WDC system functioning. The specificity of the applications integration for interdisciplinary research is that the coordination of business processes models is not required because it is, in fact, substituted with schemes for performing users' tasks. In general, the applications integration is performed on the basis of coordinated business processes models, and the integrated complex of business applications is intended to support them.

3. Definition of the Application Integration Problem for Interdisciplinary Research

Mathematical models and methods as the basis for a holistic solution should be created to power data storage and processing centres, which will provide users with versatile possibilities at the level of information integration and servers availability. The mentioned models and methods should be devoid of the drawbacks characteristic of the algorithmic approach, related to the need to reprogram data processing algorithms upon the emergence of new data types and changes in the implemented algorithms or the emergence of new ones. At the data processing level, such a solution should ensure [18,22]:

- computations distribution and the use of remote hardware resources;
- versatility and adaptation to the system load;
- system usage space;
- data supply by remote client or service;
- availability of intellectual data processing means directly to the end user with no special knowledge or skills;
- fast and simple integration of data and applications of various global information systems.

4. The Applications Integration System Architecture

The principal tenet for the creation of a distributed system is the method of organizing services interaction. Of the two known main ways of services interaction – orchestration and choreography – the more efficient for the WDC systems is the former. Indeed, orchestration that aggregates basic services into hierarchically integrated systems, subordinating them to administrators – "orchestrator" services – allows services to be unified by attributes that are convenient to WDC (science area, functionality, regional location, etc.), and to provide orchestrator services with powers in accordance with the international data exchange policies. Thus, it is possible to line up a simple and effective system, in which the search for the necessary services or the construction of their composition for the complex queries inherent in interdisciplinary research will not require a lot of time, as every Orchestrator Service has information on the functionality of inferior services, and in addition they

can coordinate their possibilities in the process of the planning and execution of users' queries. Plugging new services into the system will not present a particular problem either, as doing so will only require the service description to be laid down, entered into the registry, and assigned to a certain orchestrator service [18,22].

Such an approach will work when users know exactly their information needs and have corresponding knowledge and skills to compile chains of queries to the known orchestrators. What is more important is that the approach will provide the opportunity to work with the system for users who cannot initiate services, by determining the sequence of their work and specifying execution and interaction parameters. Users only have to know how to formulate their needs in terms of a particular subject domain. In this case, the association of services and the organization of their cooperation require an intellectual constituent. For this purpose, intellectual agents are used, which constitute the system core, implementing its functioning logic, which promotes the formation of queries, plans their execution, and organizes basic services interaction.

The introduction of intellectual agents as orchestrator services into the hierarchical structure forms a two-tier system; the bottom level consists of services performing basic tasks, the top level consists of intellectual agents that orchestrate basic services. By using the registries of subordinated basic services and their functionality and by interacting with each other, interconnected intellectual agents implement methods of logical inference. The inference result is the composition of basic agents' operations that allow user-defined tasks to be performed. This operations composition, or proof, is transferred to the lower level, where the operations necessary to solve the user's task are performed. At this point, control is handed over to the lower-level agents that only return the final result to be sent to the user.

The user's query itself is performed upon formation of the proof containing references to one or several pointers to data sources and IDs of methods that the system services should apply to the data. Some queries do not require pre-processing of data, since the execution of services methods is sufficient upon receipt of information from the data source. Others require a certain sequence of prior data operations, and as a result, the execution of other services methods. The necessary connections are described in the axioms entered into the knowledge base upon registration of respective services in the system. These operations yield the final product of the system – data that are the solution to the user's problem. Since the system operation is determined by the user's query, the system interface should help them make a query in familiar terms.

The proposed solution uses UDDI to create the service registries and WSDL for the unified description thereof. The key part – messaging – is implemented using SOAP, and services orchestration – using BPEL. Agents are described by means of JADE. The system components interaction scheme in presented in Fig. 2.

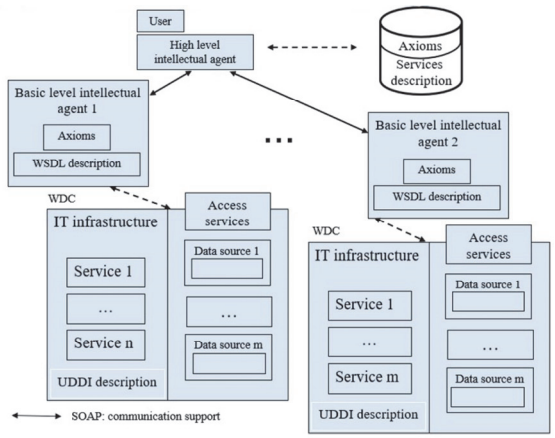


Fig. 2. System components interaction scheme

The description of the interface (WSDL) of every low-level service registered in the system and of related axioms is input into the system knowledge base and directly into the knowledge base of the Controller Agent to which the new service will be subordinated. The Controller Agent is selected in accordance with the policies adopted in the WDC system, for example, by the criteria of the geographical location of the deployed agent, in order to minimize data exchange.

At the agent level, the orchestrator agent (or several), having received the user's query, interrogates the agents in order to search for the necessary services and available resources. The user's query is processed taking into account the services of the system, their functionality (described by the axioms of the system), and the inference rules defined by logical formalism. The resulting proof is transferred to the lower level to be implemented.

At the services level, each agent invokes the required services from its set and sends processed data according to the action chains. This process is carried out in accordance with the task solution tree reconstructed by the solution tree reconstruction mechanism based on the proof.

To realize this interaction of services in the WDC system, the most appropriate is the logical approach, which allows both the level of the abovementioned requirements to be reached and the drawbacks of a traditional algorithmic approach to be eliminated. Indeed, it is only required to develop the inference method and the solution tree reconstruction mechanism that will implement the query formation processes, plan their execution, reconstruct and implement the task solution scheme. The logical approach is the most appropriate one to create and describe these constituents. The logical approach implementation requires one to:

- describe the existing applications and their functional capacities in the formal language;
- formulate the inference rules;
- determine the inference method;
- develop an algorithm of the user's query execution tree reconstruction based on the proof;
- implement these methods and mechanisms in the agents of the system.

5. The Formal Logical System

The principal tenet for the creation of a distributed system is the method of organizing services interaction [18,22,23].

Let us describe the formalism upon which the program system that will ensure solution of the formulated problem will be built. We will take the first order clausal logic for a base and describe the formal system language in accordance with the structural elements determined in [27].

Symbols:

service: (,), [,], {, }, :, <, >, ,;

constant:

1. *individual, of primary types* (int, real, char, bool) – a_1^1, a_2^1, ..., a_1^2, a_2^2, ... where each constant a_i^k pertains to type (primary type) *k; structural type* (construct) – c_1, c_2, ...; procedural type (method) – d_1, d_2, ...; objective type (problem, entity, relation) — e_1, e_2, ...;
2. *functional i-place*, for individuals of type *k* – h_1^1, h_2^1, ..., h_1^2, h_2^2, ...;
3. *predicate i-place,* for individuals of type *k* – A_1^1, A_2^1, ..., A_1^2, A_2^2, ... (this class includes taxonomic, relational and other predicates, as well as traditional relations, at least equality = and order ≥);

variable: for individuals of type *k* – x_1^1, x_2^1, ..., x_1^2, x_2^2, ..., where every variable x_i^k pertains to type *k;*

logical: $\neg, \wedge, \vee, \leftarrow, \exists, \forall, \Leftrightarrow$

Individual terms of type *k:*

1. each individual constant a_i^k of type *k* is an individual term of type *k*
2. each free variable x_i^k for individuals of type *k* is an individual term of type *k*;
3. if h_i^j is a certain functional constant for individuals of type *k* and τ_1, ..., τ_j are terms for individuals of type *k*, then h_i^j (τ_1, ..., τ_j) is an individual term of type *k;*
4. there are no other individual terms of type *k*.

The terms obtained by applying construction rules 1 or 2 of the definition will be called primary, and all the others—complex.

Formulas for individuals:

1. if A_i^j is a predicate constant for individuals and τ_1, ..., τ_j are terms for them, then A_i^j (τ_1, ..., τ_j) is the atomic formula for individuals;
2. the atomic formula for individuals is the formula for them;
3. there are no other formulas for individuals.

Hereinafter we consider that the system contains an omni-purpose transformer of formulas into the traditional for the clausal form (we are talking about the Horn clauses) view with a single → symbol, atomic formulas to its left and right, and an implicit quantifier $\forall$.

Specifiers are constructions of type τ1, where τ1 is a term for an individual object. The construct specifiers are:

if $e_1^e ... e_i^e$ are individual terms of type entity, and $e_1^r ... e_i^r$ are individual terms of type relation, and $A_1^j(a_1^1 ... a_j^1) ... A_i^j(a_1^k ... a_j^k)$ are atomic formulas for the individuals of primary types, and τ is an individual term of type construct, then τ: $(e_1^e ... e_i^e, e_1^r ... e_i^r, A_1^j(a_1^1 ... a_j^1) ... A_i^j(a_1^k ... a_j^k))$ is a construct specifier.

4. there are no other construct specifiers.

Preconditions:

1) if c_1 is a specifier of construct, and Π is a sequence of atomic formulas, then $<\tau_1:(c_1,\Pi)>$ is the elementary precondition;
2) elementary precondition – precondition;
3) if $<\tau_1>$ is a precondition, and τ_2 is an elementary precondition, then $<\tau_1, \tau_2>$ is the precondition;
4) there are no other preconditions.

Post-conditions:

1) if τ_1 is a construct specifier, and Π is a sequence of atomic formulas, then $<\tau_1:(c_1,\Pi)>$ is the elementary post-condition;
2) elementary post-condition – post-condition;
3) if $<\tau_1>$ is a post-condition, and a τ_2 is an elementary post-condition, then $<\tau_1, \tau_2>$ is the post-condition;
4) there are no other post-conditions.

Specifiers of methods:

1) if τ is an individual term of type method, and $<\tau_1>$ is a precondition, and $<\tau_2>$ is a post-condition, then τ: ($<\tau_1>$, $<\tau_2>$) is a method specifier;
2) the arity by constructs of method precondition must not be lower than the arity by constructs of method post-condition;
3) there are no other method specifiers.

Specifiers of problems:

1) if $<\tau_1>$ is a precondition, and $<\tau_2>$ is a post-condition, and τ is a term for an individual object of type Problem, then τ: $<<\tau_1>,<\tau_2>>$ is the problem specifier;
2) there are no other problem specifiers.

Clause is an expression of type $\Pi \rightarrow \Lambda$, where Π is a sequence of atomic formulas; Λ is a single atomic formula. Depending on their type, atomic formulas will be distributed into the following clauses:

1) individual (contains only atomic formulas for individuals);
2) typified (contains only atomic formulas for types);
3) general type (contains atomic formulas for individuals and types).

The system's knowledge base consists of the ontology of axioms depicting methods of system services and the ontology of data sources described in the OWL language based on RDF. The knowledge base also includes the inference rules necessary to obtain

the desired result when it is required to combine methods of a single service or of different services [18, 22, 23].

Inference rules:

1) If $\mathbf{d}_1$: ($<\boldsymbol{\tau}_1>$,$<\boldsymbol{\tau}_2>$) and $\mathbf{d}_2$: ($<\boldsymbol{\tau}_3>$,$<\boldsymbol{\tau}_1>$) then $\mathbf{d}_1$: ($<\mathbf{d}_2>$,$<\boldsymbol{\tau}_2>$)
2) If $\mathbf{d}_1$: ($<\boldsymbol{\tau}_1>$,$<\boldsymbol{\tau}_2>$) and $\mathbf{d}_2$: ($<\boldsymbol{\tau}_3 \wedge \tau_4>$,$<\boldsymbol{\tau}_1>$) then $\mathbf{d}_1$: ($<\mathbf{d}_2>$,$<\boldsymbol{\tau}_2>$)
3) If $\mathbf{d}_1$: ($<\boldsymbol{\tau}_1 \wedge \tau_2>$,$<\boldsymbol{\tau}_3>$) and $\mathbf{d}_2$: ($<\boldsymbol{\tau}_4>$,$<\boldsymbol{\tau}_1>$) and $\mathbf{d}_3$: ($<\boldsymbol{\tau}_5>$,$<\boldsymbol{\tau}_2>$) then $\mathbf{d}_1$: ($<\mathbf{d}_2 \wedge d_3>$,$<\boldsymbol{\tau}_3>$)
4) If $\mathbf{d}_1$: ($<\boldsymbol{\tau}_1>$,$<\boldsymbol{\tau}_2>$) and $\mathbf{d}_2$: ($<\boldsymbol{\tau}_3 \vee \tau_4>$,$<\boldsymbol{\tau}_1>$) then $\mathbf{d}_1$: ($<\mathbf{d}_2>$,$<\boldsymbol{\tau}_2>$)
5) If $\mathbf{d}_1$: ($<\boldsymbol{\tau}_2 \wedge \tau_2>$,$<\boldsymbol{\tau}_1>$) then $\mathbf{d}_1$: ($<\boldsymbol{\tau}_2>$,$<\boldsymbol{\tau}_1>$)
6) If $\mathbf{d}_1$: ($<\boldsymbol{\tau}_1,\boldsymbol{\tau}_2>$,$<\boldsymbol{\tau}_3>$) and $\mathbf{d}_2$: ($<\boldsymbol{\tau}_4>$,$<\boldsymbol{\tau}_1>$) and $\mathbf{d}_3$: ($<\boldsymbol{\tau}_5>$,$<\boldsymbol{\tau}_2>$) then $\mathbf{d}_1$: ($<\mathbf{d}_2, \mathbf{d}_3>$,$<\boldsymbol{\tau}_2>$)
7) If $\mathbf{d}_1$: ($<\boldsymbol{\tau}_1>$,$< \boldsymbol{\tau}_2,\boldsymbol{\tau}_3>$) and $\mathbf{d}_2$: ($<\boldsymbol{\tau}_2>$,$<\boldsymbol{\tau}_4>$) and $\mathbf{d}_3$: ($<\boldsymbol{\tau}_3>$,$<\boldsymbol{\tau}_5>$) then $\mathbf{d}_2$: ($<\mathbf{d}_1>$,$<\boldsymbol{\tau}_4>$) and $\mathbf{d}_3$: ($<\mathbf{d}_1>$, $<\boldsymbol{\tau}_5>$)

6. The Inference Method

We use the method proposed in [27] based on analogy and types of assertions. Its idea is to build a proof in abstract space and subsequently use it to control the inference process in the initial solution search space. This will increase the inference effectiveness by cutting off most of the unpromising proof branches in the initial space.

Hereinafter a multiset of literals (atomic formulas) will be called a multiclause (m-clause), and literal L will be recorded in the m-clause as many times as it is repeated. An ordinary clause will be considered as an m-clause in which every literal's multiplicity is 1. The operations $\cup$ (unification), $\cap$ (intersection), — (difference), · (concatenation) and relation $\subseteq$ (occurrence) for multisets are naturally performed.

Suppose $A_1 \in C_1$, $A_2 \in C_2$ and α_1, α_2 are such substitutions that some literal L *on the right side of* A_1 and on the left side of A_2, $A_1\alpha_1=\{L\}$, $A_2\alpha_2=\{L\}$, and if $|A_i|>1$, i = 1, 2, the corresponding substitution α_i is the most common unifier of literals from A_1. *Then the* clause *obtained from the* unification *of* clauses $C_1\alpha_1$ and $C_2\alpha_2$ *by removing L of the left and right parts is called the m-resolvent of m-*clauses C_1 *and* C_2. Suppose all the m-clauses in the original set S are ordered, and C_1, C_2 are two of them. Suppose $A_1 \in C_1$, $A_2 \in C_2$ and α_1, α_2 are such substitutions that for some literal L *on the right side of* A_1 and on the left side of A_2, $A_1\alpha_1=\{L\}$, $A_2\alpha_2=\{L\}$, and if $|A_i|>1$, i = 1, 2, the corresponding substitution α_i is the most common unifier of the literals from A_i. *Then* the *clause* obtained from the *unification* of *clauses* $C_1\alpha_1$ *and* $C_2\alpha_2$ by removing L of the right part *and the last L of the left part* and by eliminating any non-*underlined* literal that is not followed by any other literal, *is called* the ordered linear m-resolvent of ordered m-*clauses* C_1 and C_2. If the last literal on the left side of the ordered m-*clause* is unified with the underlined literal on the right side of the same *clause*, the ordered linear m-resolvent can be obtained by removing the last literal from the left part of the ordered m-*clause*, i.e., *m-clause* reduction.

The next pair whose first component is a set of proof vertices and the second is a set of triples of vertices (to select components, a pair of the function-selectors s-N(Tm) and s-M(Tm) will be respectively used) is called the m-resolution proof and denoted *Tm*. Each vertex of the m-resolution proof is characterized by a mark and the depth in the proof, so that if $n \in$ *s-N(Tm), then s-L(n)* is the mark of vertex *n* and *s-D(n) is its* proof depth. If $<n_1, n_2, n_3> \in$ *s-M(Tm),* then *s-L*(n_3) is the *m-resolvent of s-L*(n_1) and s-L(n_2). Each triple of such a form is called an m-resolution. If $n \in$*s-N(Tm)* and vertex *n is* not the third component of any of the triples from *s-M(Tm),* then *n is called* the initial vertex of proof *Tm. The mark of such* vertex *is the initial* m-clause. If $n \in$*s-N(Tm)* and vertex *n is neither the first nor the second* component of any of the triples from *s-M(Tm),* then n *is called* the terminal vertex of proof and its mark *is called* the terminal m-clause. In general, *Tm* must satisfy the following restriction: if $< n_1, n_2, n_3> \in$ *s-M(Tm), then* $< n_2, n_1, n_3> \in$ *s-M(Tm).* The depth (in the proof) of each vertex n, $n \in$*s-N(Tm),* by definition:

1) *s-D(n) = 0* for any initial vertex *n;*
2) *s-D(n) = 1* + min{max{s-D(n_1), s-D(n_2)}|$<n_1, n_2, n> \in$ s-M(Tm)} for any non-initial vertex *n.*

If the marks of the initial vertices *Tm* belong to the set of m-clauses *S*, then m-resolution proof *Tm* is *called the* proof from *S.* C is inferred from *S* if *Tm* is a proof from *S,* while *C is* the mark of one of the *Tm* vertices. Suppose the value of function Result() is defined for *Tm* and is equal to C, if terminal clause C of proof *Tm* is the only one, i.e. Result(Tm) = *C*. Suppose Tm_1 and Tm_2 *are m*-resolution proofs. Proof Tm_1 *is called the sub*-proof of Tm_2 (and denoted $Tm_1 \subseteq Tm_2$), if the following conditions are fulfilled: *s-N*(Tm_1) $\subseteq$ *s-N*(Tm_2) and *s-M*(Tm_1) $\subseteq$ *s-M*(Tm_2). Moreover, if all the initial vertices of Tm_1 *are* the initial vertices of Tm_2, then Tm_1 *is called the* initial *sub*-proof of Tm_2.

Suppose *Tm is the m*-resolution proof from *S and all its m*-clauses *are ordered*. If for any triple $< n_1, n_2, n_3>$ from *s-M(Tm) s-L*(n_3) is *the ordered linear m-resolvent of s-L(n1)* and *s-L*(n_2), then *Tm is called the ordered linear* proof from *S. Then, by the m*-resolution proof *definition, the following condition must be satisfied*: max{s-D(n_1), s-D(n_2)} = s-D(n_3) – 1.

In addition to the general properties of m-resolution proofs, *the ordered linear* proofs have their own features. For example, for *the ordered linear* m-resolution proof *the following conditions are fair*:

1) if $< n_1, n_2, n_3> \subseteq$ *s-M(Tm),* then $< n_2, n_1, n_3> \notin$ *s-M(Tm);*
2) for each non-initial vertex n_3 *with proof* depth r *there exists* a triple $< n_1, n_2, n_3>$, in which n_1 has the proof depth r-1, and n_2 corresponds to the initial vertex or is absent;
3) *Tm* contains a single terminal clause.

To manage an ordered linear proof, we will use the typification abstraction proposed in [26]. Suppose *f* is such mapping from the set of *m*-clauses into the set of *m*-clauses that:

1) if *m*-clause C_3 is the *m*-resolvent of *m*-clauses C_1 and C_2, while $D_3 \in f(C_3)$, then there *exist* such

$D_1 \in f(C_1)$ and $D_2 \in f(C_2)$ that the result of the substitution of some m-resolvent D_1 and D_2 belongs to D_3;
2) $f(\varnothing)=(\varnothing)$;
3) if the result of some substitution of m-clause C_1 belongs to m-clause C_2, then for any abstraction D_2 for C_2 there exists such abstraction D_1 for C_1 that the result of D_1 substitution belongs to D_2.

Such mapping *is called f* m-abstraction mapping, while any D from *f(C) is called* m-abstraction. The typification mapping is understood as a certain mapping ϕ from a set of literals into a set of literals that reflects each atomic formula into the formula whose terms have the type closest to the basic types in the hierarchy. Work [8] determines that typification mapping φ is m-mapping.

Work [14] determines that typification mapping φ is m-mapping. A relationship between the *ordered linear* proofs *Tm* and *Um is called the* relationship of improvement (and denoted $Tm \rightarrow^{f} Um$), where f is m-abstraction mapping, if the vertices of proofs *Tm* and *Um are in such a* relationship of *accordance R* that:
1) the following conditions are fulfilled simultaneously: $\forall n(n \in s\text{–}N(Tm))\ \exists n'(n' \in s\text{–}N(Um))(nRn')$ and $\forall n(n \in s\text{–}N(Um))\ \exists n'(n' \in s\text{–}N(Tm))(nRn')$;
2) if nRn', where $n \in s\text{–}N(Tm)$, $n' \in s\text{–}N(Um)$, then n and n' may be initial or terminal vertices only simultaneously;
3) for terminal vertices *Tm* and *Um*, relationship R is a one-to-one relationship;
4) if $<n_1, n_2, n_3> \in s\text{-}M(Tm)$, $<n'_1, n'_2, n'_3> \in s\text{-}M(Um)$ and $n_3Rn'_3$, then $n_1Rn'_1$ and $n_2Rn'_2$;
5) if n and n' are initial vertices of proofs *Tm* and *Um* respectively, and nRn', then $s\text{-}L(n') \in f(s\text{-}L(n))$;
6) if n and n' are non-initial vertices of proofs *Tm and Um respectively*, and nRn', then example $s\text{-}L(n')$ belongs to *f(s-L(n)).*

So the ordered linear proofs have the property of minimality, i.e. there is exactly one terminal vertex, and if triple $<n_1, n_2, n_3>$ belongs to the minimal ordered linear deduction, then no other triple can contain vertex n_3 as the third component.

On this property of the ordered linear proofs, the authors have built the inference mechanisms implemented by agents. Once a user has formulated the problem, the system builds a proof – in fact, by checking the possibility of performing the query with existing resources. This task is performed by Orchestrator Agents as an important part of the system's intellectual kernel. With access to the list of services, their descriptions and axioms that specify services application capabilities, these agents initiate the inference process. The inference process constructs a chain of services capable of providing the user with the desired result, starting with obtaining data from relevant sources.

The ordered linear m-resolution proof is specified by a pair of vertices sets $V = \{k_1, k_2, k_3 \ldots k_n\}$ and vertices triples $T\{<k_1, k_2, k_3>, <k_3, k_4, k_5> \ldots <k_{n-2}, k_{n-1}, k_n>\}$, where k_i is the vertices of the tree, k_n is the terminal vertex, the result of the algorithm execution. The proof is formed based on the post-condition of the formulated problem as the initial proof vertices, and the compatible axiom from the knowledge base as the lateral vertex. The axiom is compatible if the sequence of atomic formulas matching the given proof vertex or any part thereof is the axiom post-condition. The third vertex of the triple is obtained by application of the axiom to the post-condition formula, taking account of the inference rules. The third vertex of the first triple becomes the first vertex of the second triple, and the process is repeated until the terminal vertex that corresponds to the formulated problem precondition is reached. If the atomic formula corresponding to the third vertex of the triple can be reduced by using one of the inference rules, it is reduced in the following triple that will only have two vertices – the first corresponding to the formula to be reduced, and the second corresponding to the reduced formula. To search for compatible axioms, the constructs of the currently processed post-condition are compared to the constructs of the ontology axioms. Thus, the set of axioms that describe services processing the objects of the desired type and format is selected from the ontology. The inference mechanism deducts and selects axioms only in this set. If, while searching for axioms for another vertex, several compatible axioms are found in the knowledge base, each of these axioms is used for further construction of their "parallel" versions of the proof. As a result, the inference mechanism can form several proof sets with various lengths and complexity, depending on the number of axioms in the knowledge base and combinations of their application to each vertex.

After the inference mechanism's work is over, the proof with the fewest vertices triples and the fewest applied axioms is selected from among the set of obtained proofs. Longer, cyclic, or dead-end proofs are discarded.

This set of formulas and clauses, corresponding to the proof triples vertices and reflecting the application of axioms and the rules of their construction, is the result of the inference mechanism work.

The inference mechanism work algorithm is presented in Figure 3.

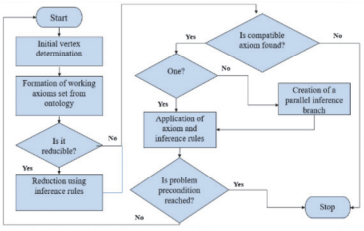


Fig. 3. Inference mechanism work algorithm

The example of inference mechanism work and solution tree reconstruction for problem P of type problem

P: <<A>,<K>>, where A and K are a precondition and a post-condition is presented in Figure 4.

Proof, a set of vertices triples: <A,$A\wedge A$>, <$A\wedge A$, $B\leftarrow A$,$B\wedge A$>, <$B\wedge A$,$C\leftarrow A$,$B\wedge C$>, <$B\wedge C$,$B\wedge C\wedge B$>, <$B\wedge C\wedge B$, $B\wedge C\wedge C\wedge B$>,<$B\wedge C\wedge C\wedge B$,$E\leftarrow B\wedge C'$,$E\wedge C\wedge B$>,<$E\wedge C\wedge B$,$D\leftarrow B$, $E\wedge C\wedge D$>, <$E\wedge C\wedge D$,$F\leftarrow E\wedge C$,$F\wedge D$>, <$F\wedge D$,$K\leftarrow F\wedge D$,K>

7. The Mechanism of Solution Tree Reconstruction

To obtain from the proof generated by the inference mechanism a functional sequence of actions to be used by our system, taking account of services features and nature, the solution scheme reconstruction mechanism should be initiated [18,22].

The solution scheme constitutes a connected directed graph with no oriented cycles with parallel directed paths from the root to the vertices; it has three types of vertices, and is specified by triple $G=<V,E,\Theta>$, where $V=V_1\cup V_2\cup V_3$, V_1 is a set of method vertices, V_2 is a set of precondition vertices and post-condition vertices, V_3 is a set of data vertices in which data is merged or split; E is a set of edges; Θ is a subset of Cartesian product E×V×V, which determines the correspondence of edges to pairs of vertices. The scheme determines the sequence and the correspondence according to the data of actions to be performed by the data processing system's executive mechanisms to obtain the result desired by the user [18,22].

Algorithm for vertex scheme reconstruction was presented on Figure 5.

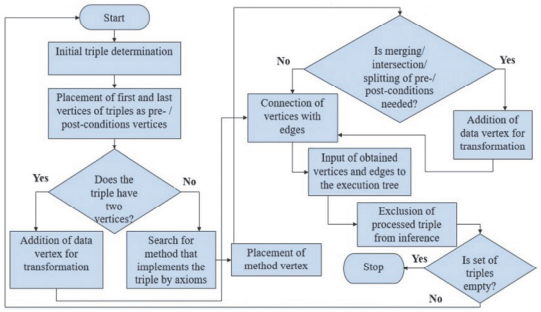


Fig. 5. Work algorithm of the solution tree reconstruction mechanism

Example of the solution tree reconstruction mechanism work for problem. Tree reconstruction by triples:

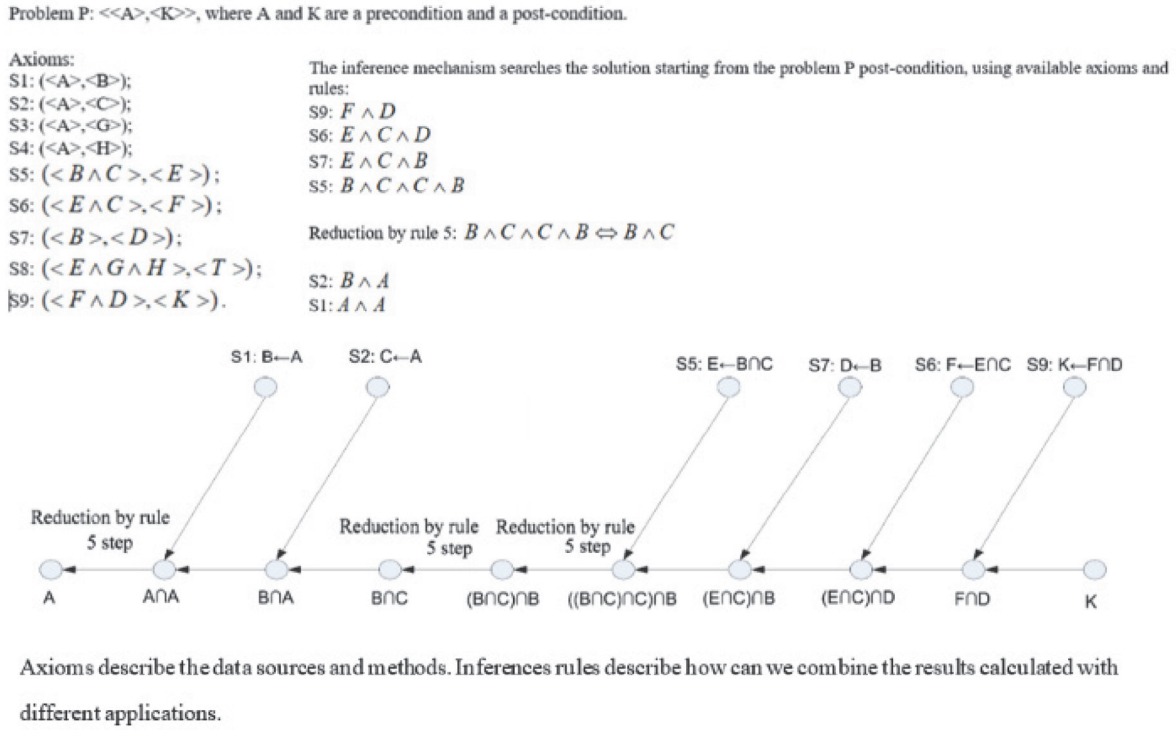


Fig. 4. The example of inference mechanism work and solution tree reconstruction

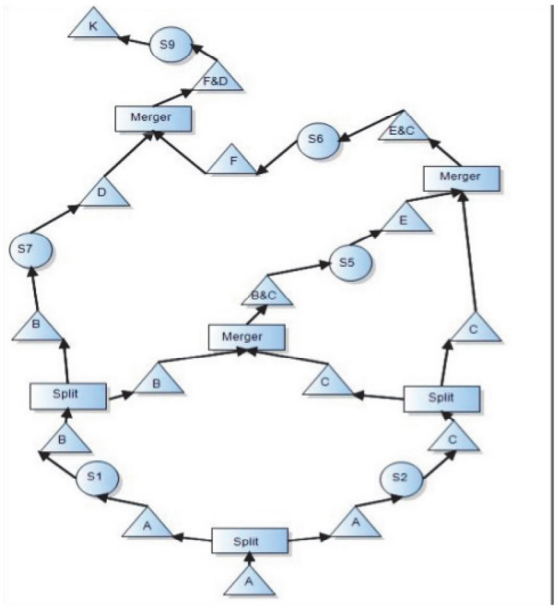


Fig. 6. The example of the solution tree reconstruction mechanism work for problem

The reconstructed solution tree is fed to the actuator input. The implementation of a particular method specified in the solution tree is represented by a construct that describes the input data (entities, connections, relations between them), method preconditions, method post-conditions, and output data. The solution tree branches downstream of the data-splitting vertex can be executed in parallel until they reach the data-merging vertex, where, after completing all the parallel branches involved in merging, they continue to be executed consecutively. The work algorithm of the solution tree reconstruction mechanism is presented in Figure 6.

The determined sequence of services performance:
S1 (<A>, <B>)
S2 (<A>, <C>)
S5 (<C∧B>, <E>)
S7 (<B>, <D>)
S6 (<E∧C>, <F>)
S9 (<F∧D>, <K>)

8. Application of Logical Approach to Problem Solution

This section is not mandatory, but can be added to the manuscript if the discussion is unusually long or complex.

The approach efficiency on the real scientific task of calculating the component of life safety and indicating critical values of threat indicators for analyzing the sustainable development of the regions of Ukraine has been demonstrated.

The life safety component formula:

$$\sqrt[3]{\sum_0^n Threat^3},$$

where the threat value is normalized by formulas (2) or (3).

The purpose of indicating critical values of threat indicators is to determine the priority in consideration thereof in the decision-making process on the level of a single region and the entire country in order to mitigate the impact of threats on sustainable development.

Suppose that for every administrative unit $i=\overline{1,n}$ there is a set of values $\langle x_{i,1},x_{i,2},...,x_{i,m}\rangle$ of indicators $X_j, j=\overline{1,m}$, which characterize the negative impact of certain phenomena on the sustainable development processes in the economic, social, and ecological spheres. Such indicators, whose essence and composition are determined by experts, will be called threat indicators.

Given the content of the critical values indication problem, the following characteristic function must be determined:

$$\Psi(x_{i,j}) = \begin{cases} 0, if\ value\ of\ x\ is\ not\ critical \\ 1, if\ otherwise \end{cases}$$

where $i=\overline{1,n},\ j=\overline{1,m}$.

It is clear that the determination of function $\Psi(x_{i,j})$ must be based upon certain criteria that take into consideration exceeding by $x_{i,j}$ a hazardous limit, the relative position of region i in the indicators rating X_j compiled for the comparison groups and for the entire country, and the degree of "hazard" of value $x_{i,j}$ in comparison to values of other indicators for region i.

To account for the relative position of the region in the entire country's ratings, the following criterion is used:

$$R_{i,j} = \left(1 + e^{\frac{a-x_{i,j}}{b}}\right)^{-1}$$

if higher values of indicator X_j correspond to a higher impact of the respective threat on the sustainable development, and:

$$R_{i,j} = 1 - \left(1 + e^{\frac{a-x_{i,j}}{b}}\right)^{-1}$$

if lower values of indicator X_j correspond to a higher impact. In formulas (2)-(3), parameters a and b are calculated by the following formulas:

$$a = \overline{X_j} = \frac{1}{n}\sum_{i=1}^{n} x_{i,j},\ b = \sigma(X_j) = \sqrt{\frac{\sum_{i=1}^{n}(x_{i,j} - \overline{X_j})^2}{n}}$$

Criterion $R_{i,j}$ is a dimensionless number that assumes values within [0,1]. Values of around 0.5 correspond to average values of X_j in the selection, and values higher than 0.75 correspond to values that exceed the average ones by more than a standard deviation. In this case, the characteristic function (1) taking account of one criterion $R_{i,j}$ may be expressed as follows:

$$\Psi_R(x_{i,j}) = \begin{cases} 0, R_{i,j} < 0,75; \\ 1, R_{i,j} \geq 0,75. \end{cases}$$

Criterion $P_{i,j}$ that takes into account a region's relative position in the comparison group may be calculated by formulas (2)-(3) taking account of the fact

that parameters a and b are calculated by formula (4) independently for each comparison group.

Criteria $R_{i,j}$ and $P_{i,j}$ are dimensionless numbers and are of similar nature and can therefore be aggregated through a weighted sum:

$$K_{i,j}=w_R R_{i,j}+w_P P_{i,j};\ w_R+w_P=1,$$

where weighting factors w_R and w_P are determined by experts.

Thus, for region $i=\overline{1,n}$ we have a set $\langle K_{i,1},K_{i,2},...,K_{i,m}\rangle$ of values of the aggregated criterion that accounts for a region's relative position in ratings compiled for comparison groups and for the entire country. Now, among values $K_{i,j}$, $j=\overline{1,m}$, the worst must be determined, which may also be performed by formula (2):

$$I_{i,j}=\left(1+e^{\frac{a-K_{i,j}}{b}}\right)^{-1}$$

where parameters a and b are calculated in selection $K_{i,j}$, $j=\overline{1,m}$.

For values of criterion $I_{i,j}$ the same remarks apply as for criterion $R_{i,j}$. Therefore, characteristic function (1) may be expressed as follows:

$$\Psi_I(x_{i,j})=\begin{cases}0, I_{i,j}<0,75;\\ 1, I_{i,j}\geq 0,75.\end{cases}$$

Thus, values $\Psi_I(x_{i,j})$ correspond to the highest priority of attention to be paid to the value of indicator X_j in the administrative decision-making process on the level of single region i.

Let us pass to the example.

Task: to calculate the indication of critical values of threat indicators for the entire country, in comparison with the group of a macro-district, and regarding a specific region.

Given: $X_{i,j}$ – a set of threat indicators values

Problem: $<<X_{i,j}>,<R_{i,j},R'_{i,j},P_{i,j},P'_{i,j},I_{i,j},I'_{i,j}>>$ where $R'_{i,j}$, $P'_{i,j}$, $I'_{i,j}$ are criteria that underwent indication.

Axioms:

Linear normalizing for the entire country: $(<D_{i,j}>,<R_{i,j}>)$

Linear normalizing for a macro-district: $(<D_{i,j}>,<P_{i,j}>)$

Aggregation of criteria $R_{i,j}$, $P_{i,j}$: $(<R_{i,j},P_{i,j}>,<K_{i,j}>)$

Linear normalizing for a specific region: $(<K_{i,j}>,<I_{i,j}>)$

Indication of critical threats values for the country: $(<R_{i,j}>,<R'_{i,j}>)$

Indication of critical threats values for a macro-district: $(<P_{i,j}>,<P'_{i,j}>)$

Indication of critical threats values for a region: $(<I_{i,j}>,<I'_{i,j}>)$

Search for a solution:

Axiom 7: $(<I_{i,j}>,<I'_{i,j}>)$

Reduction by rule 5:

$R_{i,j},R'_{i,j},P_{i,j},P'_{i,j},I_{i,j},I_{i,j}\Leftrightarrow R_{i,j},R'_{i,j},P_{i,j},P'_{i,j},I_{i,j}$

Axiom 4: $(<K_{i,j}>,<I_{i,j}>)$

Axiom 3: $(<R_{i,j},P_{i,j}>,<K_{i,j}>)$

Reduction by rule 5:

$R_{i,j},R'_{i,j},P_{i,j},P'_{i,j},R_{i,j},P_{i,j}\Leftrightarrow R_{i,j},R'_{i,j},P_{i,j},P'_{i,j},P_{i,j}$

Reduction by rule 5: $R_{i,j},R'_{i,j},P_{i,j},P'_{i,j},P_{i,j}\Leftrightarrow R_{i,j},R'_{i,j},P_{i,j},P'_{i,j}$

Axiom 5: $(<R_{i,j}>,<R'_{i,j}>)$

Reduction by rule 5: $R_{i,j},R_{i,j},P_{i,j},P'_{i,j}\Leftrightarrow R_{i,j},P_{i,j},P'_{i,j}$

Axiom 1: $(<D_{i,j}>,<R_{i,j}>)$

Axiom 6: $(<P_{i,j}>,<P'_{i,j}>)$

Reduction by rule 5: $D_{i,j},P_{i,j},P_{i,j}\Leftrightarrow D_{i,j},P_{i,j}$

Axiom 2: $(<D_{i,j}>,<P_{i,j}>)$

Reduction by rule 5: $D_{i,j},D_{i,j}\Leftrightarrow D_{i,j}$

The proof constitutes a set of vertices triples:
<D; D, D>, <D, D; (<D>,<P>); D, P>, <D, P; D, P, P>, <D, P, P; (<P>,<P'>); D, P, P'>, <D, P, P'; (<D>,<R>); R, P, P'>, <R, P, P'; R, R, P, P'>, < R, R, P, P'; (<R>,<R'>); R, R', P, P'>, < R, R', P, P', R, R', P, P', P>, < R, R', P, P', P; R, R', P, P', R, P>, < R, R', P, P', R, P; (<R, P>, <K>); R, R', P, P', K>, < R, R', P, P', K; (<K>,<I>); R, R', P, P', I>, < R, R', P, P', I; R, R', P, P', I, I>, < R, R', P, P', I, I; (<I>,<I'>); R, R', P, P', I, I'>

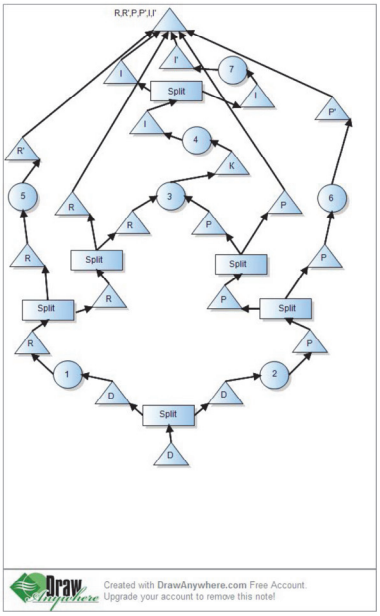


Fig. 7. The example of the solution tree reconstruction mechanism work

The execution sequence is found:

Axiom 2: $(<D_{i,j}>,<P_{i,j}>)$

Axiom 6: $(<P_{i,j}>,<P'_{i,j}>)$

Axiom 1: $(<D_{i,j}>,<R_{i,j}>)$

Axiom 5: $(<R_{i,j}>,<R'_{i,j}>)$

Axiom 3: $(<R_{i,j},P_{i,j}>,<K_{i,j}>)$

Axiom 4: $(<K_{i,j}>,<I_{i,j}>)$

Axiom 7: $(<I_{i,j}>,<I'_{i,j}>)$

9. Conclusion

Based on the analysis of existing data centres, their equipment and software, a high-quality solution is offered that provides a simple and flexible way to integrate heterogeneous information systems and their services into the World Data System. One of the key features of the proposed solution is the automation of algorithm construction of the actions sequence that executes users' queries. A logical formalism has been created to describe this solution, and on its basis an inference method and a solution tree reconstruction mechanism have been developed.

The analysis of available technologies used for the implementation of distributed systems allowed the use of such a set of software solutions for practical implementation of this solution: UDDI for creating a registry of services entered into the system; WSDL for a unified description of services; SOAP for exchanging notifications between services; BPEL for the overall coordination of services. Intellectual agents can be implemented using JADE.

The implementation of the proposed solution will provide an opportunity to use all the integrated computing capacities and data storage systems of the World Data System in a comprehensive manner. Thus, users will be able to easily gain access to all the necessary resources and services available to the system.

ACKNOWLEDGEMENTS

The research was conducted at the Faculty of Electrical and Computer Engineering, Cracow University of Technology and was financially supported by the Ministry of Science and Higher Education, Republic of Poland (grant no. E-1/2022).

AUTHORS

Grzegorz Nowakowski* – Faculty of Electrical and Computer Engineering, Cracow University of Technology, Cracow, Poland, e-mail: gnowakowski@pk.edu.pl.

Sergii Telenyk – Faculty of Electrical and Computer Engineering, Cracow University of Technology, Cracow, Poland, e-mail: stelenyk@pk.edu.pl.

Kostiantyn Yefremov – World Data Center for Geoinformatics and Sustainable Development, Kyiv, Ukraine, e-mail: k.yefremov@wdc.org.ua.

Volodymyr Khmeliuk – National Technical University of Ukraine "Igor Sikorsky Kyiv Politechnic Institute" Kyiv, Ukraine, e-mail: hmelyuk@gmail.com.

* Corresponding author

REFERENCES

[1] "About the Common Object Request Broker Architecture Specification Version 3.4," The Object Management Group (OMG), https://www.omg.org/spec/CORBA Accessed on: 2022-08-30.

[2] M. Juric, "BPEL and Java," 2021, https://www.theserverside.com/news/1364554/BPEL-and-Java Accessed on: 2022-08-30.

[3] "Data Integration Definition," https://www.heavy.ai/technical-glossary/data-integration Accessed on: 2022-08-30.

[4] "About the OMG System Modeling Language Specification Version 1.6," The Object Management Group (OMG), https://www.omg.org/spec/SysML/About-SysML/ Accessed on: 2022-08-30.

[5] "SOAP Version 1.2 Part 1: Messaging Framework (Second Edition)," World Wide Web Consortium (W3C), https://www.w3.org/TR/soap12-part1/ Accessed on: 2022-08-30.

[6] "UDDI Version 3.0.2," http://www.uddi.org/pubs/uddi_v3.htm Accessed on: 2022-08-30.

[7] "Windows Communication Foundation Architecture Overview," Microsoft Corporation, http://msdn.microsoft.com/en-us/library/ aa480210.aspx Accessed on: 2022-08-30.

[8] "Taverna – Apache Incubator," https://incubator.apache.org/projects/taverna.html Accessed on: 2022-08-30.

[9] S. A. El-Seoud, H. F. El-Sofany, M. A. F. Abdelfattah and R. Mohamed, "Big Data and Cloud Computing: Trends and Challenges", *International Journal of Interactive Mobile Technologies*, vol. 11, no. 2, 2017, 10.3991/ijim.v11i2.6561.

[10] S. Graham, *Building Web services with Java: making sense of XML, SOAP, WSDL, and UDDI*, Sams Publishing, 2005.

[11] J. Greer, *Web Services Description Language: 55 Most Asked Questions – What You Need To Know*, Emereo Publishing, 2014.

[12] M. R. Kogalovsky and L. A. Kalinichenko,"Conceptual and ontological modeling in information systems", *Programming and Computer Software*, vol. 35, no. 5, 2009, 241–256, 10.1134/S0361768809050016.

[13] J. Laznik, *BPEL and Java Cookbook: Over 100 Recipes to Help You Enhance Your SOA Composite Applications with Java and BPEL*, Packt Pub., 2013.

[14] A. Y. Levy, "Logic-Based Techniques in Data Integration". In: J. Minker (eds.), *Logic-Based Artificial Intelligence*, 2000, 575–595, 10.1007/978-1-4615-1567-8_24.

[15] P. Maslianko, "Fundamentals of the Methodology of System Design of Information and Communication Systems", *Research Bulletin of the National Technical University of Ukraine "Kyiv Polytechnic Institute"*, vol. 6, 2007, 54–60.

[16] G. Nowakowski, "Open source relational databases and their capabilities in constructing a web-based system designed to support the functioning of a health clinic", *Czasopismo Techniczne*, vol. Automatyka Zeszyt 1-AC (2), 2013, 53–65.

[17] G. Nowakowski, "Rest API safety assurance by means of HMAC mechanism", *Information Sys-*

tems in Management, vol. 5, no. 3, 2016, 358–369.

[18] G. Nowakowski, S. Telenyk, K. Yefremov and V. Khmeliuk, "The Approach to Applications Integration for World Data Center Interdisciplinary Scientific Investigations". In: *2019 Federated Conference on Computer Science and Information Systems*, 2019, 539–545, 10.15439/2019F71.

[19] S. Telenyk and O. Pavlov, "Algorithmization and IT in management," Kyiv: Technics, 2002.

[20] N. Shakhovska, M. Medykovskyj, V. Lytvyn, "Dataspace Class Algebraic System for Modeling Integrated Processes," *Journal of Applied Computer Science*, vol. 20, no. 1, 2012, 69–80, https://it.p.lodz.pl/file.php/12/2012-1/JACS-1-2012-Shakhovska.pdf Accessed on: 2022-09-11.

[21] N. Shakhovska, "Methods of Processing Data Using Consolidated Data Space", *Problems of Development, National Academy of Sciences of Ukraine, Institute of Software of NAS of Ukraine*, vol. 4, 2011, 72–84.

[22] S. Telenyk, G. Nowakowski, K. Yefremov and V. Khmeliuk, "Logics based application integration for interdisciplinary scientific investigations". In: *2017 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, 2017, 1026–1031, 10.1109/IDAACS.2017.8095241.

[23] S. Telenyk, G. Nowakowski, E. Zharikov and J. Vovk, "Conceptual Foundations of the Use of Formal Models and Methods for the Rapid Creation of Web Applications". In: *2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, 2019, 512–518, 10.1109/IDAACS.2019.8924416.

[24] M. Ulema, B. Wu, J. Hwang, F. Lin and J.-H. Yi, "Next generation service overlay networks (NGSON)", *IEEE Communications Magazine*, vol. 50, no. 1, 2012, 52–53, 10.1109/MCOM.2012.6122532.

[25] H. G. Miller and J. Veiga, "Cloud Computing: Will Commodity Services Benefit Users Long Term?", *IT Professional*, vol. 11, no. 6, 2009, 57–59, 10.1109/MITP.2009.117.

[26] M. Zgurovsky, O. Akimova, A. Boldak, S. Voitko, O. Havrysh, O. Gluhanyk, I. Dzhygyrey, K. Yefremov, A. Ishchenko, A. Kovalchuk, M. Kokorina, S. Lazareva, I. Makodym, T. Matorina, Y. Matsuki, A. Melnychenko, I. Pyshnohrayev, A. Prakhovnik, G. Statyukha and S. Tulchinska, "Part 2. Ukraine in Indicators of Sustainable Development (2011-2012)". In: *Analysis of Sustainable Development – Global and Regional Contexts*, 2012.

[27] M. Z. Zgurovsky, A. D. Gvishiani, K. V. Yefremov and A. M. Pasichny, "Integration of the Ukrainian science into the world data system", *Cybernetics and Systems Analysis*, vol. 46, no. 2, 2010, 211–219, 10.1007/s10559-010-9199-9.

[28] V. Ivanov and K. Smolander, "Implementation of a DevOps Pipeline for Serverless Applications". In: M. Kuhrmann, K. Schneider, D. Pfahl, S. Amasaki, M. Ciolkowski, R. Hebig, P. Tell, J. Klünder and S. Küpper (eds.), *Product-Focused Software Process Improvement. PROFES 2018, Lecture Notes in Computer Science*, 2018, 48–64, 10.1007/978-3-030-03673-7_4.

[29] C. Vassallo, S. Proksch, A. Jancso, H. C. Gall and M. Di Penta, "Configuration smells in continuous delivery pipelines: a linter and a six-month study on GitLab". In: *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2020, 327–337, 10.1145/3368089.3409709.

[30] N. Siegmund, N. Ruckel and J. Siegmund, "Dimensions of software configuration: on the configuration context in modern software development". In: *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2020, 338–349, 10.1145/3368089.3409675.

[31] S. P. Gilroy and B. A. Kaplan, "Furthering Open Science in Behavior Analysis: An Introduction and Tutorial for Using GitHub in Research", *Perspectives on Behavior Science*, vol. 42, no. 3, 2019, 565–581, 10.1007/s40614-019-00202-5.

[32] M. Auch, M. Weber, P. Mandl and C. Wolff, "Similarity-based analyses on software applications: A systematic literature review", *Journal of Systems and Software*, vol. 168, 2020, 10.1016/j.jss.2020.110669.

The 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications
18-21 September, 2019, Metz, France

Conceptual Foundations of the Use of Formal Models and Methods for the Rapid Creation of Web Applications

Sergii Telenyk [1], Grzegorz Nowakowski [2], Eduard Zharikov [3], Jewhenii Vovk [4]
[1] Department of Theoretical Electrical Engineering and Computer Science, Faculty of Electrical and Computer Engineering, Cracow University of Technology, Cracow, Poland, stelenyk@pk.edu.pl, https://www.pk.edu.pl
[2] Department of Automatic Control and Information Technology, Faculty of Electrical and Computer Engineering, Cracow University of Technology, Cracow, Poland, gnowakowski@pk.edu.pl, pk.edu.pl
[3,4] Department of Automation and Control in Technical Systems, National Technical University of Ukraine "Igor Sikorsky Kyiv, Polytechnic Institute", Kyiv, Ukraine, [3]zharikov.eduard@acts.kpi.ua, [4]kafedra@acts.kpi.ua, kpi.ua/en

***Abstract*—The problem of the rapid creation of effective web applications of one class using formal means is considered. A conceptual approach to its solution is offered based on an analysis of the peculiarities of the construction of web applications. Our approach is based on defining the standard web application architecture and selecting its components using formal methods in accordance with user requirements. A formal logical system is proposed that uses the design of web applications as a process of outputting a formula specified according to the needs of the user, which defines the schemes of execution of the modules of the system. An important feature of this approach is the ability to visualise the process of designing the system in 3D, which creates the conditions for effective interaction between developers and machine development tools.**

Keywords—web-applications, application construction, business processes, mathematical logic

I. INTRODUCTION

Today, the creation of information systems is based on modern methodological concepts that have inherited the most important ideas of classical methodologies such as SADT and IDEF0, while enriching them with new ideas.

Historically, the most important directions for improving classical methodologies have been: (i) structuring software that has opened the way for teamwork between programmers and the reuse of software code; (ii) describing the possibilities of the created system using the language of conceptual modelling, which was accompanied by the development of automated design systems; (iii) structuring the design process, which has enriched the industry with the concept of the system life cycle and repository, and unified models of role design systems; and (iv) prototyping, based on which the concepts of parallel development of parts of the system, and the design and implementation of the system by parts were developed.

In general, modern methodologies for information systems design, for example Agile methodology, have successfully used different combinations of these concepts, filling them with real content and complementing the elements of psychology.

Formalisation is very specific in modern methodologies [1]-[4]. In classical methodologies, it was considered the basis of automation, for example CASE tools. Today, formalisation is more closely related to describing the architecture of the system, and descriptions are mainly used to set tasks for project participants, discuss intermediate results, and define the next steps [1]-[2].

This article attempts to use previously developed formal models and effective methods to build a technology for the automated generation of software systems using modern methodologies.

As this problem involves very large scales, we have identified one of the most advanced components of modern information systems: the development of systems with web presentation.

II. THE ESSENCE OF THE PROBLEM

The expediency of the class automation of tasks is explained by the considerable complexity of the design and implementation of the subtasks into which they are decomposed. Subproblems are duplicated in each project, and changes are subject to data but not their main transformation.

Developers should agree on a universal architecture; the application will then have the same overall appearance and differ from other applications only in terms of the presence or absence of certain components, depending on the functional requirements for the application.

Adhering to the principles of the three-tier architecture, we identify three levels in the typical application: representation, business logic and access to data. The level of data access determines the ability to work with data, and in general, this is the implementation of the *Repository* template. We define access to data, if

978-1-7281-4069-8/19/$31.00 ©2019 IEEE

necessary, based on business logic in the traditional way: (i) certain generalised behaviour is associated with each entity, which is inherent in all elements of the level of access to data; and (ii) behaviour is expanded by adding unique methods for the current entity.

Accordingly, a specific repository class inherits its interface, obtaining basic methods that are common to all classes and its own methods from its own interface; it also inherits the class that implements the basic behaviour, the fact of which contains the implementation of only their methods. The template has a specific architecture, and is expanded by adding new classes and interfaces in order to work with each entity.

Business logic is responsible for processing the data [5] that are obtained based on the level of access, through the use of behavioural patterns such as strategy and behavioural classes, in certain cases of commands. Each component implements the end-user functionality that the user expects. Using this concept, components can be expanded by adding classes with narrowly oriented behaviour.

The *Strategy* template aims to implement a set of different behaviours depending on the needs of the component that is level higher and appeals to this strategy. A strategy is described by an appropriate method that takes the required parameters as input, and several implementations of the method.

The *Command* template provides the implementation of a certain behaviour on demand, allowing abstraction from the logic of the command itself within the mechanism.

Another mechanism that is used is an interface with behaviour descriptions and a implementation class that contains the data processing elements and the formation of the result required for a component that is in the architecture of the level above.

A web presentation is a system component that is accessible from the outside and is visible to all users. It is used to receive and analyse queries and to perform necessary actions by calling business logic methods and displaying data on the client side. This component is mainly built on an MVC template, consisting of representations and methods that are directly invoked following a request from the processing client.

The MVC template is built in accordance with its purpose: the model determines which data should be displayed; the representation is a graphical component that can display a particular model; and the controller determines how to get the model, which presentation to submit it and how to process the request.

By following this description, the task of creating any web application can be equated to the task of filling in a particular template. However, building a template for the correct architecture, and writing long, monotonous code require a great deal of time. The problem of automating the creation of a software system therefore appears from the automated implementation of a system template based on user requirements. Solving this problem involves the following steps: designing a database model and setting levels of access to it; realising data processing through built-in strategies at the user's request; and generating a presentation.

III. Existing Approaches to Creating Web-Based Applications

Attempts to automate the creation of web applications have been made for some time, and it is worth mentioning previously developed solutions such as CMS. The logic of the behaviour of these constructors is to use the finished system for the entire range of tasks, at the expense of the means of choosing exclusively the theme of the graphic design and adding or disconnecting those or other pages. This approach has a number of shortcomings.

More promising is an approach to creating web applications based on models of the processes to be automated. Many model-driven engineering approaches have been proposed to develop web applications. The auto-generation of applications is also a mature research field, and several approaches have been proposed.

A great deal of work has been devoted to the auto-generation of web application code. For example, in [6], the authors introduce a research agenda for assurance at runtime with models at runtime (M@RT) as a foundation, and emphasise that research into the use of M@RT is fundamental to the development of runtime assurance techniques. They define the types of information that can be captured by M@RT, outline key research challenges for assurance at runtime, characterise assurance methods, and propose selected application areas in which M@RT could provide significant benefits beyond existing assurance techniques for adaptive systems.

In [7], the authors follow the same approach, but due to the complexity and heterogeneity of systems that are the object of development, suggest the use of several models to take into account the overall complexity of the problem.

Another solution is the composite automation of the capabilities of individual modules through the use of frameworks built on appropriate technology.

The authors of [8] propose a complexity less approach for the automated generation of web-based applications in the data management domain. This can be used to create very complex applications by autogenerating them from the definitions of data models. Using the proposed approach, CRUD operations, integrity checks and user documentation can be automatically generated from a UML-based data model. The results of evaluation show that the proposed approach allows the total project cost to be reduced due to the autogeneration of code, and this can make it easier to maintain, debug and track a generated web-based application. The emphasis is on the complexity of the problem of autogenerating web applications and the need to use several interrelated models, based on data model definitions in UML.

In [9], the general idea of a model-driven approach to the creation of web-based applications is developed using technology based on the concept of prototyping. There are services at the level of the technological core unit, and reuse of templates is considered as a means of simplifying and cheapening the design. This approach is suitable for rapid prototyping and model verification, and is compatible with common programming languages. The authors present a meta-model for designing web services and a related tool for verifying the consistency of a model. The use of high-level constructions for the definition of web-services allows web service patterns to be formally verified and safely reused in other contexts. The overall approach was illustrated with a case study, using the proposed tool to demonstrate its advantages, such as robustness and flexibility.

In [10], the authors consider the development of cross-platform web applications with the aim of finding a technology unifier. They analyse the foundations of progressive web applications (PWAs) in cross-platform development, scrutinise the status quo of current possibilities, investigate the idea of unified development and discuss several open questions. The authors also introduce several technologies and their underpinnings, and examine the possibilities of PWAs, finally suggesting a balanced approach combining experimental and qualitative work.

The study in [11] conceptually resonates with the work in [3,4]. The emphasis is on prototyping, although here this is based on the simulation of flows of information. The development workflow forms the basis for the co-evolution of the model and code. The authors introduce IFMLEdit.org, an online tool for the rapid prototyping of web applications, and describe the user's interaction with the application by means of flows of information in reaction to user events. The proposed tool allows the user to edit interaction flow modelling language (IFML) specifications, to generate the first version of the code of a web application from the model, to improve the generated code with manually added details, to evolve the original IFML model and to re-generate the code for the updated version. The proposed approach solves the well-known problem of model-driven forward engineering through the rapid modification of the application model and the generation of realistic prototypes that can easily be turned into deployable applications.

The work in [12] emphasises the model-driven engineering trend, but with an additional aspect of the agile methods effects study. The authors summarise several research activities focusing on the development and use of models in software and systems engineering. They describe engineering software and software-intensive systems to understand system conditions and optimisation potentials that require a precise but abstract understanding of these heterogeneous systems. The authors also consider the benefits of agile methods in language engineering, the application of model-based engineering techniques to different domains, and tools and languages developed using the language workbench MontiCore.

The study in [13] is interesting as it takes into account the logging of interaction events when creating web applications. The proposed original technique for the in-page logging of interaction events can help interaction designers make more informed design decisions. This technique facilitates overall user satisfaction by measuring content engagement using low-level logging on the client side of the application. Thus, the workload on web servers caused by server-side logging can be significantly decreased. On the research side, in-page logging of interaction events allows the user to analyse meaningful to designers by capturing a rich representation of the events in the user interface at multiple levels of granularity.

The article considers other approaches to automating the creation of web applications, and concludes that they are partially compliant with the solution to the problem of generating the code of web applications, taking into account the possibility of managing the automation process at each level and the appropriate capability of selecting and integrating existing components based on formal user requirements and descriptions of these components.

For a more detailed consideration of the architecture of web applications, an appropriate alternative allowing their rapid creation is to create templates for their implementation at each level of the architecture, using existing frameworks developed for a particular type of technology.

However, there are several obvious aspects of solving this problem. The level of access to data is subject to templating based on the entities of the database as separate components. Similarly, the creation of a REST API [14] can be customised based on process descriptions.

The automation of the rapid creation of web applications using this approach leaves programmers with the key problem of combining the resulting fragments into a general application, which requires appropriate changes to the autogenerated product code.

IV. Conceptual Foundations for Creating Web-Applications Based on Templates

The main idea behind the development is to build a real model based on the basic notion of architecture, in order to model the application development process.

A set of template processes can be implemented in one application at the expense of means of using the frameworks and specifications for a certain form of technology. The application is multilevel, and consists of template modules. Each module is constructed of classes and interfaces that have their own structural units, and the appearance and methods of the application depend on the tasks to be solved in accordance with the requirements of users.

The design of any system begins with the implementation of business processes and the structure of the data warehouse to meet the objectives. The user needs to create a schema model for the database, in order to describe the entities and indicate their links.

Following this, the system builds the basic functionality in the form of a data model with scripts for database generation at the appropriate level of data access, using scenarios to work with users and authorisations that the user can choose depending on their needs.

The next stage of the work is the direct generation of the data component, i.e. data selection and data preparation, with a number of conditions created by combining existing conditions using the proposed template, or the user creating their own. After the user has set the conditions, a semantic description of the corresponding transformation is created and the data are processed in accordance with the current business process.

The final stage is to generate the presentation with the appropriate selection and to select and edit the most suitable template from the proposed alternatives.

The result is a ready-made application that is generated by creating and assembling the individual modules of the system.

V. MATHEMATICAL MODEL FOR CONSTRUCTING BASIC FUNCTIONALITY

To select and integrate the components of the complete solution, we will use first-order causal logic, the structural elements of which are described here for the integration of applications in the work [1].

Symbols:
service: (,), [,], {, }, :, <, >, ,;
constant:

- individual, of primary types (int, real, char, bool) - a_1^1, a_2^1, …, a_1^2, a_2^2, ... where each constant a_i^k pertains to type (primary type) *k*; structural type (construct) - c_1, c_2, ...; procedural type (method) - d_1, d_2, ...; objective type (problem, entity, relation) — e_1, e_2, ...;
- *functional i-place*, for individuals of type *k* - h_1^1, h_2^1, …, h_1^2, h_2^2, ...;
- *predicate i-place*, for individuals of type *k* - A_1^1, A_2^1, …, A_1^2, A_2^2, ... (this class includes taxonomic, relational and other predicates, as well as traditional relations, at least equality = and order ≥);

variable: for individuals of type k - x_1^1, x_2^1, …, x_1^2, x_2^2, ..., where every variable x_i^k pertains to type *k*;

logical: $\neg, \wedge, \vee, \leftarrow, \exists, \forall, \Leftrightarrow$.

Individual terms of type *k*:

- each individual constant a_i^k of type *k* is an individual term of type *k*;
- each free variable x_i^k for individuals of type *k* is an individual term of type *k*;
- if h_i^j is a certain functional constant for individuals of type *k* and τ1, …, τj are terms for individuals of type *k*, then h_i^j (τ1, …, τj) is an individual term of type *k*;
- there are no other individual terms of type *k*.

Formulas for individuals:

- if A_i^j is a predicate constant for individuals and τ1, …, τj are terms for them, then A_i^j (τ1, …, τj) is the atomic formula for individuals;
- the atomic formula for individuals is the formula for them;
- there are no other formulas for individuals.

Hereinafter we consider only the systems of Horn clauses (with a single → symbol, atomic formulas to its left and right, and an implicit quantifier $\forall$).

Specifiers, preconditions, post-conditions, specifiers of methods, specifiers of problems, clause are presented in detail in [1].

The knowledge base of the system consists of three main parts. In the first part, ontological axioms describe the methods and other components of the framework that are available for use. The second part provides an ontology of the system, and is described using OWL-based languages based on RDF. The third part summarises the output rules that are required to achieve the desired result for the user. There are two types of rules for the output. The first type are used to integrate several methods and other components into a single application. They take into account the specific nature of the problem, the preconditions, the stages, the descriptions of methods and other components of the system. Realised with the desired output proof in a certain way defines the architecture of the application that is being created within a defined class of architectures.

Inference rules of first type:

- **If** d1: (<τ1>,< τ2>) **and** d2: (<τ3>,< τ1>) **then** d1: (<d2>,< τ2>)
- **If** d1: (<τ1>,< τ2>) **and** d2: (<τ3 ∧ τ4>,< τ1>) **then** d1: (<d2>,< τ2>)
- **If** d1: (<τ1 ∧ τ2>,< τ3>) **and** d2: (<τ4>,< τ1>) **and** d3: (<τ5>,< τ2>) **then** d1: (<d2 ∧ d3>,< τ3 >)
- **If** d1: (<τ1>,< τ2>) **and** d2: (<τ3 ∨ τ4 >,< τ1>) **then** d1: (<d2>,< τ2>)
- **If** d1: (<τ2 ∧ τ2>,< τ1>) **then** d1: (<τ2>,< τ1>)
- **If** d1: (<τ1, τ2>,< τ3>) **and** d2: (<τ4>,< τ1>) **and** d3: (<τ5>,< τ2>) **then** d1: (<d2, d3>,< τ2>)

- **If** d1: (<τ1>,< τ2,τ3>) **and** d2: (<τ2>,< τ4>) **and** d3: (<τ3>,< τ5>) **then** d2: (<d1>,< τ4>) **and** d3: (<d1>,< τ5>)

The second type of inference rules are used to speed up the inference process. They take into account the semantics of the user's problem and the semantics of the methods and the other components of the system, expressed in terms of the system's ontology [15]. In essence, these rules are used to reduce the search for the rules and axioms of the first part of the knowledge base of the system. They describe ways of combining methods that are semantically acceptable at the level of input and output and other components of the system in more functionally complete combinations. These combinations can be used to deduce in the space of rules of the first type. The meaningful use of the second type of inference rules can be illustrated by the process of 3D visualisation of the combination of methods and other components of the system, using spatial structures based on common entities for their inputs/outputs.

VI. The Semantically Controlled Inference Mechanism

For the inference we will use the approach proposed by the authors of [1]. This defines the procedures for inference and restoring the inference tree. However, we will add to this approach a preliminary procedure for searching the spatial structure of the associated by input/output methods and other components of the system, which at inputs has entities defined by the user as input information, and at its outputs are entities defined by the user as the source information. Define the necessary concepts.

Here, we focus on the means of knowledge representation, and to reduce the search, we use the knowledge base and particularly information on effective inference schemes for frequently performed queries and its ability to structure, factorise and abstract. This is a combined inference strategy, at the lower levels of which the birth of a lossy resolvents is blocked. Formally, this is a question of combining the approach of Kowalski [3] and the analogue method of Plaisted [4]. In this case, the proof in the abstract space is based exclusively on the axioms and inference rules of the knowledge base, and the result is then used to control the inference in the initial search space.

We cut off the hopeless branches of the inference in the initial space in two stages. In the first stage, we formulate the constructions of the associated by input/output methods and other components of the system. In the second stage, we use the inference in the abstract space. For the reflection of the transition to the abstract space, we use taxonomic connections. The inference in the abstract space will then be reduced to the inference in the system of types (classes of entities) with a gradual deepening of the system of subtypes up to individuals. As before, to reduce the search in both the abstract (for classes and types) and in the initial (for individuals) spaces, we will use modifications to the Robinson resolution strategy [16]. The properties needed for the mutual adaptation of the method of analogy and modification of the resolution strategy the used reflection do naturally acquire in a responsibly structured knowledge base.

For the analogy method, we use the concept of a multi clause (*m*-clause) as a multiset of atomic formulas (the atomic formula *L* will be recorded in *m*-clause as many times as it is repeated). The operations $\cup$ (union), $\cap$ (intersection), — (difference), · (concatenation) and relation $\subseteq$ (occurrence) for multisets are performed (note that the operations are performed separately for the left- and right-hand parts of the clause).

Suppose $A_1 \in C_1$, $A_2 \in C_2$ and α_1, α_2 are substitutions that allow us to obtain the most common unifier for the atomic formulas A_1 and A_2. Then, the clause obtained from the unification of clauses $C_1\alpha_1$ and $C_2\alpha_2$ by removing *L* to the left and right of the symbol → is called the *m*-resolvent of *m*-clauses C_1 and C_2. If the *m*-clauses C_1 and C_2 are ordered and the *m*-resolvent is obtained by eliminating the non-underlined atomic formula in both parts, and no other atomic formula follows it on the left, then we get the ordered linear *m*-resolvent. If the last atomic formula to the left of the → symbol in the ordered *m*-clause is unified with the underlined atomic formula to the right of the → symbol of the same *m*-clause, the ordered linear *m*-resolvent is obtained by reduction of the the *m*-clause.

Definitions of *m*-clauses and *m*-resolvents are used to define the ordered linear *m-resolutional* proof. We start with the definition of the *m-resolutional* proof *Tm* as a pair <*V*, *Th*>, where *V* is the set of proof vertices and *Th* is the set of vertex triples. In the future, the first and second components of *Tm* will be allocated using the *s-N*(*Tm*) and *s-M*(*Tm*) selector functions, respectively. Each vertex $n \in$ *s-N*(*Tm*) of the proof *Tm* is characterised by the mark *s-L*(*n*) and the depth *s-D*(*n*). If $<n_1, n_2, n_3> \in$ *s-M*(*Tm*), then *s-L*(n_3) is the *m*-resolvent *s-L*(n_1) and *s-L*(n_2), and each triple of this type is called *m*-resolution. In the proof *Tm*, the vertex $n \in$ *s-N*(*Tm*) is called the initial if it is not the third component of any of the triples of *s-M*(*Tm*) (its mark is the initial *m*-clause), or the terminal if it is neither the first nor the second component of any of the triples of *s-M*(*Tm*) (its mark is the terminal *m*-clause).

We call the *m-resolutional* proof *Tm* the proof from S, if the marks of the initial vertices *Tm* belong to the set of *m*-clause S. From S, we deduce C if *Tm* is a proof from S, and C is a mark of one of the vertices of *Tm*.

Finally, the ordered linear *m-resolutional* proof from S is called *m-resolutional* proof *Tm* from S, all *m*-clauses of which are ordered and for an arbitrary triple $<n_1, n_2, n_3>$ from *s-M*(*Tm*) *s-L*(n_3) is an ordered linear *m*-resolvent *s-L*(n_1) and *s-L*(n_2).

To manage an ordered linear inference, we will use the typification abstraction proposed in [4]. Suppose f is a mapping from the set of m-clauses into the set of m-clauses such that: (i) if m-clause C_3 is the m-resolvent of m-clauses C_1 and C_2, while $D_3 \in f(C_3)$, then $D_1 \in f(C_1)$ and $D_2 \in f(C_2)$ exists such that the result of the substitution of some m-resolvent D_1 and D_2 belongs to D_3; (ii) $f(\varnothing) = \{\varnothing\}$; and (iii) if the result of some substitution of m-clause C_1 belongs to m-clause C_2, then for any abstraction D_2 for C_2 an abstraction D_1 exists for C_1 such that the result of D_1 substitution belongs to D_2. This is called f m-abstraction mapping, while any D from $f(C)$ is called m-abstraction. The typification mapping is understood as a certain mapping ϕ from a set of literals into a set of literals, which transforms each atomic formula into the formula for which the terms have the type closest to the basic types in the hierarchy.

Construction of the proof begins with the formulation of the problem by the user. In essence, the possibility of executing the request is checked, taking into account all available resources. Having access to the descriptions of all modules and other components and axioms that determine the possibilities of their application, the inference mechanism combines the modules and other components into a structure (proof) that ensures the receipt of the result from the input data.

Let the ordered linear *m-resolutional* proof be obtained by definition as a pair of sets: vertices $V = \{k_1, k_2, k_3, \ldots k_n\}$ and vertex triples $Th = \{< k_1, k_2, k_3 >, < k_3, k_4, k_5 > \ldots < k_{n-2}, k_{n-1}, k_n >\}$, where k_i, $i = 1,\ldots,$ n are the vertices of proof, and k_n is the terminal vertex. The proof is formed using the initial vertices of the problem postcondition and, as a lateral vertex, the appropriate axiom from the knowledge base. The axiom is appropriate if the sequence of atomic formulas corresponding to this vertex of the proof, or any part of it, is an axiom postcondition. The third vertex of this triple is obtained by applying the axiom to the formula postcondition, taking into account the inference rules. The third vertex of the first triple becomes the first vertex of the second triple, and the process repeats until the terminal vertex is reached, i.e. the problem precondition. If the atomic formula corresponding to the third vertex of the triple can be simplified by applying one of the inference rules, it is simplified in the next triangle, which will have only two vertices: a vertex with a formula, which must be simplified, and a vertex with a simplified formula.

The search for the corresponding axioms is done by comparing the construct of the postcondition that is currently being processed with the axiom constructs from the ontology. Thus, a set of axioms is chosen from the ontology that describes the methods that process the type and format of the objects we need.

If there are several corresponding axioms in the knowledge base for the next vertex, then each of these axioms is used to build its own "parallel" version of the proof. Thus, during the operation of the inference mechanism, a number of proofs of varying length and complexity can be formed. Upon completion of the inference mechanism, we select from the set of proofs the one with the smallest number of triples of vertices and the smallest number of applied axioms.

To shorten the search, we can pre-select the structures of related methods and other components, the entities of which inputs and outputs are common. After the structures have been selected, the preconditions and postconditions are checked and the architecture of the solution is determined only in the second stage.

The output of the inference mechanism is a set of formulas and a clause that corresponds to the vertices of the proof triples and reflects the application of the axiom of the first part and the inference rules of the first type. To restore the structure, we use the proof tree recovery algorithm proposed by the authors of [1].

Then, using the constructor, we generate the user representation using the finished elements, to which the output points of the business logic of the application are tied in the form of a data source.

VII. IMPLEMENTATION

This technology is realised using modern tools. The user-friendly web interface allows the end user to build a database model that is used to generate the base framework of the application (CRUD data operations with the subsequent generation of the interface as a REST API [14]). The next step is to display this general view of the application with the ability to generate the logical behaviour and implementation of the business processes of the required application, by means of a semantic assignment of relationships between the relevant models of the application.

Pre-set templates are integrated into the application with pointing inputs and outputs. Additional functionality includes the ability to create and store the user's own templates as well as the ability to put them in the public domain. Formally, template integration can take place at different levels of the architecture, ranging from method integration to the integration of an entire component consisting of classes.

The applications are composed of modules, which are composed of classes, which in turn are composed of methods, etc. Definition by the user of the input and output entities and their characteristics allows the necessary methods and their interconnections to be found and the template model of the lowest architectural level to be obtained, presented in the form of classes.

At the highest level, the composition model is maintained, but the corresponding classes are already merged through the calls of other methods and the generation of behaviour classes, which is the source of the

module. Accordingly, the combination of modules follows from the composition of their outgoing points and the established connections.

VIII. CONCLUSION

In the article, the problem of the rapid creation of effective web applications of one class is considered. Our analysis of the problem and existing decisions has confirmed the relevance of the problem and allowed us to formulate it and identify the key aspects of the study.

A complex approach is proposed to the solution of the problem, based on autogenous template solutions for information systems that can be integrated into the process of components at different architectural levels. The main advantages of the proposed solution are that it offers a formal basis for creating applications, based on templating the design and implementation processes, and the ability to fine-tune the process [17]. In addition, the model adopted for constructing template solutions simplifies the work of developers and helps automate the basic process of developing web applications for a wide range of business information systems with web representation.

An analysis of the problem and the existing solutions enabled the choice of appropriate formal means. The mathematical apparatus for the proposed solution includes means for the semantic description of business processes and a formal logical system. The first provides a description of business processes and allows them to be decomposed into subprocesses from the highest level of architecture to non-decomposable atomic components, indicating the input and output entity, and the characteristics based on which the choice of the corresponding functional unit is made - a template for the processing of the subprocess, or the creation of the user's own template based on the proposed methods. A formal logic system facilitates the selection and integration of components into a complete solution.

Future research will involve the 3D visualisation of the processes of creating web applications, using the proposed theoretical model for building a comprehensive solution and based on the template for the process of creating an information system with a web representation.

ACKNOWLEDGMENT

Presented results of the research, which was carried out under the theme No. E-3/586/2018/DS, were funded by the subsidies on science granted by Polish Ministry of Science and Higher Education.

REFERENCES

[1] S. Telenyk, G. Nowakowski, K. Yefremov and V. Khmeliuk, "Logics based application integration for interdisciplinary scientific investigations," *Proceedings of the 2017 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS'2017)*, Bucharest, 2017, pp. 1026-1031. DOI: 10.1109/IDAACS.2017.8095241

[2] A.Y. Levy, "Logic-based techniques in data integration, in: logic based artificial intelligence," Edited by J. Minker: Kluwer Publishers, 2000, pp. 575-595. DOI: 10.1007/978-1-4615-1567-8_24

[3] R. Kowalski, "Computational logic and human thinking: how to be artificially intelligent," Cambridge University Press, 2011, pp. 213-226. DOI: 10.1017/CBO9780511984747

[4] D. A. Plaisted, "History and prospects for first-order automated deduction," *Proceedings of the International Conference on Automated Deduction*, Springer, Cham, 2015, pp. 3-28. DOI: 10.1007/978-3-319-21401-6_1

[5] J. Su, V. Vysotska, A. Sachenko, V. Lytvyn, Y. Burov, "Information resources processing using linguistic analysis of textual content," *Proceedings of the 2017 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS'2017)*, Bucharest, 2017, pp. 573-578. DOI: 10.1109/IDAACS.2017.8095038

[6] B. Cheng et. al., "Using models at runtime to address assurance for self-adaptive systems," *In Models@ run time*, Springer, Cham, 2014, pp. 101-136.

[7] K. Wakil and D. Jawawi, "A new adaptive model for web engineering methods to develop modern web applications," *Proceedings of the 2018 ACM International Conference on Software Engineering and Information Management*, 2018, pp. 32-39.

[8] T. Panetti and A. D'Ambrogio, "A complexity-less approach for automated development of data-intensive web applications," *Proceedings of the 2018 IEEE International Symposium on Networks, Computers and Communications (ISNCC'2018)*, 2018, pp. 1-6.

[9] D. Sferruzza, J. Rocheteau, C. Attiogbe, A. Lanoix, "A model-driven method for fast building consistent web services in practice," Modelsward, INSTICC, 2018, pp. 1-12.

[10] T. Majchrzak, A. Biørn-Hansen, T. Grønli, "Progressive web apps: the definite approach to cross-platform development?" *Proceedings of the 51st Hawaii International Conference on System Sciences*, 2018, pp. 5735-5744.

[11] E. Falzone, and C. Bernaschina, "Model based rapid prototyping and evolution of web application," *Proceedings of the International Conference on Web Engineering*, Springer, Cham 2018, pp. 496-500.

[12] K. Hölldobler, J. Michael, J. Ringert, B. Rumpe, A. Wortmann, "Innovations in model-based software and systems engineering," *Journal of Object Technology*, vol. 18, no. 1, pp. 1-60, 2019.

[13] A. Edmonds, R. White, D. Morris, S. Drucker, "Instrumenting the dynamic web," *Journal of Web Engineering*, vol. 6, pp. 243-260, 2019.

[14] G. Nowakowski, "Rest Api safety assurance by means of HMAC mechanism," *Information Systems in Management*, vol. 5, no. 3, pp. 358-369, 2016.

[15] N. Garanina, E. Sidorova, I. Kononenko, S. Gorlatch, "Using multiple semantic measures for coreference resolution in ontology population," *International Journal of Computing*, vol. 16, issue 3, pp. 166-176, 2017.

[16] J. A. Robinson, "A machine-oriented logic based on the resolution principle," *Journal of the Association for Computing Machinery*, vol. 12, pp. 23–41, 1965.

[17] N. O. Hodas et. Al., "Beyond fine tuning: Adding capacity to leverage few labels," *Proceedings of the 31st Conference on Neural Information Processing Systems*, Long Beach, CA, USA, 2017, pp. 1-7.

BULLETIN OF THE POLISH ACADEMY OF SCIENCES
TECHNICAL SCIENCES, Vol. 72(4), 2024, Article number: e150117
DOI: 10.24425/bpasts.2024.150117

An analysis of the influence of famous people's posts on social networks on the cryptocurrency exchange rate

Sergii TELENYK[1], **Grzegorz NOWAKOWSKI**[1]*, **Olena GAVRILENKO**[2], **Mykhailo MIAHKYI**[2], and **Olena KHALUS**[2]

[1] Faculty of Electrical and Computer Engineering, Cracow University of Technology, Warszawska 24, 31-155 Cracow, Poland
[2] National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute," Prosp. Peremohy 37, Kyiv, Ukraine

Abstract. In this work, the level of influence of the posts published by famous people on social networks on the formation of the cryptocurrency exchange rate is investigated. Celebrities who are familiar with the financial industry, especially with the cryptocurrency market, or are somehow connected to a certain cryptocurrency, such as Elon Musk with Dogecoin, are chosen as experts whose influence through social media posts on cryptocurrency rates is examined. This research is conducted based on statistical analysis. Real cryptocurrency exchange rate forecasts for the selected time period and predicted ones for the same period, obtained using three algorithms, are utilized as a dataset. This paper uses methods such as statistical hypotheses regarding the significance of Spearman's rank correlation coefficient and Pearson's correlation. It is confirmed that the posts by famous people on social networks significantly affect the exchange rates of cryptocurrencies.

Keywords: cryptocurrency exchange rate; forecasting algorithms; posts on social networks; methods of statistical analysis; information technology of intellectual analysis.

1. INTRODUCTION

Today, people can obtain the information they need through the Internet without leaving their homes. But there are problems with availability – the Internet is filled with a large quantity of non-informative data that has no value for the vast majority of users, or in other words, "garbage data". That is why every day it becomes increasingly difficult to find relevant information because huge amounts of data must be reviewed and analyzed.

This problem is inherent in all spheres of human activity, whether it is building one's own business or participating in society through governing bodies. In addition, the appearance of a large amount of non-informative data is, as always, the fault of humanity itself, because every entrepreneur seeks to preserve as much information as possible about the state of their business and at the same time strives to remain competitive, which is a vital factor for any enterprise.

In view of all this, methods for Internet content analysis must be constantly developed to facilitate finding the necessary data conveniently, and most importantly, quickly. The most expensive commodity today is information, and at the moment the main sources are social networks, through which people share their thoughts and plans. However, to use these posts for specific purposes, the data must be analyzed.

The collection and processing of data related to the financial sphere is of particular interest, especially the exchange rates of various currencies and cryptocurrencies. Cryptocurrency is gaining more popularity every day due to the relative ease of entry and the large amount of recommended information available about the process. Buying and selling cryptocurrency is an interesting process because certain conditions are met, an individual can increase their wealth several times, or even replace their main job with this activity. However, to truly make money from this process, it is necessary to conduct research regarding the chosen cryptocurrency, as well as its exchange rate and news concerning it.

The relevance of this research is due to the growing popularity of investing in cryptocurrency. Posts by famous people who have their own interest in this process have a significant influence on the formation of the prices of certain cryptocurrencies. When traders create forecasts regarding changes in the exchange rate of certain cryptocurrencies, they need an information system that can analyze the impact of such publications on changes in the exchange rates of cryptocurrencies and provide recommendations on market behaviour. This will increase the accuracy of the prediction created.

This work presents research that can be used by financial market participants to obtain high-quality forecasts of cryptocurrency rates, based on which they can make decisions about its purchase (or sale).

*e-mail: gnowakowski@pk.edu.pl

Manuscript submitted 2023-06-01, revised 2024-02-26, initially accepted for publication 2024-04-04, published in July 2024.

2. ANALYSIS OF LITERARY SOURCES AND FORMULATION OF THE PROBLEM

In general, the task of analyzing Internet posts is very important, because a well-analyzed publication can provide much more information than a simple question to the author such as: "What did you mean by this publication?". Detailed analysis of posts on social networks will allow for obtaining information about preferences and professional activity, as well as the users' circle of communication and their mutual influence.

In the paper [1], the process of the computer detection and categorization of opinions expressed in a piece of text to determine whether the writer's attitude towards a certain topic, product, etc., is positive, negative, or neutral is examined. Within the framework of the research presented in this paper, a detailed study of the analysis of moods and the cause-and-effect relationship of moods was carried out. In addition, using sentiment analysis, a generalized event can be identified based on mood and time. The results of the causality analysis can be used not only to determine the causes and effects, respectively, but also for their subsequent prediction of user sentiments. The main part of the publication is an overview of the combination of these two approaches, combined into a model that allows for determining the mood during future events, as well as creating a time forecast about the length of the interval between certain events. The average relative error was used to assess accuracy.

To view posts, you need to choose a place where a high number are available and stored in a single text format. A social network such as Twitter is useful for this. Special linguistic analysis and Twitter statistics are discussed in detail in [2]. This study aimed to identify criminal elements in the United States by modelling discussion topics and then incorporating them into a crime prediction model. A study was conducted on the impact of social media posts on future crimes.

In [3], a comprehensive reference for researchers and practitioners was considered, as well as coverage of all areas that contribute to the construction and analysis of social networks.

The paper [4] presents an integrated framework that offers the infrastructure needed to access, integrate, and analyze multilingual user-generated content from various social media sites.

The paper [5] demonstrates that Twitter messages (tweets) can be reliably classified based on flu-related keywords; the spread of flu can be predicted with high accuracy; and there is a way to monitor the spread of flu in selected cities in real time. We propose an approach to efficiently mine and extract data from Twitter streams, reliably classify tweets based on their sentiment, and visualize the data using an interactive real-time map.

The paper [6] shows what topics are being voiced by individuals and groups about the pandemic. It is determined whether there are any noticeable thematic trends, and if so, how these topics change over time and in response to important events. Using an improved sequential latent distribution model, the twelve most popular topics present in the Twitter dataset collected from April 3 to 13, 2020 in the United States are identified and their growth and changes are discussed.

It is also impossible not to highlight Internet blogs, in which many people express their own opinions and visions of certain problems, etc. Therefore, in [7], a study was conducted on the identification of hate groups. The proposed approach is semi-automatic and consists of four modules, namely: blog spider, information retrieval, network analysis, and visualization. This investigation was conducted on the blogging site Xanga. The results of the analysis revealed some interesting demographic and topological characteristics in hate groups and identified at least two large communities in addition to the smaller ones. The proposed approach is also appropriate for the examination of hate groups and other related blog communities.

In terms of business and the financial market, the process of analyzing large amounts of data and understanding the needs of the majority of people is very important, as it directly affects the income of the company and individuals. In [8], a study of the dominant factors that lead to currency crises was conducted. Within the framework of the research presented in this publication, the nature and characteristics of currency crises were identified and the forecasting of possible currency crises at an early stage was conducted. This can save managers some time in improving crisis management policies and corrective actions.

The work of [9] investigated the dynamics of linear and nonlinear serial dependencies in financial time series within the framework of a moving window. In particular, the focus was on identifying episodes of statistically significant two- and three-tribal correlation in the returns of several leading exchange rates, which may offer some potential for their predictability. A moving-window approach was used to capture correlation dynamics for different window lengths and to analyze the distribution of periods with statistically significant correlations. It was found that for sufficiently large window lengths, these distributions correspond well to the power law. Predictability itself was measured by the hit rate, i.e. the level of agreement between the actual return features and their predictions obtained using a simple correlation-based predictor.

It should be noted that in all the cited works, the research is general and the results of forecasting currency rates, in particular cryptocurrencies, were not provided. Accordingly, the factors affecting them were not investigated.

In [10], a study of the main macroeconomic indicators of the influence on the US dollar exchange rate in Ukraine was carried out, considering the purchase/sale of cash and non-cash currency, the balance of these, inflation in the current year, and nominal and real gross domestic product, as well as purchases/sales by bank clients, transactions between banks, gross and net international reserves, unemployment rates, and accounting (interest) rates, in addition to the balance of foreign exchange interventions, and the volume of nominal value transactions. Using the method of main components, the main economic components of the formation of the exchange rate were determined. With the help of an autoregressive integrated moving average (ARIMA), exponential smoothing, and singular spectrum analysis statistical models, the values of the selected influencing factors were predicted. The values of currency rates were predicted using regression models built by fast tree, fast forest, fast tree Tweedie, and generalized additive model algorithms, and the obtained values were studied for accuracy. This work did not forecast the exchange rates of cryptocurrencies in particular and did not

investigate the influence of factors such as posts on social networks.

The paper [11] analyzed methods, areas of application, and approaches to the analysis of publications and forecasting events based on the collected data, as well as the concept of the influence of publications on changes in the cryptocurrency exchange rate. The justification of the topicality of the topic was presented and the possibilities of appropriate application of the results of the work were described. The main stages of working with event forecasting data were defined, namely: the pre-processing of data, their further analysis, and forecasting. This work did not investigate the level of influence of posts by famous people on social networks on the cryptocurrency exchange rate. Within the framework of the research presented in the previously mentioned papers [10, 11], information systems were created to implement the above-described tasks of intellectual data analysis.

The above discussion indicates that the influence of certain factors on the exchange rate of cryptocurrencies, especially that of posts by famous people on social networks, has not been sufficiently studied to date and requires further study.

3. THE PURPOSE AND OBJECTIVES OF THIS STUDY

The purpose of this study is to examine the level of influence of posts by famous people on social networks on the cryptocurrency exchange rate. This will make it possible to increase the reliability of the forecast of the exchange rate of cryptocurrencies. To achieve this purpose, the following aims were set:

- To form statistical samples based on real exchange rates of the selected cryptocurrency, as well as their forecasts based on ARIMA, exponential smooth algorithms, and an algorithm for taking into account posts on social networks (ATAPSN).
- To calculate forecasting errors for all considered models and choose the most accurately forecasted cryptocurrency rates.
- To establish the level of dependence between the predicted and real exchange rates of cryptocurrency.
- To determine the statistical significance of the forecasted cryptocurrency exchange rates.

4. RESEARCH MATERIALS AND METHODS

The object of this study is the level of influence of posts on social networks on the exchange rate of cryptocurrencies. The necessary information includes forecasts for a selected period, taking into account the influence of posts by famous people on social networks (ATAPSN) [11], using the ARIMA [12] and exponential smoothing algorithms [13], as well as real cryptocurrency rates for the same period.

Time-limited impact: The first and most important argument is that publications and posts have the greatest impact on decision-making during the period of greatest interest. In the first hours, days, or weeks after publication, information can be very relevant and important for decision-making.

Patterns of influence: Some studies show that the impact of publications and posts on decision-making decreases rapidly over time. People may react more quickly to news and information that has appeared recently, and their reactions may be more intense.

Competition for attention: In an information society, competition for consumer attention is intense. Over time, new events and information can displace old ones, and the impact of publications can be significantly reduced.

Importance of freshness of information: Some topics or industries may require fresh information to make accurate decisions. For example, in the financial sector, information can become outdated quickly, and making decisions based on outdated information can be more risky.

Of course, the length of the period for a study depends on the specific context and objectives of the study. For example, when providing certain services, it is necessary to take into account the opinions of users from social networks based on their posts.

The application of a mathematical apparatus using the rank correlation coefficient allows for using this information to determine the relationship between the real and predicted cryptocurrency rates. Therefore, the following information is required to form the dataset: the real exchange rates of the chosen cryptocurrency for a certain period, as well as those predicted using various algorithms.

For ATAPSN, well-known individuals who are knowledgeable in the field of finance in general and cryptocurrencies in particular, or whose activities are somehow related to a certain cryptocurrency, were chosen as experts. As part of the conducted research, the Dogecoin cryptocurrency was selected to form the dataset and Elon Musk's posts on the Twitter social network were taken into account.

A fragment of the dataset is given in Table 1.

Table 1
A fragment of the input dataset

Hours	Real rates	Algorithm ATAPSN	Algorithm ARIMA	Exponential smoothing
1	467	466.19	497.8707174	502.2597
2	475	473.44	502.1439974	463.9021
3	516	513.78	506.4172773	490.0742
4	533	534.78	510.6905573	515.4008
5	508	508.31	514.9638372	523.1350
6	510	508.60	519.2371172	528.5218
7	525	525.67	523.5103971	518.3018
8	512	510.85	527.7836771	515.8382
9	514	512.55	532.0569570	529.6971
10	514	515.43	536.3302370	521.8438

Table 1 shows forecasts of Dogecoin rates, which were obtained using three algorithms (ATAPSN, ARIMA, exponential smoothing), for 10 hours, as well as its real rates for the specified period. The data were taken from the site of the crypto exchange Binance [14].

The dataset assembled in this way formed the input information for this study. As part of the conducted research, the following steps were necessary:

- Obtain forecasts based on ATAPSN and classic ARIMA and exponential smoothing algorithms and then form statistical samples based on them.
- Calculate forecasting errors for all considered models and choose the most accurately forecasted cryptocurrency rates.
- Establish the level of dependence between the predicted and real exchange rates of cryptocurrency using Spearman and Pearson correlation coefficients.
- Determine the statistical significance of the predicted cryptocurrency rate using the Student's t-test.

The use of the specified methods of statistical analysis guarantees obtaining reliable results when forecasting cryptocurrency rates and researching how they are influenced by the posts of famous people on social networks.

Software was developed to perform statistical analysis and obtain results based on the specified methods. It consisted of two main parts: the client (site) and the server. The client part was implemented using the Angular framework and was designed to display the resulting data, as well as to configure the search part.

The server part was implemented using the Python language and consisted of the following main blocks: the search block of publications – with the main goal to collect data about the publications of a certain person for a certain time; the cryptocurrency search block – chiefly intended to collect data on cryptocurrency exchange rates for the specified period; the algorithmic block – aimed at analyzing and forecasting future cryptocurrency rates; the communication unit – with the main purpose of data transfer between the client and server part and other blocks.

5. STUDY OF THE LEVEL OF INFLUENCE OF POSTS BY FAMOUS PEOPLE ON SOCIAL NETWORKS ON THE CRYPTOCURRENCY RATE

5.1. Formation of statistical samples based on real rates of the selected cryptocurrency, as well as their forecasts based on the ARIMA, exponential smoothing, and ATAPSN algorithms

Formulation of the problem: Based on the input dataset, form the following statistical samples X, Y_i $(i = \overline{1,3})$ of volume n each (n is the number of experiments (forecasts) made during the selected time [15]. The value of n is set depending on the desired accuracy of forecasting: hourly, daily, etc.):

X is a set of real cryptocurrency rates.

Y_1 is a set of forecasted cryptocurrency rates obtained using ATAPSN.

Y_2 is a set of forecasted cryptocurrency rates obtained using the ARIMA algorithm.

Y_3 is a set of predicted cryptocurrency rates obtained using the exponential smoothing algorithm.

Justification. Algorithm for forecasting the cryptocurrency exchange rate taking into account posts on social networks (ATAPSN) [11]. The idea of the algorithm is to calculate the coefficient of significance of the expert's post c_j, which is calculated according to the formula:

$$c_j = k_j \cdot ch_j, \tag{1}$$

where ch_j is the assessment of the tonality of the expert's post:

$$ch_j = \begin{cases} 1, & \text{if the post is positive,} \\ 0, & \text{if the post is neutral,} \\ -1, & \text{if the post is negative,} \end{cases} \tag{2}$$

k_j is the correctness of the forecast made at the previous moment in time:

$$k_j = \left| y_j - x_j \right|, \tag{3}$$

where y is the predicted value of the cryptocurrency exchange rate obtained using time series; x_j is the actual value of the cryptocurrency exchange rate; j is the moment in time.

After determining the coefficient c_j from (1)–(3) a forecast of the change in the cryptocurrency exchange rate will be created based on the available data on the expert's posts in the selected social network for the specified time:

$$y'_{j+1} = y_{j+1} + c_j. \tag{4}$$

Algorithm for forecasting the cryptocurrency exchange rate using the ARIMA algorithm [12]. The ARIMA(p, d, q) model for a non-standard time series looks like this:

$$\Delta^d y_j = c + \sum_{k=1}^{p} a_k \Delta^d y_{j-k} + \sum_{k=1}^{q} b_l \varepsilon_{j-l} + \varepsilon_j, \tag{5}$$

where ε_j is a stationary time series; c, a_k, b_l are model parameters; Δ^d is a time series difference operator of order d.

Algorithm for forecasting the cryptocurrency exchange rate using the exponential smoothing algorithm [13]. The exponential smoothing model is as follows:

$$y'_j = \begin{cases} y_1 : & j = 1, \\ y'_{j-1} + \alpha \left(y_j - y'_{j-1} \right) : & j > 1, \end{cases} \tag{6}$$

where y'_j is smoothed series; y_j is primary series; α is the smoothing coefficient.

Statistical samples X, Y_i $(i = \overline{1,3})$ of volume n each are formed based on forecasts obtained n times during the considered time using formulas (4)–(6).

The representation of cryptocurrency rates, both real and forecasted, in the form of a statistical sample, facilitates the application of statistical analysis methods to calculate forecasting errors, determine the level of dependence between forecasted and real rates of the selected cryptocurrency, and establish the statistical significance of forecasted cryptocurrency rates.

The choice of forecasting algorithms is determined by the need to compare forecasts using ATAPSN [11], the focus of this study, with classical forecasts made utilizing time series [16] including ARIMA and exponential smoothing [12, 13].

It should also be noted that the quality of the obtained samples depends on the accuracy of forecasting. Therefore, before conducting this research, to obtain correct forecasts of cryptocurrency exchange rates, it is necessary to conduct pre-processing of the data (reject abnormal forecasts, establish a sufficient volume of samples, normalize data, etc.).

Result. Based on the dataset presented in Table 1, the following statistical samples in volume were obtained $n = 10$:

$$X = (467.0, 475.0, 516.0, 533.0, 508.0, 510.0, 525.0, 512.0, 514.0, 514.0),$$
$$Y_1 = (466.19, 473.44, 513.78, 534.78, 508.31, 508.6, 525.67, 510.85, 512.55, 515.43),$$
$$Y_2 = (497.8707174289546, 502.14399737658755, 506.4172773242205, 510.69055727185344, 514.9638372194864, 519.2371171671194, 523.5103971147524, 527.7836770623854, 532.0569570100184, 536.3302369576514),$$
$$Y_3 = (502.259726, 463.902197, 490.07416, 515.400778, 523.134984, 528.521778, 518.301774, 515.838192, 529.697092, 521.84384).$$

5.2. Calculating the forecasting error for all considered models and choosing the most accurately forecasted cryptocurrency rates

Formulation of the problem. Let the pairs of statistical samples (X, Y_i), $i = \overline{1,3}$, be given (see Section 5.1).

For each such pair of samples, it is necessary to calculate the deviation of the elements of the sample X from the corresponding elements of the sample Y_i.

Justification. The deviation of the elements of the sample X from the corresponding elements of the sample Y_i is calculated as a relative error according to the formula:

$$R_{ij} = \frac{|x_j - y_{ij}|}{x_j} 100\%, \tag{7}$$

where x_j are elements of the sample X; y_{ij} are elements of the sample Y_i; $i = \overline{1,3}$, $j = \overline{1,n}$; n is the volume of samples X and Y_i [17].

Then, the average relative error of the sample is calculated by the formula:

$$R_i = \frac{\sum_{j=1}^{n} R_{ij}}{n}. \tag{8}$$

The proposed approach makes it possible to assess the accuracy of cryptocurrency forecasts to assess the quality of samples Y_1, Y_2, and Y_3. It also allows for choosing from the proposed algorithms the one that gives the highest accuracy, to further use the forecasts obtained with the help of this particular algorithm.

Results. Values for relative errors and their average values obtained using formulas (7) and (8) for the samples presented in Section 5.1 are provided in Table 2.

Table 2
Values of forecasting errors

X	Y_1	Y_2	Y_3
467	466.19	497.8707174	502.2597
475	473.44	502.1439974	463.9021
516	513.78	506.4172773	490.0742
533	534.78	510.6905573	515.4008
508	508.31	514.9638372	523.1350
510	508.60	519.2371172	528.5218
525	525.67	523.5103971	518.3018
512	510.85	527.7836771	515.8382
514	512.55	532.0569570	529.6971
514	515.43	536.3302370	521.8438
R_i	0.23406855	3.27736587	3.1429464

In Table 2, it is easy to see that the sample Y_1 has the smallest average relative error (0.23%), followed by the sample Y_3 (3.14%), and Y_2 has the largest error (3.28%).

5.3. Setting the level of dependence between the predicted and real cryptocurrency rates

Formulation of the problem. Let the pairs of statistical samples (X, Y_i), $i = \overline{1,3}$, be given (see Section 5.1).

It is necessary to establish the level of dependence between pairs of samples (X, Y_i), taking into account the sequence of elements in the sample.

Justification. In general, the Pearson correlation coefficient is used to establish the level of dependence between two samples [18]:

$$r_i = \frac{\sum_{j=1}^{n} (x_i - \bar{x})(y_{ij} - \bar{y_j})}{\sum_{j=1}^{n} (x_i - \bar{x})^2 \sum_{j=1}^{n} (y_{ij} - \bar{y_j})^2}, \tag{9}$$

where x_j are elements of sample X; y_{ij} are elements of the sample Y_i; $\bar{x}$, $\bar{y_j}$ are their sample averages; $i = \overline{1,3}$; $j = \overline{1,n}$; n is the volume of samples X and Y_i.

Since the sequence of items in the samples is important for this research (real and forecasted cryptocurrency rates are considered and compared at the appropriate time), using the rank correlation coefficient, in particular the Spearman coefficient, is recommended [19].

Spearman's rank correlation coefficient is calculated using the formula:

$$\rho_i = 1 - \frac{6 \sum_j d_j^2 + T_X + T_{Y_i}}{n(n^2 - 1)}, \tag{10}$$

where $d_j = x_j - y_{ij}$ is the rank difference for the j-th element of the sample Y_i; $i = \overline{1,3}$; $j = \overline{1,n}$; n is the volume of samples X and Y_i.

If there are connected ranks in the samples, i.e. the ranks of the elements of the sample are repeated:

$$T_X = \frac{N_X^3 - N_X}{12} \quad \text{and} \quad T_{Y_i} = \frac{N_{Y_i}^3 - N_{Y_i}}{12}, \tag{11}$$

(N_X and N_{Y_i}, respectively, is the number of repeated ranks in the samples X and Y_i, $i = \overline{1,3}$). For a sample that does not have repeated elements, the corresponding coefficient from formula (11) is zero.

The value of both correlation coefficients lies in the interval $[-1, 1]$.

It should be noted that Spearman's rank correlation coefficient (10) is less accurate compared to Pearson's correlation coefficient (9) since its calculation does not take into account the quantitative values of sample elements, but only their order. This indicates the need to consider the Pearson correlation coefficient in a more in-depth determination of the level of dependence of samples X and Y_i, $i = \overline{1,3}$. It is implemented using Table 3.

Table 3

Chaddock's scale of identifying the strength of the relationship by the value of the paired correlation coefficient (rank correlation)

The value of the correlation coefficient (by module)	0.1–0.3	0.3–0.5	0.5–0.7	0.7–0.9	0.9–0.99
Characteristics of the bond strength	Weak	Moderate	Notable	Strong	Very strong

Note that if the correlation coefficient (rank correlation) is equal to 1, there is a very strong (up to functional) directly proportional relationship between random variables. If it is equal to –1, then there is a very strong (to functional) inversely proportional relationship between random variables.

The proposed approach makes it possible to assess how close a correlation exists between the real and forecasted exchange rates of the chosen cryptocurrency for the specified period of time.

Result. For the samples presented in Section 5.1, the results are shown in Table 4.

In this table, the correlation coefficients are calculated according to formulas (9)–(11), and the strength of the connection with the sample X is established using Table 3.

The data given in Table 2 emphasizes that the sample Y_1 has the strongest correlation with the sample X among all considered samples (very strong relationship). The sample Y_2 has an average level of connection with the sample X, while the sample Y_3 has the worst indicators.

5.4. Calculating the forecasting error for all considered models and choosing the most accurately forecasted cryptocurrency rates

Formulation of the problem. Let (X, Y_i), $i = \overline{1,3}$, be a given pair of samples. We know the coefficient of rank correlation between them (Spearman's rank correlation coefficient ρ_i or Pearson correlation coefficient r_i, the values of which were obtained in Section 5.3).

Table 4

Correlation analysis between pairs of samples (X, Y_i), $i = \overline{1,3}$

The sample	Spearman's rank correlation coefficient	Pearson's correlation coefficient	The strength of bond with the sample X
Y_1	0.97828283	0.99829258	A very strong connection both in terms of the similarity of the elements of the samples and in terms of taking into account their order of follow-up
Y_2	0.39040404	0.581244474	A noticeable connection from the point of view of the similarity of sample elements, and moderate from the point of view of taking into account their sequence
Y_3	0.09949495	0.58275805	A noticeable connection in terms of the similarity of the elements of the samples, but almost absent in terms of taking into account their order of follow-up

It is necessary to test the hypothesis about the significance of the corresponding rank correlation coefficient at the level of significance α.

Justification. To solve the formulated problem, the following rules were used, which are part of the Student's t-test [20, 21].

Rule 1. To test the hypothesis about the significance of the Spearman rank correlation coefficient at the significance level α, it is necessary to calculate the observed value

$$t_i = t_{\text{cr}}(\alpha, k)\sqrt{\frac{1-\rho_i^2}{n-2}}, \tag{12}$$

where the critical value $t_{\text{cr}}(\alpha, k)$ is taken from the Student distribution table according to the level of significance α and the degree of freedom $k = n-2$; ρ_i is the Spearman's rank correlation coefficient; $i = \overline{1,3}$, n is the volume of samples X and Y_i.

If $|\rho_i| \le t_i$, then the hypothesis is accepted. If not, it is rejected, that is, there is a significant correlation between the samples.

Rule 2. To test the hypothesis of the significance of the Pearson correlation coefficient at the significance level α, it is necessary to calculate the observed value:

$$t_i = \sqrt{\frac{r_i^2(n-2)}{1-r_i^2}}, \tag{13}$$

where n is the volume of samples X and Y_i; r_i is Pearson correlation coefficient; $i = \overline{1,3}$.

Next, it is necessary to compare it with the tabular critical value of this criterion $t_{\text{cr}}(\alpha, k)$ (which is taken from the Student

distribution table [20, 21] taking into account the given level of significance α, which is a sufficient level to obtain reliable results, and the number of degrees of freedom $k = n - 2$).

If $t_i \leq t_{cr}(\alpha, k)$, then the hypothesis is accepted. If not, it is rejected, that is, there is a significant correlation between the samples.

By a significant correlation between the predicted rates of cryptocurrency, i.e. samples Y_i are well consistent with real rates at similar moments of time, i.e. with the sample X. In particular, if we consider a pair of samples (X, Y_1), within the framework of the proposed approach, it is possible to track how much the forecast of the cryptocurrency exchange rate, taking into account the posts by famous people on social networks, correlates with the real exchange rates for a similar period. This makes it possible to assess the level of influence of these posts on cryptocurrency rates, which is the main goal of this study.

Result. For the samples presented in Section 5.1, taking into account the rank correlation coefficients defined in Section 5.3, the results are shown in Tables 5 and 6.

Table 5

Results of testing the hypothesis about the significance of the Spearman rank correlation coefficient between pairs of samples (X, Y_i) at $\alpha = 0.05$ and $t_{cr}(0.05, 8) = 2.31$

The sample	Spearman's rank correlation coefficient ρ_i	Value of t_i	Conclusions about accepting or rejecting the hypothesis
Y_1	0.97828283	0.16	Significant correlation
Y_2	0.39040404	0.75	Non-significant correlation
Y_3	0.09949495	0.81	Non-significant correlation

Table 6

Results of testing the hypothesis about the significance of the Pearson rank correlation coefficient between pairs of samples (X, Y_i) at $\alpha = 0.05$ and $t_{cr}(0.05, 8) = 2.31$

The sample	Pearson correlation coefficient r_i	Value of t_i	Conclusions about accepting or rejecting the hypothesis
Y_1	0.99829258	48.34	Significant correlation
Y_2	0.581244474	2.02	Non-significant correlation
Y_3	0.58275805	2.03	Non-significant correlation

In these tables, correlation coefficients are calculated according to formulas (9)–(11), and critical values according to formulas (12) and (13), respectively.

The calculation results shown in Tables 5 and 6 confirm that there is a significant correlation in the pair of samples (X, Y_1), while the correlation in the pairs of samples (X, Y_2) and (X, Y_3) is insignificant.

6. THE INTRACTABLE PROBLEM OF MINIMIZING THE TOTAL TARDINESS OF PARALLEL MACHINE COMPLETION TIMES REGARDING THE COMMON DUE DATE WITH MACHINE RELEASE TIMES

The use of statistical information, the formation of a dataset (Table 1), and the calculation of informative indicators based on the methods of statistical analysis allow for solving the formulated research tasks.

It is worth emphasizing that the fact that both samples were obtained from real and predicted cryptocurrency rates over the same period and at the same moments is crucial. In addition, the order in which they were obtained should be taken into account. This makes it possible to justify the legitimacy of the transition from time series to statistical samples and, thus, to correctly calculate the average relative forecasting error.

The accuracy of the obtained results depends on the size and quality of samples X, Y_1, Y_2, and Y_3, presented in Section 5.1. In this regard, it is recommended to carry out thorough data processing before starting the research.

The classic ARIMA and exponential smoothing algorithms, which have proven themselves in solving forecasting problems, were used to create samples Y_2 and Y_3. This study showed that these algorithms do not make it possible to track the influence of any specific factor, unlike the ATAPSN algorithm, based on which the sample Y_1 was created. The advantage of the proposed method of presenting the dataset (Table 1) in the form of statistical samples is the possibility of expanding the area of use of the results of this study far beyond the boundaries of finance. An example of the application of the proposed approach to the analysis of public services is given in [22]. This approach was used to study the relationship between some pillars of European regional competitiveness depending on the quality of regional institutions [23].

The proposed approach can also be applied in various studies related to the design, implementation, management, and development of services and in general supporting the life cycle of services in high-tech industries [24, 27], and to determine which appliances can be considered for the Demand Response programme [25].

Checking the accuracy of the received forecasts, and therefore the quality of samples Y_1, Y_2, and Y_3, was carried out by calculating the average value of the relative error (see Section 5.2). The results of the analysis of the received forecasting errors showed that forecasts based on the sample Y_1 are the most accurate (error 0.23%) (see Table 2). That is, the algorithm, which takes into account posts by famous people on social networks, makes it possible to obtain better results in terms of accuracy.

Considering the fact that the input dataset (see Table 1) was formed from predicted and real cryptocurrency rates for a certain period when studying the correlation between pairs of samples (X, Y_i), $i = \overline{1, 3}$, it is appropriate to use rank correlation coefficients. The Pearson correlation coefficient was also utilized to clarify the obtained results.

According to the results obtained in Section 5.3, it can be stated that the largest correlation exists between the samples from the pair (X, Y_1), which indicates that the predictions of the

Dogecoin cryptocurrency exchange rates, taking into account the posts by Elon Musk on social networks, are most consistent with the real values of these rates. At the same time, it is worth noting that a very strong correlation between samples from the pair (X, Y_1) was confirmed using both correlation coefficients, which guarantees the correctness and accuracy of the obtained results.

Obtaining high-quality forecasts by the ATAPSN algorithm, which are used to form the sample Y_1, depends on the correctness of the selected experts, and for this purpose, it is necessary to choose the social network that best suits the specifics of the study [26].

In addition, the data given in Table 4 confirm that for pairs of samples (X, Y_2) and (X, Y_3) Spearman's rank correlation coefficient is smaller than Pearson's correlation coefficient. This is because the elements in samples Y_2 and Y_3 do not differ too much from the elements of the sample X, which is confirmed by the value of 3% relative error (see Table 2) for both samples. But the orders of placement of elements for pairs of samples (X, Y_2) and (X, Y_3) have moderate and almost no correlations, respectively, because samples Y_2 and Y_3 were formed using algorithms that are based on the use of time series that are not able to track the influence of a specific factor on the dependent variable, in contrast to ATAPSN. Therefore, such forecasts achieved a sufficient level of accuracy, but at the same time, they did not take into account the trends of constant growth or decline of cryptocurrency rates.

The assessment of the correlation between the real and predicted cryptocurrency rates can also be carried out using the Kendall rank correlation coefficient [19], which could be the subject of further research.

To check the statistical significance of the Pearson correlation coefficient (Spearman's rank correlation) (see Section 5.4), classical stationary criteria were used at the significance level $\alpha = 0.05$ (see Tables 5 and 6). As a result of testing these hypotheses, it was established that the existing correlation is significant in the pair of samples (X, Y_1), and this fact was confirmed by two statistical criteria at once. This means that a factor such as a post made by famous people on social networks significantly affects the formation of the cryptocurrency rate at the time when the post was made. It should also be noted that for sample pairs (X, Y_2) and (X, Y_3), the correlation is insignificant.

The advantage of the performed research is the relative simplicity in the formation of the input dataset and the correctness of the proposed methods of statistical analysis. In addition, this approach can be used when checking the level of influence of any other factor on the formation of the rate of the selected cryptocurrency.

Disadvantages include the need to prepare a high-quality dataset since an insufficient number of forecasts, the presence of anomalous data (atypical forecasts), or inaccurate forecasts in the input dataset can significantly affect the accuracy of the results. It should also be noted that the accuracy of the forecast can be negatively affected by an incorrectly selected time interval for which the forecast is made since it is not known in advance how long the expert's post will affect the cryptocurrency rate. This indicates the need for the constant monitoring of both cryptocurrency rates and experts' posts on social networks.

An alternative way to determine the level of influence of one or another factor on the cryptocurrency exchange rate is the method of principal components. But to use this method, it is necessary to have a clear list of all factors that in one way or another affect the cryptocurrency rate. That is, the process of forming the input dataset would be much more complicated. In addition, this method would be more difficult from the point of view of software implementation.

7. CONCLUSIONS

1. As part of this research, it has been established that posts by famous people are an influential factor in the formation of the cryptocurrency exchange rate.
2. For this study, a model was created that best suited the goals and objectives of the research. The authors did not consider such well-known models as Granger Causality and LSTM, as they contradict the main goal of the study, which is to identify the impact of celebrity posts on cryptocurrency rates.
3. Since the current research investigates the influence of only one factor, namely, the posts of celebrities on social media on the cryptocurrency rate, there is no need to use multivariate models that would complicate the forecasting process.
4. The obtained results have been used to increase the accuracy of forecasting cryptocurrency rates.
5. The reliability of the obtained results is guaranteed by the quality of the input dataset and the correctness of the statistical analysis methods used.
6. In the course of the research, software has been developed that, in combination with other software developed by the authors of the study, can form an information system for forecasting cryptocurrency rates and studying the impact of various factors on their formation.
7. In further research, the authors plan to study the impact of other factors on cryptocurrency rates and investigate the impact of posts by several experts on a particular cryptocurrency, as well as the impact of posts by the same expert (group of experts) on the rates of different cryptocurrencies. In addition, it is planned to consider a wider range of cryptocurrencies and influencers to increase the generalizability and relevance of the study, conduct sentiment analysis to determine the tone and sentiment of social media posts and their correlation with market movements and conduct comparative case studies involving different social media platforms or financial markets to provide a broader context for the findings.

ACKNOWLEDGEMENTS

Funding: This research was funded by the Faculty of Electrical and Computer Engineering, Cracow University of Technology, and the Ministry of Science and Higher Education, Republic of Poland (grant no. E-1/2024).

REFERENCES

[1] P. Preethi, V. Uma, and A. Kumar, "Temporal Sentiment Analysis and Causal Rules Extraction from Tweets for Event Prediction," *Procedia Comput. Sci.*, vol. 48, pp. 84–89, 2015, doi: 10.1016/j.procs.2015.04.154.

[2] M.S. Gerber, "Predicting crime using Twitter and kernel density estimation," *Decis. Support Syst.*, vol. 61, pp. 115–125, 2014, doi: 10.1016/j.dss.2014.02.003.

[3] R. Alhajj and J. Rokne, *Encyclopedia of Social Network Analysis and Mining*, Springer, 2020, p. 2200, doi: 10.1007/978-1-4614-7163-9.

[4] Y. Dang *et al.*, "An integrated framework for analyzing multilingual content in Web 2.0 social media," *Decis. Support Syst.*, vol. 61, pp. 126–135, 2014, doi: 10.1016/j.dss.2014.02.004.

[5] K. Byrd, A. Mansurov, and O. Baysal, "Mining Twitter data for influenza detection and surveillance," *SEHS '16: Proceedings of the International Workshop on Software Engineering in Healthcare Systems*, 2016, pp. 43–49, doi: 10.1145/2897683.2897693.

[6] G. ChengHe, "Dynamic topic modeling of twitter data during the COVID-19 pandemic," *PLoS ONE,* vol. 17, no. 5, p. 22, 2022, doi: 10.7910/DVN/YXIAEK.

[7] M. Chau and J. Xu, "Mining communities and their relationships in blogs: A study of online hate groups," *Int. J. Hum.-Comput. Stud.*, vol. 65, no. 1, pp. 57–70, 2007.

[8] D. Karahoca, A. Karahoca, and Ö. Yavuz, "An early warning system approach for the identification of currency crises with data mining techniques," *Neural Comput. Appl.*, vol. 23, pp. 2471–2479, 2013, doi: 10.1007/s00521-012-1206-9.

[9] M. Žukovič, "Dynamics of episodic transient correlations in currency exchange rate returns and their predictability," *Open Phys.*, vol. 10, pp. 615–624, 2012, doi: 10.2478/s11534-011-0120-6.

[10] O. Gavrilenko, K. Novakivska, and O. Shumeiko, "Selection of the most influential economic factors for forecasting the US dollar exchange rate," *NTU Bull.*, vol. 54, pp. 26–35, 2022, doi: 10.33744/2308-6645-2022-4-54-026-035.

[11] O. Gavrylenko, M. Miahkyi, and Y. Zhurakovskyi, "The task of analyzing publications to build a forecast for changes in cryptocurrency rates," *Adapt. Syst. Autom. Control,* vol. 2 no. 41, pp. 90–99, 2022, doi: 10.20535/1560-8956.41.2022.271349.

[12] ARIMA Model – Complete Guide to Time Series Forecasting in Python – Resources [Online] Available: https://www.machinelearningplus.com/%20time-series/arima-model-time-series-forecasting-python/. (Accessed 05.04.2024).

[13] R.J. Hyndman, A.B. Koehler, J.K. Ord, and R. D. Snyder, *Forecasting with Exponential Smoothing: The State Space Approach*. Springer, 2008, p. 362, doi: 10.1007/978-3-540-71918-2.

[14] Binance cryptocurrency exchange – Resources [Online] Available: https://www.binance.com/ (Accessed 05.04.2024).

[15] M. Kartashov, *Probability, processes, statistics*. Kyiv University, 2007, p. 504.

[16] O. Valenzuela, F. Rojas, L.J. Herrera, H. Pomares, and I. Rojas, *Theory and Applications of Time Series Analysis and Forecasting*. Springer, 2023, p. 333, doi: 10.1007/978-3-031-14197-3.

[17] A.D. Helfrick and W.D. Cooper, *Modern Electronic Instrumentation and Measurement Techniques*. Prentice Hall of India, 2008, p. 460.

[18] H. Abdi, V. Guillemot, A. Eslami, and D. Beaton, *Canonical Correlation Analysis*. Springer, 2017, pp. 1–16, doi: 10.1007/978-1-4614-7163-9_110191-1.

[19] Rank correlation – Resources [Online] Available: https://moodle.znu.edu.ua/ (Accessed 05.04.2024).

[20] W.J. Ewens and K. Brumberg, *Introductory Statistics for Data Analysis*. Springer, 2023, p. 273, doi: 10.1007/978-3-031-28189-1.

[21] B. Blaine, *Introductory Applied Statistics*. Springer, 2023, p. 190, doi: 10.1007/978-3-031-27741-2.

[22] O. Gavrilenko *et al.*, "The principle for forming a portfolio of public services based on the analysis of statistical information," *East.-Eur. J. Enterprise Technol.*, vol. 3, no. 3(117), pp. 57–64, 2022, doi: 10.15587/1729-4061.2022.260136.

[23] J. Ascorbebcitia, E. Ferrcira, and S. Orbc, *Testing conditional multivariate rank correlations: the effect of institutional quality on factors influencing competitiveness*. Springer, 2022, pp. 931–949, doi: 10.1007/s11749-022-00806-1.

[24] V. Chymshyr, S. Telenyk, Î. Rolik, and E. Zharikov, "The platform for supporting the life cycle of services in the information systems of information and communication service providers," *Adapt. Syst. Autom. Control*, vol. 1 no. 42, pp. 205–226, 2023, doi: 10.20535/1560-8956.42.2023.279172.

[25] P. Kapler, "The application of the fuzzy rule-based Bayesian algorithm to determine which residential appliances can be considered for the demand response program," *Bull. Pol. Acad. Sci. Tech. Sci.*, vol. 71, no. 4, p. e146106, 2023, doi: 10.24425/bpasts.2023.146106.

[26] O. Gavrilenko and M. Myagkyi, "Forecasting the cryptocurrency exchange rate based on the ranking of expert opinions," *Innov. Technol. Sci. Sol. Ind.*, no. 1, no. 27, pp. 18–25, 2024, doi: 10.30837/ITSSI.2024.27.018.

[27] G. Nowakowski, S. Telenyk, and Y. Vovk, "Chatbots Lifecycle Support Platform," *2023 IEEE 12th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, Germany, 2023, pp. 308–319, doi: 10.1109/IDAACS58523.2023.10348794.

2025 IEEE 13th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS) | 979-8-3315-8045-2/25/$31.00 ©2025 IEEE | DOI: 10.1109/IDAACS68557.2025.11322226

The 13th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications
4-6 September, 2025, Gliwice, Poland

Joint Punctuation Restoration and Text Capitalisation with a Hybrid XLM-RoBERTa–LSTM Model

Volodymyr Shymkovych [1], Grzegorz Nowakowski [2], Sergii Telenyk [1,2]
[1] Department of Information Systems and Technologies, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Prospect Beresteiskyi 37, Kyiv, 03056, Ukraine, v.shymkovych@kpi.ua
[2] Faculty of Electrical and Computer Engineering, Cracow University of Technology, Warszawska 24, Cracow, 31-155, Poland, grzegorz.nowakowski.edu.pl, stelenyk@pk.edu.pl

Abstract— **Punctuation restoration is a common post-processing task in text generation and automatic speech recognition systems. It is crucial for enhancing the readability and interpretability of transcribed text for human users. Recent state-of-the-art approaches tackle this problem using various deep learning architectures, with transformer-based models demonstrating particularly strong performance. In this work, we propose a hybrid model that combines transformer architectures with long short-term memory (LSTM) layers, leveraging both global attention mechanisms and sequential dependencies. The model is trained on the IWSLT 2012 dataset and achieves results comparable to current state-of-the-art systems. Its performance is summarised by the following metrics: accuracy - 0.96, precision - 0.93, recall - 0.92, and F1-score - 0.92. Our findings suggest that punctuation restoration serves as a valuable auxiliary task, enhancing structural understanding of text and contributing to more robust natural language representations.**

Keywords*— *machine learning, natural language processing, punctuation restoration, token classification, transformers, neural networks

I. Introduction

The rapid advancement of artificial intelligence technologies and associated implementation tools has facilitated their widespread adoption across diverse domains of human activity, significantly contributing to their increased sophistication [1–3]. Among these developments, machine learning has emerged as a particularly prominent area, experiencing remarkable growth and popularity in recent years [4, 5, 21]. Notably, machine learning methods and technologies have achieved substantial progress in fields such as computer vision [6] and natural language processing (NLP) [7].

One of the key areas in NLP is the processing of voice data [8, 9]. However, text generated by most speech recognition systems typically lacks punctuation, which hinders subsequent text processing tasks [9]. Similar challenges arise in other forms of textual data, such as user-generated content on websites, where punctuation may be incomplete, inconsistent, or entirely missing. Incomplete or ambiguous linguistic data also poses a well-known challenge in information systems and database querying [10], necessitating robust interpretation mechanisms.

Punctuation structures grammatical constructs and disambiguates sentence meaning in written language. It plays a crucial role in all aspects of text perception, particularly with regard to readability and comprehension, both for humans and for downstream NLP models. This is especially important in dialogue systems, which have recently seen widespread adoption.

A prominent solution to this problem is punctuation restoration (also referred to as *punctuation correction*), which aims to automatically reinsert appropriate punctuation marks into unstructured or poorly punctuated text. Recent work also demonstrates improved accuracy when jointly leveraging text and audio modalities for punctuation recovery [24]. Despite its significance, punctuation restoration remains a complex and relatively underexplored task within NLP. Moreover, the effectiveness of downstream processes—such as text summarisation, machine translation, and overall text readability—largely depends on the accuracy of punctuation recovery.

In general, the literature on capitalisation and punctuation restoration—such as [9]—identifies seven principal methodological approaches: rule-based methods, *n*-gram language models, treating capitalisation as a discriminative task, hidden event language models, boosting-based methods, probabilistic models based on conditional random fields (CRFs), and neural network approaches. Among these, the two most widely adopted techniques are CRFs and neural networks.

Punctuation restoration typically involves several core steps: preprocessing the input text, tokenising it into words or subword units, and applying a machine learning model to recover a predefined set of punctuation marks. However, many existing models fail to account for the full range of punctuation symbols. This limitation often

stems from the infrequent use of certain marks—such as semicolons—which makes it difficult for models to learn reliable insertion patterns and to adjust the associated capitalisation accordingly.

II. LITERATURE REVIEW AND PROBLEM STATEMENT

A range of methods has been proposed to address the task of punctuation restoration in text. One notable example is DeepCorrect [11], a widely recognised neural network model specifically designed for punctuation recovery. DeepCorrect follows the sequence-to-sequence (Seq2Seq) paradigm, transforming an input sequence of word tokens into an output sequence of punctuation labels—such as full stops, commas, or placeholders indicating the absence of punctuation—assigned after each word.

The model architecture employs bidirectional Long Short-Term Memory (BiLSTM) layers in both the encoding and decoding stages. This design allows the model to capture contextual dependencies in both forward and backward directions, thereby enhancing its ability to infer appropriate punctuation marks based on the surrounding words within a sentence.

Another widely recognised approach is punctuation restoration using transformer models [12], which combines a transformer-based encoder—specifically, a Bidirectional Encoder Representations from Transformers (BERT) model—with a bidirectional Long Short-Term Memory (BiLSTM) decoder. The transformer component encodes rich contextual information from the input text, while the BiLSTM layer processes this representation to predict appropriate punctuation marks, leveraging the full sentence context. Further improvements have been achieved by leveraging external data to enhance transformer-based punctuation recovery for ASR transcripts [23].

Another noteworthy contribution is FullStop [13], a collection of punctuation restoration models based on fine-tuning pre-trained BERT-type transformers. These models demonstrate that transfer learning from large language models can be effectively adapted to punctuation recovery tasks, often requiring only minimal architectural modifications.

An important feature of the FullStop models is that they were trained on the Europarl dataset, which contains parallel texts in multiple languages, including English, French, Italian, and German. Comparative studies have shown that models trained on multilingual corpora outperform those trained on monolingual datasets in terms of punctuation restoration accuracy. This suggests that exposure to cross-linguistic structures can improve a model's generalisation ability.

It is also worth noting that the aforementioned models are available as Python software libraries, which facilitates their integration into NLP pipelines. However, none of the reviewed approaches fully integrates all state-of-the-art techniques for punctuation restoration. Moreover, many existing models remain primarily research prototypes, developed to evaluate specific methods rather than to provide production-ready solutions. As such, they often require further refinement and optimisation before they can be effectively deployed in real-world applications.

Article [14] presents a multilingual approach to sentence segmentation and punctuation prediction. The objective is to train NLP models capable of identifying sentence boundaries and inserting appropriate punctuation marks into automatically generated or transcribed text. The authors demonstrate that these tasks benefit significantly from cross-lingual transfer by leveraging multilingual deep language models. Their model achieves an average F1-score of 0.94 for sentence boundary detection and 0.78 for punctuation prediction across English, German, French, and Italian texts.

For the analysis of texts from social networks, sentence segmentation is also critically important, as most social media texts are poorly punctuated. The study in [15] proposes a punctuation restoration algorithm based on a transformer model architecture. Various BERT-based models were evaluated as encoders, in combination with neural network classifiers. Among the tested configurations, RoBERTa-large combined with an LSTM decoder achieved the highest accuracy: 97% on the Amazon dataset and 90% on the Telekom dataset.

In [16], the authors present an approach to automatic punctuation restoration using BERT models for English and Hungarian. For English, experiments were conducted on the TED Talks dataset, a widely used benchmark in this domain. For Hungarian, evaluation was carried out on the Szeged Treebank dataset. The best-performing models achieved macro-averaged F1-scores of 79.8 for English and 82.2 for Hungarian.

Despite the impressive generative capabilities of recent large language models, their ability to capture syntactic and semantic structure remains limited. The authors of [17] hypothesise that this limitation stems from insufficient modelling of linguistic structure in current pre-training objectives. Focusing on English, they demonstrate that using punctuation restoration as an auxiliary learning objective improves performance on structure-related tasks such as named entity recognition, open information extraction, chunking, and part-of-speech tagging. In 16 out of 18 experiments across six of seven tasks, punctuation restoration led to an average performance improvement of 2%, highlighting its effectiveness in enhancing structure-aware language representations in base-sized models. Additional research explores context-aware multimodal generation using enhanced retrieval pipelines [22].

In [18], the authors investigate various deep learning-based punctuation restoration approaches for the Ukrainian language. The selected model employs a Bidirectional Recurrent Neural Network (BiRNN) to estimate the probabilities of potential punctuation

placements. The input data consists of publicly available text corpora. Through experimentation, the authors identify an effective text processing strategy and demonstrate a significant improvement in punctuation prediction when incorporating word embeddings. The model achieved an F1-score of 51.8% and a Sentence Error Rate (SER) of 66.2% on the test set.

Most contemporary approaches are limited to predicting only a few of the most frequently used punctuation marks—typically full stops, commas, and question marks—and only a single punctuation mark per word. However, natural written language employs a broader range of punctuation characters (e.g., parentheses, hyphens) and combinations (e.g., a parenthesis followed by a full stop), which cannot always be unambiguously reduced to a basic subset

In [19], the authors evaluate several models on the task of comprehensive punctuation reconstruction. They conduct experiments on parallel corpora in two typologically distinct languages: English, with relatively simple morphology, and Polish, which exhibits complex morphological structure. They also analyse the impact of comprehensive punctuation modelling on the performance of basic punctuation restoration tasks.

Building upon insights from the reviewed literature, this article proposes and investigates a novel approach to punctuation restoration that integrates the most effective techniques identified in prior work. At the core of the proposed method is a hybrid model that combines three well-established components: a bidirectional LSTM layer, fine-tuning of a BERT-type transformer model, and support for multilingual input. This combination is designed to enhance contextual understanding, capture sequential dependencies, and improve generalisation across languages.

III. THE STUDY MATERIALS AND METHODS

The input to the proposed model is a sequence of tokens $x = \{x_1, x_2, \dots, x_n\}$, where n denotes the sequence length. Each token x_i may represent a complete word, a subword unit, a special symbol (e.g., start-of-sequence or end-of-sequence markers), or a padding token used for sequence alignment.

The task of punctuation restoration is formulated as a token classification problem: for each token x_i, where $i = 1, \dots, n$, the model estimates the probability that it belongs to one of m predefined classes $j = 1, \dots, m$, where each class corresponds to a specific punctuation mark or capitalization label. As a result, the model outputs a matrix $y \in R^{n \times m}$, where each row represents the predicted class distribution for the corresponding token.

This formulation frames punctuation and capitalisation recovery as a multi-class classification problem at the token level, making it well suited for implementation using neural sequence models.

As mentioned earlier, tokens in the input sequence may correspond to entire words, parts of words, or special symbols that do not represent linguistic content. Consequently, not all tokens require classification. To account for this, we introduce an additional binary mask sequence $y_{mask} = \{y_{mask,1}, y_{mask,2}, \dots, y_{mask,n}\}$ of the same length n as the input sequence x. Each element $y_{mask,i}$ is set to 1 if the corresponding token x_i represents a complete word or the final subword unit of a word; otherwise, $y_{mask,i} = 0$.

The binary sequence y_{mask}, defined as described above, is used during model training to control which tokens contribute to the loss computation and parameter updates. Specifically, if $y_{mask,i} = 0$ the classification output $\hat{y}_i$ for token x_i is excluded from the loss calculation. Conversely, if $y_{mask,i} = 1$, the output $\hat{y}_i$ is included in the loss function and used to update the model parameters.

In addition, the input sequence x may contain padding tokens, which do not carry any linguistic information and are used solely to equalise input lengths within a batch. These tokens should also be ignored during punctuation prediction. To handle this, we introduce another binary sequence $x_{mask} = \{x_{mask,1}, x_{mask,2}, \dots, x_{mask,n}\}$, where $x_{mask,i} = 1$ indicates that token x_i contains meaningful input, and $x_{mask,i} = 0$ denotes padding. During both training and inference, both y_{mask} and x_{mask} are applied to ensure that only relevant tokens influence model predictions and learning.

If $x_{mask,i} = 0$, the corresponding input token x_i is excluded from the model's computation of the output y. This ensures that padding tokens do not influence the prediction or training process.

Training a punctuation restoration model requires the selection of an appropriate corpus for both training and evaluation. In this work, we use a dataset derived from the IWSLT 2012 corpus [20], which contains transcriptions of TED Talks lectures. This dataset spans a wide range of topics and has been widely used in prior research on punctuation restoration. Importantly, training on transcribed speech aligns with the intended use case of the model—enhancing the readability of text produced by automatic speech recognition systems.

The dataset used in this study contains over 140,000 text sequences. Each sequence is presented on a separate line and may consist of one or more complete sentences.

A. *Dataset structure*

The dataset is structured as a table with four fields. Its core component is a list of tokens x, generated from the input text using a tokenizer. The tokenizer used in this work relies on a vocabulary of 250,002 tokens, so each element in the list is represented as an integer in the range [0, 250,001], corresponding to a token index in the vocabulary.

The second crucial field is the list of punctuation and capitalisation classes y, aligned with the tokens in x. Each element in y is an integer in the range [0,9], representing a specific punctuation–capitalisation

combination. These classes are detailed in Table I.

The third field, x_{mask}, is a binary mask indicating whether each token in the sequence corresponds to actual input text or a padding token. After tokenisation, padding tokens are appended to ensure consistent sequence lengths across batches. Each element in x_{mask} takes a value of 1 (valid token) or 0 (padding).

The fourth field, y_{mask}, is another binary mask that indicates whether a given token is the final subword unit of a word. Since punctuation marks are placed only after complete words, only tokens for which $y_{mask,i} = 1$ are relevant during training and inference. Elements of this mask take the values 1 (end of word) or 0 (not end of word or padding).

TABLE I. DESCRIPTION OF DATASET CLASSES

№	Label	Punctuation	Capitalized word?
0	_	Absent	No
1	,	,	No
2	.	.	No
3	?	?	No
4	-	-	No
5	_Up	Absent	Yes
6	,_Up	,	Yes
7	._Up	.	Yes
8	?_Up	?	Yes
9	-_Up	-	Yes

B. Model architecture and training

The proposed model consists of three main components:

- *XLM-RoBERTa Transformer* – a pre-trained multilingual transformer model (*xlm-roberta-base*) that encodes high-level contextual representations of the input sentence. This component captures general linguistic and semantic information for each token.
- *Bidirectional LSTM* – a BiLSTM layer that processes the contextual embeddings from the transformer and captures sequential dependencies in both forward and backward directions. This layer decodes token-level information relevant for punctuation and capitalisation prediction.
- *Linear Classification Layer* – a fully connected layer that maps the output of the BiLSTM to one of the 10 predefined punctuation–capitalisation classes (as described in Table I).

The detailed architecture and dimensionality of each layer are presented in Table II.

TABLE II. MODEL ARCHITECTURE

Layer name	Input data size	Output data size	Number of hidden layers
XLM-RoBERTa Transformer	384	768	12
Bidirectional LSTM	768	1536	1
Linear Layer	1536	10	-

To train the model, the *Adafactor* optimiser was employed. This choice was motivated by practical considerations: preliminary experiments using the *Adam* optimiser resulted in excessive memory consumption and instability during training. *Adafactor* is particularly well suited for large transformer-based models due to its significantly lower memory footprint.

The model's loss was computed using the standard *CrossEntropyLoss* function, commonly applied in multi-class classification tasks. Since the *PyTorch implementation* of *CrossEntropyLoss* internally includes a softmax operation, the model outputs raw (unnormalised) scores. During inference, the class with the highest score is selected for each token to determine the predicted punctuation and capitalisation label.

To improve training efficiency, a staged training schedule was adopted. The full model—including the XLM-RoBERTa encoder—was trained during the first epoch. In the remaining three epochs, only the BiLSTM and linear classification layers were updated, while the transformer component was frozen. This strategy reduced training time while maintaining high performance.

The model was developed and trained in Google Colaboratory using a Tesla T4 GPU. Total training time was approximately 20 minutes.

IV. RESEARCH RESULTS

While accuracy is a widely used metric for evaluating model performance, relying on it exclusively can lead to misleading conclusions—particularly in tasks involving imbalanced data. Accuracy measures the ratio of correct predictions to the total number of predictions. However, it has several notable limitations when used as a standalone metric:

- *Imbalanced Data*: Real-world datasets are rarely balanced. In scenarios where one class is significantly more prevalent than others, a model may achieve high accuracy by simply predicting the majority class, without truly learning meaningful patterns. Therefore, accuracy should always be interpreted in the context of class distribution and complemented with additional metrics.
- *Error Analysis*: Understanding the types of errors a model makes is essential for improving its performance. A detailed confusion matrix provides insight into class-wise misclassifications, helping to identify systematic weaknesses and refine the model accordingly.

The performance of the model was evaluated using standard classification metrics: accuracy, precision, recall, and F1-score. In addition, a confusion matrix was generated to analyse class-wise prediction discrepancies.

We begin with accuracy, which reflects the overall proportion of correct predictions across all tokens. It is computed as shown in Equation (1):

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

where TP (*true positives*) and TN (*true negatives*) denote the number of correctly predicted positive and negative instances, respectively, and FP (*false positives*) and FN (*false negatives*) represent the number of incorrect positive and negative predictions.

The second metric, precision, measures the proportion of predicted positive instances that are actually correct. A higher precision indicates a lower false positive rate and greater reliability in the model's predictions for a given class. It is computed as shown in Equation (2):

$$precision = \frac{TP}{TP+FP} \quad (2)$$

The third metric, recall, quantifies the proportion of actual positive instances that are correctly identified by the model. A higher recall value indicates that the model successfully captures most relevant instances of a given class. It is calculated as shown in Equation (3):

$$recall = \frac{TP}{TP+FN} \quad (3)$$

The fourth metric, the F1-score, is the harmonic mean of precision and recall. It provides a balanced measure that accounts for both false positives and false negatives, making it particularly useful when working with imbalanced datasets. The F1-score is calculated using the following formula, shown in Equation (4):

$$\text{F1-score} = 2 \cdot \frac{precision \cdot recall}{precision+recall} \quad (4)$$

The confusion matrix provides a detailed breakdown of the model's predictions by showing how often the model predicted class i when the true class was j. Accordingly, the values along the main diagonal of the matrix represent the number of correct classifications for each class. The higher these diagonal values, the better the model's performance, as they indicate a greater number of accurate predictions for the corresponding classes. The confusion matrix for the proposed model is presented in Table III.

TABLE III. MODEL CONFUSION MATRIX

Label \ *Class*	*Prediction*									
	_	,	.	?	-	_Up	,_Up	._Up	?_Up	-_Up
_	31630	112	26	21	206	124	90	3	3	8
,	98	3003	13	7	192	7	79	5	0	10
.	5	53	2325	19	89	1	2	20	7	5
?	0	3	11	259	4	0	0	2	10	0
-	21	85	3	3	1562	1	2	0	0	5
_Up	105	20	2	1	14	4755	107	7	6	40
,_Up	0	15	0	0	7	21	734	16	7	28
._Up	0	0	12	0	1	1	11	307	12	6
?_Up	0	0	0	1	0	0	1	4	21	1
-_Up	0	1	0	0	1	3	10	2	0	89

The data presented in Table III provide several important insights into the behaviour and performance of the proposed model across individual punctuation–capitalisation classes.

First, the model demonstrates high reliability in predicting capitalisation, as most misclassifications occur within the same capitalisation group as the correct label (i.e., either within the lowercase or capitalised subset). This suggests that the model successfully learns structural cues associated with capitalisation.

Second, in some cases, the model correctly predicts the presence of punctuation but misidentifies the specific punctuation mark. This is particularly evident in the mid-range F1-scores for classes such as comma and dash, which are frequently confused due to their similar syntactic functions.

Third, the model appears to struggle with distinguishing between commas and dashes, especially in their lowercase variants. Both punctuation marks typically occur within a sentence and are not followed by a capital letter, which may contribute to the confusion.

Notably, the *?_Up* class exhibits the lowest precision (0.318) and F1-score (0.446), likely due to its low frequency in the training corpus. The rarity of this class in the IWSLT dataset limits the model's ability to learn a robust representation for it.

In addition to aggregate metrics, the model also reports class-wise precision, recall, and F1-score, providing more granular insight into its performance across individual punctuation–capitalisation classes. These per-class evaluation results are summarised in Table IV and are also presented in Figure 1.

TABLE IV. MODEL TESTING RESULTS FOR EACH CLASS

Class	*Indicator*		
	Precision	*Recall*	*F1-score*
_	0.993	0.982	0.987
,	0.912	0.880	0.896
.	0.972	0.920	0.946
?	0.833	0.896	0.863
-	0.752	0.929	0.831
_Up	0.995	0.940	0.967
,_Up	0.971	0.886	0.927
._Up	0.839	0.877	0.858
?_Up	0.318	0.750	0.447
-_Up	0.464	0.840	0.597

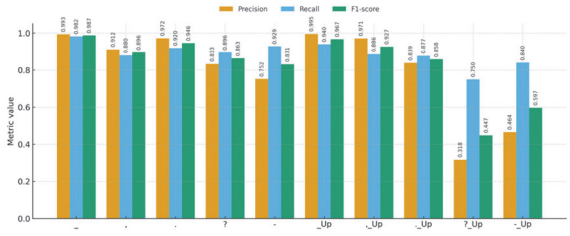


Figure 1. Per-class Precision/Recall/F1. Best F1 for no punctuation (0.987), _Up (0.967) and period (0.946); weakest for ?_Up (0.447) and -_Up (0.597).

TABLE V. COMPARISON OF XLM-ROBERTA-LSTM WITH EXISTING MODELS – PER-CLASS F1-SCORES

Model	*F1-score*				
	_	,	.	?	-
[13]	**0.991**	0.819	**0.948**	**0.890**	0.425
[14]	0.99	0.76	0.88	0.82	-
[15]	-	0.678	0.84	0.876	-
[17]	-	0.625	0.373	0.268	-
[18]	-	0.809	0.938	0.846	-
XLM-RoBERTa-LSTM	0.987	**0.896**	0.946	0.863	**0.831**

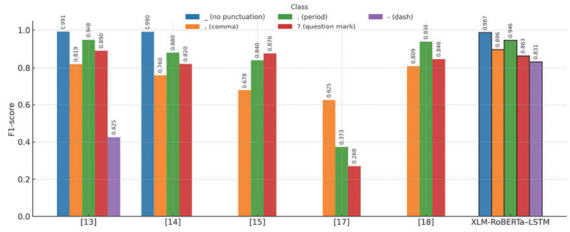


Figure 2. Per-class F1 comparison across models. Our XLM-RoBERTa–LSTM is top-1/top-2 in all five classes and shows the biggest gain for dash (0.831 vs ~0.43 in prior work).

Table V and Figure 2 present a per-class comparison of the proposed XLM-RoBERTa–LSTM model with existing approaches from the literature. The comparison is based on F1-scores for five punctuation–capitalisation classes: no punctuation (_), comma (,), full stop (.), question mark (?), and dash (–).

Table VI presents a comparison of the overall performance of the proposed model with existing approaches, using key evaluation metrics: accuracy, precision, recall, and F1-score, both including and excluding class 0 (the *no punctuation* baseline class). Including class 0 typically inflates evaluation scores due to class imbalance; therefore, both perspectives are reported for completeness and transparency.

TABLE VI. COMPARISON OF XLM-ROBERTA-LSTM WITH EXISTING MODELS – OVERALL METRICS

Model	*Accuracy/ + 0 class*	*Precision/ + 0 class*	*Recall/ + 0 class*	*F1-score/ + 0 class*
[13]	-	-	-	-/**0.94**
[14]	-/**0.97**	0.813/0.858	0.83/0.87	0.82/0.863
[15]	-	0.758/-	0.851/-	0.798/-
[17]	-	0.607/0.661	0.341/0.426	0.422/0.518
[18]	-	0.887/0.912	0.844/0.88	0.864/0.894
XLM-RoBERTa-LSTM	0.923/0.959	**0.906/0.927**	**0.899/0.919**	0.901/0.923

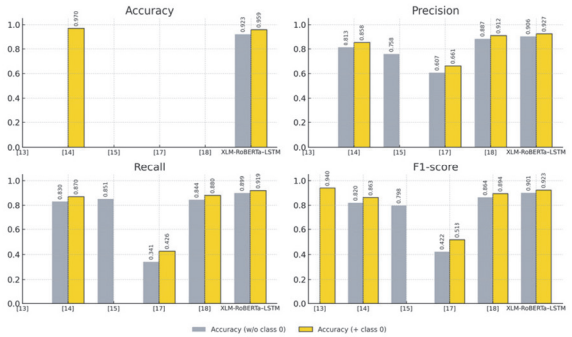


Figure 3. Overall Accuracy/Precision/Recall/F1 reported with and without class 0 ("no punctuation"). Our model reaches 0.923/0.959 (Acc), 0.906/0.927 (Prec), 0.899/0.919 (Rec), 0.901/0.923 (F1)—competitive with SOTA. Including class 0 inflates scores, so both views are shown.

It is important to note that the overall values for accuracy, precision, recall, and F1-score are computed by aggregating the true positives (TP), false positives (FP), and false negatives (FN) across all classes, *excluding* class 0. This exclusion is justified by the fact that class 0—representing the absence of punctuation and lowercase formatting—is heavily overrepresented in the dataset. Including this class in the aggregate metrics would artificially inflate performance scores due to the resulting class imbalance.

In contrast, a second set of metrics is calculated *including* class 0, providing a more general view of overall model performance across all tokens, regardless of class distribution. Both variants are reported in Table VI to provide a balanced and comprehensive perspective on the model's capabilities and are also presented in Figure 3.

An example of punctuation restoration and text capitalisation generated by the proposed XLM-RoBERTa–LSTM model is presented in Figure 4.

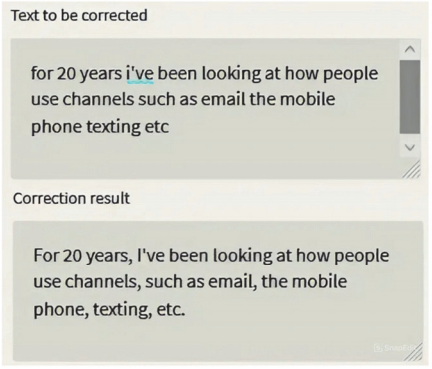


Figure 4. Example of punctuation and capitalisation correction performed by the XLM-RoBERTa-LSTM model.

V. CONCLUSIONS

In this work, we presented a neural network model for automatic punctuation restoration and text capitalisation in English-language texts. The proposed architecture consists of a pre-trained XLM-RoBERTa-base transformer encoder, a bidirectional LSTM layer for contextual decoding, and a linear classification layer that outputs punctuation–capitalisation classes.

The model was trained for four epochs using the CrossEntropyLoss function and the Adafactor optimiser. We reimplemented a state-of-the-art approach and evaluated it on the IWSLT 2012 TED Talks dataset. Notably, the model performs punctuation restoration and capitalisation prediction jointly, using a unified label space of 10 classes—five for lowercase and five for capitalised tokens, each paired with a specific punctuation mark or its absence.

Our model achieved results comparable to current state-of-the-art methods, demonstrating the effectiveness of combining transformer-based multilingual representations with sequential modelling.

Future work will focus on improving the efficiency of the XLM-RoBERTa–LSTM model. One important direction is the development of text augmentation methods tailored to punctuation restoration and capitalisation, which would help address class imbalance and significantly enhance accuracy on real-world data. Another key objective is the effective integration of the trained model into a production-level software application.

DECLARATION ON GENERATIVE AI

During the preparation of this work, the authors used ChatGPT in order to: grammar, spelling check and reword. After using this tool, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

REFERENCES

[1] J. Arcila-Diaz, D. Altamirano-Chavez, L. Arcila-Diaz, and C. Valdivia, "Real-time identification of rice leaf diseases using convolutional neural networks," *International Journal of Computing*, vol. 23, no. 4, pp. 709–714, Dec. 2024, doi: 10.47839/ijc.23.4.3773.

[2] P. Kravets, A. Novatskyi, V. Shymkovych, A. Rudakova, Y. Lebedenko, and H. Rudakova, "Neural network model for laboratory stand control system controller with parallel mechanisms," in *Lecture Notes on Data Engineering and Communications Technologies*, Springer Nature Switzerland, pp. 47–58, 2023, doi: 10.1007/978-3-031-36118-0_5.

[3] Golovko, V., Egor, M., Brich, A., Sachenko, A. (2017). A Shallow Convolutional Neural Network for Accurate Handwritten Digits Classification. In: *Krasnoproshin, V., Ablameyko, S. (eds) Pattern Recognition and Information Processing. PRIP 2016. Communications in Computer and Information Science*, vol 673. Springer, Cham. https://doi.org/10.1007/978-3-319-54220-1_8.

[4] M. Sirola and J. E. Hulsund, "Machine-learning methods in prognosis of ageing phenomena in nuclear power plant components," *International Journal of Computing*, vol. 20, no. 1, pp. 11–21, Mar. 30, 2021, doi: 10.47839/ijc.20.1.2086.

[5] V. Hamolia, V. Melnyk, P. Zhezhnych, and A. Shilinh, "Intrusion detection in computer networks using latent space representation and machine learning," *International Journal of Computing*, vol. 19, no. 3, pp. 442–448, 2020, doi: 10.47839/ijc.19.3.1893.

[6] K. Khotin, V. Shymkovych, P. Kravets, A. Novatsky, and L. Shymkovych, "Convolutional neural network for dog breed recognition system," *Adaptive Systems of Automatic Control*, vol. 2, no. 45, pp. 3–14, Oct. 10, 2024, doi: 10.20535/1560-8956.45.2024.313022.

[7] M. Mishchenko and M. Dorosh, "Detection of Windows portable executable malware using NLP techniques and proxy-server," *International Journal of Computing*, vol. 23, no. 4, pp. 663–672, Dec. 2024, doi: 10.47839/ijc.23.4.3765.

[8] M. Norval and Z. Wang, "Speech emotion recognition using hybrid architectures," *International Journal of Computing*, vol. 23, no. 1, pp. 1–10, 2024, doi: 10.47839/ijc.23.1.3430.

[9] H. Ahlawat, N. Aggarwal, and D. Gupta, "Automatic speech recognition: A survey of deep learning techniques and approaches," *International Journal of Cognitive Computing in Engineering*, vol. 6, pp. 201–237, Dec. 2025, doi: 10.1016/j.ijcce.2024.12.007.

[10] G. Nowakowski, "Fuzzy queries on relational databases," in *Proc. 2018 Int. Interdisciplinary PhD Workshop (IIPhDW)*, Świnoujście, Poland, 2018, pp. 293–299, doi: 10.1109/IIPHDW.2018.8388376.

[11] V. Păiş and D. Tufiş, "Capitalization and punctuation restoration: A survey," *Artificial Intelligence Review*, vol. 55, no. 3, pp. 1681–1722, 2021, doi: 10.1007/s10462-021-10051-x.

[12] P. Bedapudi, "DeepCorrection2: Automatic punctuation restoration," *Medium*, [Online]. Available: https://praneethbedapudi.medium.com/deepcorrection2-automatic-punctuation-restoration-ac4a837d92d9. [Accessed: Jul. 27, 2025].

[13] T. Alam, A. Khan, and F. Alam, "Punctuation restoration using transformer models for high- and low-resource languages," in Proc. 6th Workshop on Noisy User-generated Text (W-NUT 2020), Association for Computational Linguistics, 2020, doi: 10.18653/v1/2020.wnut-1.18.

[14] O. Guhr, A.-K. Schumann, F. Bahrmann, and H.-J. Böhme, "FullStop: Multilingual deep models for punctuation prediction," *CEUR Workshop Proc.*, vol. 2957, pp. 1–9, 2021.

[15] A. M. Bakare, K. S. M. Anbananthen, S. Muthaiyah, J. Krishnan, and S. Kannan, "Punctuation restoration with transformer model on social media data," *Applied Sciences*, vol. 13, no. 3, p. 1685, Jan. 28, 2023, doi: 10.3390/app13031685.

[16] A. Nagy, B. Bial, and J. Ács, "Automatic punctuation restoration with BERT models," *arXiv preprint*, arXiv:2101.07343, 2021, doi: 10.48550/arXiv.2101.07343.

[17] J. Min, M. Lee, W. Lee, and Y. Lee, "Punctuation restoration improves structure understanding without supervision," *arXiv preprint*, 2024, doi: 10.48550/arXiv.2402.08382.

[18] M. Sazhok, A. Poltieva, V. Robeiko, R. Seliukh, and D. Fedoryn, "Punctuation restoration for Ukrainian broadcast speech recognition system based on bidirectional recurrent neural network and word embeddings," *CEUR Workshop Proc.*, vol. 2870, pp. 300–310, 2021.

[19] M. Pogoda and T. Walkowiak, "Comprehensive punctuation restoration for English and Polish," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 4610–4619.

[20] M. Federico, M. Cettolo, L. Bentivogli, M. Paul, and S. Stüker, "Overview of the IWSLT 2012 evaluation campaign," in *Proc. 9th Int. Workshop on Spoken Language Translation: Evaluation Campaign*, 2012, pp. 12–33.

[21] V. Turchenko, L. Grandinetti and A. Sachenko, "Parallel batch pattern training of neural networks on computational clusters," *2012 International Conference on High Performance Computing & Simulation (HPCS), Madrid, Spain, 2012*, pp. 202-208, doi: 10.1109/HPCSim.2012.6266912.

[22] V. Chymshyr, O. Zhdanova, O. Havrylenko, et al. „Models and methods for forming service packages for solving of the problem

of designing services in information systems of providers", *Information, Computing and Intelligent systems, Art. nr 5, 2024*, doi: 10.20535/2786-8729.5.2024.316432.

[23] X.-Y. Fu, C. Chen, M. T. R. Laskar, S. Bhushan, and S. Corston-Oliver, "Improving punctuation restoration for speech transcripts via external data," in *Proc. 7th Workshop on Noisy User-generated Text (W-NUT 2021)*, pp. 168–174, 2021. doi: 10.18653/v1/2021.wnut-1.19.

[24] Y. Zhu, L. Wu, S. Cheng, and M. Wang, "Unified multimodal punctuation restoration framework for mixed-modality corpus (UniPunc)," arXiv preprint arXiv:2202.00468, 2022, doi: 10.48550/arXiv.2202.00468.

Generative Data Augmentation by Dataset Distillation

Yuri Gordienko[1(✉)], Grzegorz Nowakowski[2], Yuriy Kochura[1], Vladyslav Taran[1], and Sergii Stirenko[1]

[1] National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine
{gord,kochura,taran,stirenko}@comsys.kpi.ua
[2] Cracow University of Technology, Krakow, Poland
grzegorz.nowakowski@pk.edu.pl

Abstract. Dataset distillation aims to create compact synthetic datasets that retain the generalization properties of real datasets. This study employs dataset distillation by matching training trajectories (DDMTT), a novel approach that utilizes expert trajectories (precomputed sequences of network parameters trained on the full dataset) to guide the distillation process. Experiments with the extremely increased number of images per class (IPC) were conducted using standard datasets such as CIFAR-10 and CIFAR-100, as well as medical benchmarking datasets from MedMNIST. The proposed method of generative data augmentation by dataset distillation (GDADD) demonstrated that, for CIFAR datasets, their smaller distilled versions containing 40,000 images achieved higher validation accuracy than the full datasets with 50,000 images, surpassing the original dataset's performance by 3.1% for CIFAR-10 and 2.9% for CIFAR-100. For the considered MedMNIST datasets (PathMNIST, DermaMNIST, RetinaMNIST), some distilled datasets (PathMNIST) exceeded the performance of models trained on full datasets, confirming the method's robustness across different domains and demonstrating the better results for the well balanced datasets.

Keywords: dataset distillation · classification · data augmentation · CIFAR-10 · CIFAR-100 · MedMNIST · ConvNet · health care

1 Introduction

Dataset distillation (DD) has emerged as a critical technique to address the increasing computational and storage demands of deep learning (DL) approaches [13]. Unlike traditional deep neural networks (DNN) that require vast amounts of data and computing power [5,7–9,12,16,19,20,25], DD synthesizes a small set of highly informative data points that retain the knowledge of large datasets [22]. Models trained on these distilled datasets can achieve performance comparable to those trained on full datasets, making DD a promising approach for efficient

C. Tommasino et al. (Eds.): AIBIO 2025, CCIS 2696, pp. 105–118, 2026.
https://doi.org/10.1007/978-3-032-17216-7_9

learning and storage. Overall, dataset distillation presents an efficient alternative to traditional data reduction, offering improved model training efficiency, storage savings, and generalization through synthetic data generation. DD's ability to compress and abstract key information makes it an essential tool in optimizing DL across explainability, design, and healthcare while improving efficiency and security. In this context, the challenges of scalability, multimodality, labeling complexity for the real datasets used in the practical setups are not resolved yet fully. That is why the further research of DD approaches should be performed for the DD better understanding. In this work, several experiments were performed to investigate the limits of DD with regard to its application to the standard and specific datasets to bring to the light the new ways for data augmentation and their effect of the dataset quality on the performance of classification tasks.

2 Background and Related Work

Due to the high-dimensional nature of DL data, distilling knowledge into a few points is challenging. DD methods can be broadly divided in several categories [13]. In meta-learning approaches, distilled data are treated as hyperparameters optimized in a nested loop to improve generalization [22]. Data-matching methods, on the other hand, update synthetic data by imitating the influence of real data on model training, either in parameter space or feature space [2,30,31].

Factorized dataset distillation (DD) leverages the idea that high-dimensional image data lie on a low-dimensional manifold and can be reconstructed using specific decoders [1,13,28]. Instead of directly storing synthetic images, factorized DD optimizes compact feature representations (codes) along with corresponding decoders, reducing storage needs and redundancy. This approach improves compression while preserving essential information. The factorized DD is categorized into three types: code-based DD, decoder-based DD, and code-decoder DD, each offering different trade-offs in efficiency and reconstruction quality, making it a powerful strategy for dataset distillation.

DD is used across various domains that demand efficient training and storage. It has been particularly useful in continual learning [18,24] and neural architecture search [15,23], where handling large datasets is computationally expensive. Additionally, since dataset distillation preserves gradient information, it plays a role in privacy preservation, federated learning, and adversarial robustness by reducing the exposure of raw data during training [6,21]. One notable application of DD is in explainable AI (XAI) algorithms [17]. Due to their small size, synthetic datasets allow precise measurement of how training examples influence test predictions. If both test and training images rely on the same synthetic images, it becomes easier to trace model decision-making, making dataset distillation a valuable tool for XAI. DD has also been employed in visual design where DD used to generate representative textures by cropping synthetic images and to model outfit compatibility, leveraging synthetic images to capture essential visual patterns [3,4]. In the health care, DD facilitates medical image sharing where DD can enhance data anonymization and secure data exchange between

hospitals [14]. It allows healthcare institutions to collaboratively train AI models for computer-aided diagnosis with reduced privacy risks and lower costs.

3 Methodology

3.1 Datasets

The following benchmarking datasets were used in the work: 32×32 pixel CIFAR-10 and CIFAR-100 datasets that widely used for dataset distillation research [10,11], and 28×28 pixel specific MedMNIST datasets, actually PathMNIST, DermaMNIST, and RetinaMNIST [26,27].

3.2 Dataset Distillation by Matching Training Trajectories

Here the dataset distillation by matching training trajectories (DDMTT) is used which introduces a novel approach to dataset distillation by leveraging expert trajectories to guide the learning process [2]. Expert trajectories are sequences of network parameters obtained by training a neural network on the full, real dataset. These trajectories represent an upper bound on performance for the distilled dataset, as they reflect the ideal learning process of a model trained on real data. To implement DDMTT, a large number of networks are trained on the real dataset, and their parameter snapshots are recorded at each epoch. These precomputed expert trajectories serve as references for guiding the learning process of a student model, which is trained on synthetic images. The student parameters) evolve during training, and the goal is to distill a synthetic dataset that induces a trajectory matching the real dataset's trajectory, ensuring that models trained on the distilled data perform similarly to those trained on the full dataset. A key advantage of DDMTT is that expert trajectories are computed before distillation, enabling faster and more efficient training. This precomputation allows for rapid experimentation, as the same expert trajectories can be used across multiple distillation attempts. By explicitly aligning the synthetic dataset with long-range network parameter evolution, DDMTT produces high-quality distilled datasets, ensuring that models trained on them generalize similarly to models trained on real data.

3.3 Network Architecture

The network architecture used for dataset distillation is based on ConvNet architectures [26] and primarily follows the intentionally simple design to enable direct analysis of the effectiveness of the distillation method while ensuring comparability with previous studies [2]. The network consists of several convolutional blocks, where the number of convolutional blocks is adapted based on the dataset resolution, with specific configurations defined for each dataset used in experiments. Each convolutional block contains 3×3 convolutional layer with 128 filters, ReLU activation functions, 2×2 average pooling and stride 2 and the more

implementation details are given elsewhere [2]. By using this lightweight yet effective architecture, this work isolates the impact of the distillation method rather than introducing confounding effects from complex architectures. This design choice ensures that the results remain directly comparable to the previous works [2] and similar results [26,27], facilitating fair evaluation and benchmarking of the dataset distillation process.

3.4 Workflow

The dataset distillation workflow follows a standard evaluation protocol, where a randomly initialized neural network is trained from scratch on the distilled dataset and then evaluated on a validation set. This ensures a fair comparison of different distillation methods by assessing how well models generalize when trained on synthetic rather than real data. To generate distilled images, the workflow implements the distillation process and follows established techniques described in [2]. This includes applying a suite of differentiable augmentations, similar to those used in prior studies [29,30]. These augmentations help improve the generalization of the distilled dataset by introducing variations that mimic real-world transformations. The key hyperparameters for the distillation process, such as the number of real epochs per iteration, synthetic updates per iteration, and image learning rate, are documented elsewhere [2]. These hyperparameters are crucial in determining the efficacy of distillation, balancing training efficiency and synthetic data quality.

4 Results

4.1 Synthesized and Reconstructed Images

CIFAR Datasets. The dataset distillation results for CIFAR-10 (Fig. 1) and CIFAR-100 reveal that the synthetic images effectively capture class-specific information, even when highly compressed. The synthesized and reconstructed visualizations for 1 image per class (IPC) (Fig. 1, a and b) demonstrate vague, but recognizable representations, for example from the left to right: bird (1st), car (2nd), bird (3rd), cat (4th), deer (5th), dog (6th), ... for CIFAR-10 (Fig. 1). It indicates that the optimization process efficiently extracts key distinguishing features. Since the task is constrained to a single synthetic IPC, the model is forced to compress essential information into just one representation. When the constraint is relaxed to allow multiple synthetic IPC (e.g., 10 IPC), the optimization distributes the class's distinguishing features across multiple samples. This results in a more diverse and structured set of synthetic images (Fig. 1, c and d). For example, instead of a single compressed representation, the dataset contains varied examples of objects within a class, such as different versions of birds, cars, deers, etc. This outcome highlights that dataset distillation benefits from increased synthetic samples per class, as it enables the model to better approximate real-world diversity while maintaining a compact dataset size. The results support the effectiveness of dataset distillation in generating highly informative and expressive synthetic datasets with minimal storage requirements.

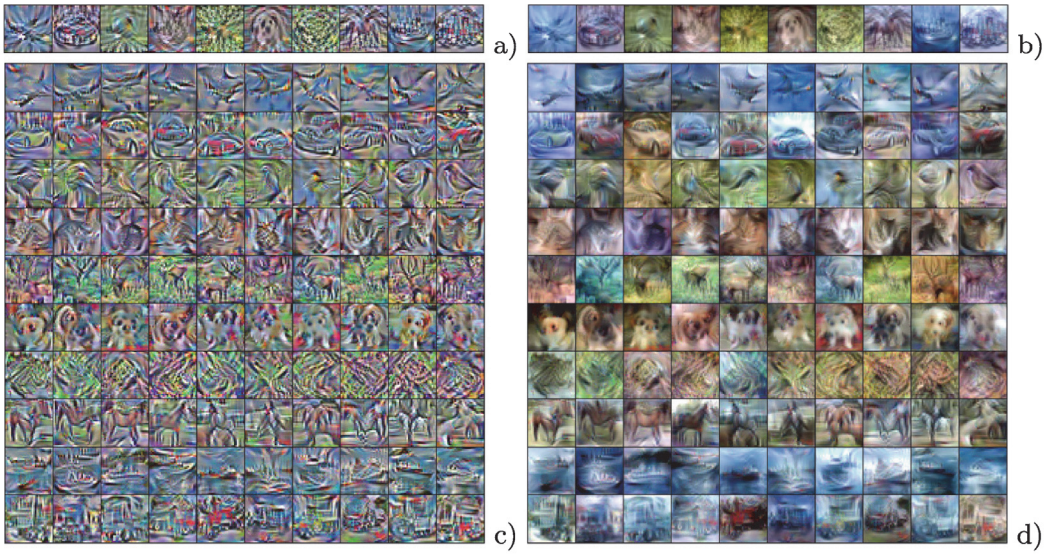


Fig. 1. Examples of synthesized (left) and reconstructed (right) images from CIFAR-10 dataset for various images per class: 1 (a,b) and 10 (c,d).

MedMNIST Datasets. But for the specific datasets like MedMNIST the synthesized and reconstructed images, for instance for PathMNIST (Fig. 2), are not so easily recognizable even for the higher values of IPCs due to the quite specific and complex patterns of some diseases.

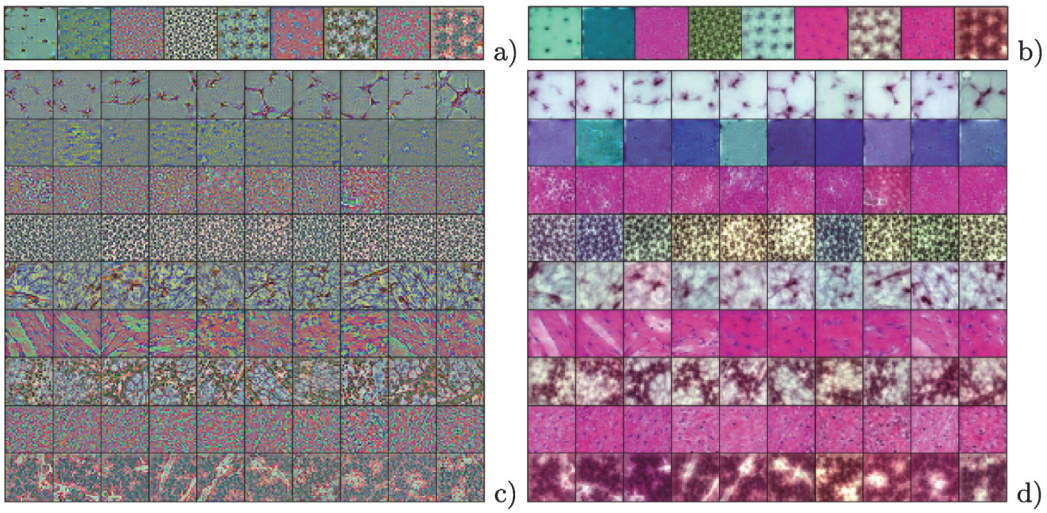


Fig. 2. Examples of synthesized (left) and reconstructed (right) images from PathMNIST dataset for various images per class: 1 (a,b) and 10 (c,d).

4.2 Impact of Distilled IPC on Model Performance: CIFAR

In this part of research 5 trials of training on the distilled IPC were performed for the different IPC numbers from 1 up to 10000 to estimate the impact of IPC number on the actual performance of the model. The saturation (i.e. no decrease of the validation accuracy) was observed for various numbers of epochs, for example, for CIFAR-10 (Fig. 3): for IPC from 1 to 100 - up to 5000 epochs, for IPC from 1000 to 2000 - up to 1000 epochs, for IPC from 4000 to 5000 - up to 500 epochs, and for IPC 10000 - up to 120 epochs.

CIFAR-10. For each IPC (from 1 up to 10000 for CIFAR-10) the maximal validation accuracy values were plotted versus the IPC values in Fig. 3 (right). It should be noted that training on the IPC numbers from 1 to 50 (that were used in [2]) give the same results with regard to the maximal validation accuracy value obtained by averaging over 5 trials after the number of epochs when accuracy saturation was observed. Moreover, after some IPC value (4000 here) the maximal accuracy (0.879 ± 0.001) becomes higher than the accuracy obtained after training on the full original dataset (0.848 ± 0.001) (denoted by "+" symbol in Fig. 3, right). Here, "+" symbol corresponds the IPC number (5000) for the full dataset but the improvement beyond the standard deviation (0.001) was observed for the lower IPC equal to 4000. It means that distilled version of CIFAR-100 dataset with 40000 training images can provide the better performance than the original version of CIFAR-100 dataset with 50000 training images. For CIFAR-10 the overall improvement of the validation accuracy by DD is better than $3.1 \pm 0.1\%$ (Table 1).

CIFAR-100. Similarly, for CIFAR-100, after some IPC value (400 here) the maximal accuracy (0.591 ± 0.003) becomes higher than the accuracy obtained after training on the full original dataset (0.562 ± 0.003) (denoted by "+" symbol in Fig. 4, right). Here, "+" symbol corresponds the IPC number (500) for the full dataset but the improvement beyond the standard deviation (0.003) was observed for the lower IPC equal to 400. Again, it means that distilled version of CIFAR-100 dataset with 40000 training images can provide the better performance than the original version of CIFAR-100 dataset with 50000 training images. For CIFAR-100 the overall improvement of the validation accuracy by DD is better than $2.9 \pm 0.3\%$ (Table 1).

In general, DD used for CIFAR-10 and CIFAR-100 as a generative data augmentation by dataset distillation (GDADD) technique allows to use smaller synthetic (40000 images) datasets in comparison to the larger original (50000 images) datasets and get the accuracy which is higher by $\sim 3 \pm 0.3\%$.

4.3 Impact of Distilled IPC on Model Performance: MedMNIST

It worth to note that the results reported above relate to the standard datasets and their potential practical value should be checked for the more specific dataset

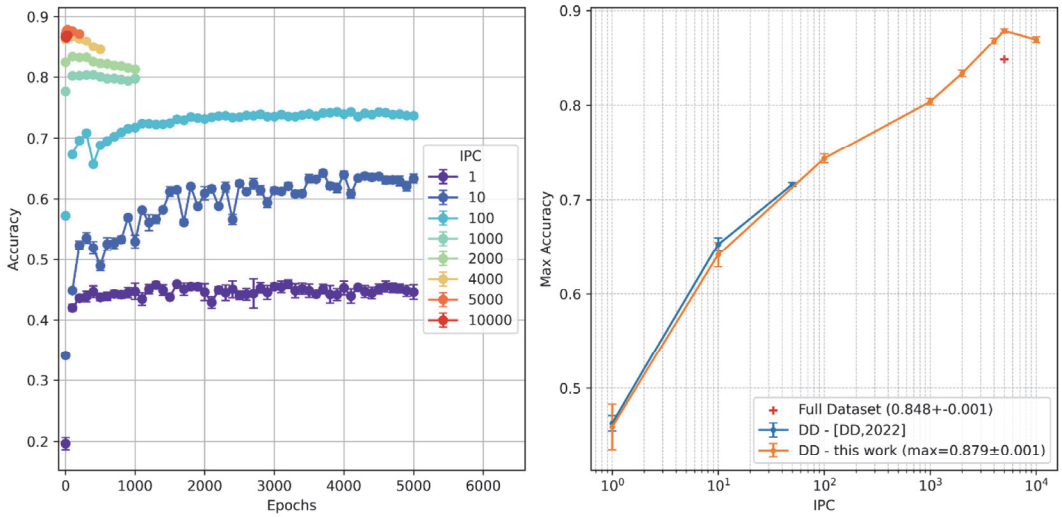


Fig. 3. CIFAR-10 dataset. Validation accuracy vs. epochs where the legend contains the distilled "images per class" (IPC) values (left) and maximal validation accuracy (observed for the stated IPC) vs. IPC (right) where the legend contains the reference values from [2]. The error bars denote the standard deviations of accuracy after 5 trials.

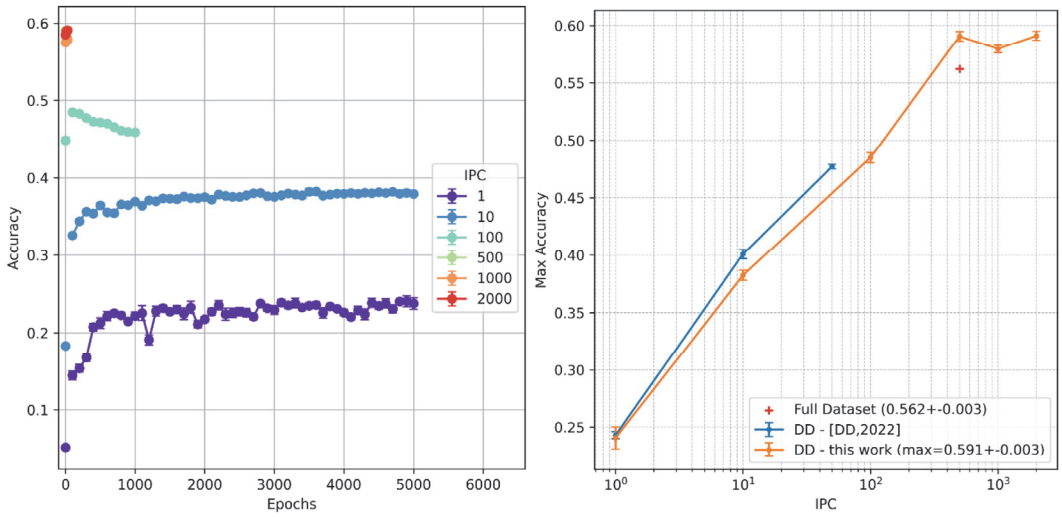


Fig. 4. CIFAR-100 dataset. Validation accuracy vs. epochs where the legend contains the distilled "images per class" (IPC) values (left) and maximal validation accuracy (observed for the stated IPC) vs. IPC (right) where the legend contains the reference values from [2]. The error bars denote the standard deviations of accuracy after 5 trials.

which are more closely reflect the real-world setups. For this purpose, the medical benchmarking datasets from MedMNIST collection were used [26,27].

PathMNIST. In this work, the performance improvement was obtained for ConvNet architecture and it is compared with the previously published results

Table 1. The maximal validation accuracy values for the balanced CIFAR datasets.

Experiment	CIFAR-10	CIFAR-100
DD, full original dataset [2]	0.848 ± 0.001	0.562 ± 0.003
Ours, distilled smaller dataset (Fig. 3, 4)	$\mathbf{0.879} \pm 0.001$	$\mathbf{0.591} \pm 0.003$

[26,27] obtained for various other methods including ResNet-18, ResNet-50, auto-sklearn, AutoKeras, and Google AutoML Vision (Fig. 5).

After some IPC value (100 here) the maximal accuracy values obtained by GDADD on the smaller synthetic version of PathMNIST dataset gradually become the better than accuracy values obtained by the aforementioned methods on the full original PathMNIST dataset. After some IPC value (10000 here) the maximal accuracy (0.917 ± 0.003) becomes higher than the best accuracy obtained after ResNet-50 training on the full original dataset (0.911) (denoted by the "·" symbol in Fig. 5, right). Here, "·" symbol corresponds the IPC number (10000) for the full dataset but the improvement at the brim of the standard deviation (0.003) was observed for the same IPC equal to 10000. Again, it means that distilled version of PathMNIST dataset with 900000 training images can provide the slightly better performance than the original version of PathMNIST dataset with same 90000 training images. For PathMNIST the overall improvement of the validation accuracy by DD is better than $0.6 \pm 0.3\%$ (Table 2). In comparison to the well balanced CIFAR-10/CIFAR-100 datasets this improvement is smaller and it can be explained by the slightly imbalanced distribution of images among classes in PathMNIST (see the details below).

DermaMNIST. Here the slight performance improvement was obtained with IPC growth and it was smaller in comparison with the previously published results [26,27] obtained for various other methods including ResNet-18, ResNet-50, auto-sklearn, AutoKeras, and Google AutoML Vision (Fig. 6). In comparison to the well balanced CIFAR-10/CIFAR-100 datasets and the slightly imbalanced PathMNIST even this improvement is much smaller and, again, it can be explained by the imbalanced images distribution among classes.

RetinaMNIST. This dataset is similar to DermaMNIST by its imbalanced distribution of images among classes and the slight performance improvement with IPC growth that also is smaller in comparison with the previously published results [26,27] obtained for various other methods including ResNet-18, ResNet-50, auto-sklearn, AutoKeras, and Google AutoML Vision (Fig. 7).

For the slightly (PathMNIST) and highly (DermaMNIST, RetinaMNIST) imbalanced datasets the impact of GDADD on the maximal validation accuracy values is summarized in Table 2.

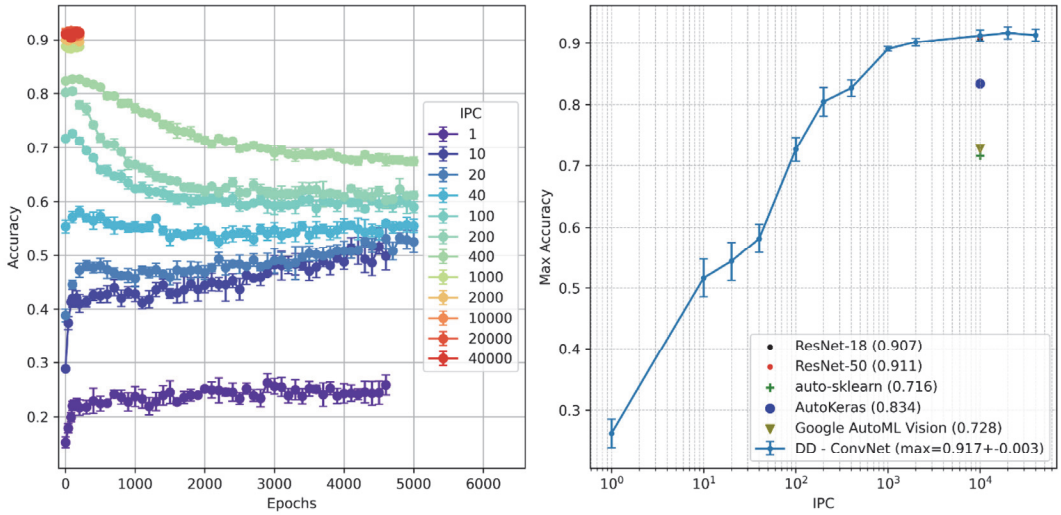


Fig. 5. PathMNIST dataset. Validation accuracy vs. epochs where the legend contains the distilled "images per class" (IPC) values (left) and maximal validation accuracy (observed for the stated IPC) vs. IPC (right) where the legend contains the reference values from [26,27]. The error bars denote the standard deviations of accuracy after 5 trials.

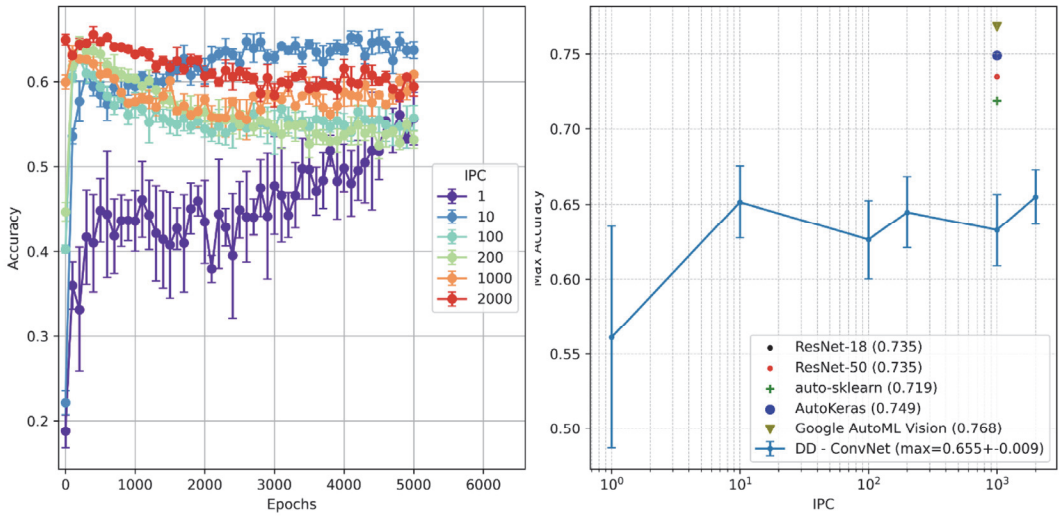


Fig. 6. DermaMNIST dataset. Validation accuracy vs. epochs where the legend contains the distilled "images per class" (IPC) values (left) and maximal validation accuracy (observed for the stated IPC) vs. IPC (right) where the legend contains the reference values from [26,27]. The error bars denote the standard deviations of accuracy after 5 trials.

5 Discussion

The analysis was performed to calculate the overall maximal validation accuracy values determined over all IPC values for all datasets (synthetic and original)

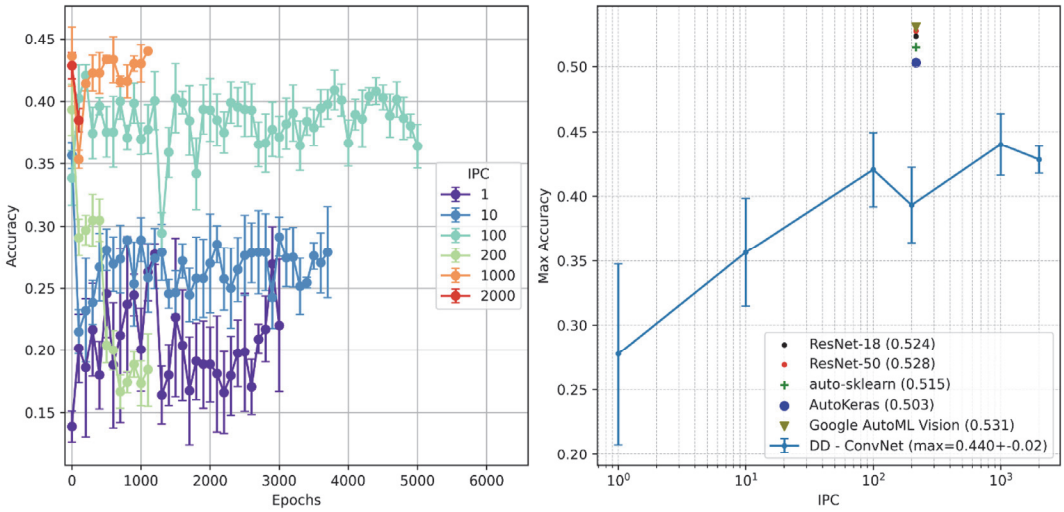


Fig. 7. RetinaMNIST dataset. Validation accuracy vs. epochs where the legend contains the distilled "images per class" (IPC) values (left) and maximal validation accuracy (observed for the stated IPC) vs. IPC (right) where the legend contains the reference values from [26,27]. The error bars denote the standard deviations of accuracy after 5 trials.

Table 2. The maximal validation accuracy values for the slightly (PathMNIST) and highly (DermaMNIST, RetinaMNIST) imbalanced datasets.

Model	Balanced	Imbalanced	
	PathMNIST	DermaMNIST	RetinaMNIST
ResNet-18 (28)	0.907	0.735	0.524
ResNet-50 (28)	0.911	0.735	0.528
auto-sklearn	0.716	0.719	0.515
AutoKeras	0.834	0.749	0.503
Google AutoML Vision	0.728	**0.768**	**0.531**
Ours (GDADD) (Fig. 5, 6 and 7)	**0.917** $\pm$ 0.003	0.655 $\pm$ 0.009	0.440 $\pm$ 0.020

and plot them versus the IPC values for which this overall maximal validation accuracy values were observed (Fig. 8).

It should be noted that the overall maximal validation accuracy values for synthetic versions of the datasets (the open symbols in Fig. 8) are aligned along a line (the dotted line in Fig. 8). But the overall maximal validation accuracy values for original versions of the datasets (the filled symbols in Fig. 8) are not aligned along a line and follow the broken line (the dashed line in Fig. 8). The error bars (red - for DermaMNIST and green - for RetinaMNIST) denote the standard deviations of IPC distribution for the original datasets after 5 trials. They characterize the imbalance of IPC distribution, the point's belonging to the trend line within the limits of standard deviation, and inability of the current GDADD setup to take it into account to improve the accuracy.

Actually, by this plot GDADD allow to obtain visual comprehension about advantages of the current GDADD setup to improve the accuracy for the well balanced datasets and disadvantages when the imbalance of IPC distribution is not considered (in the current setup).

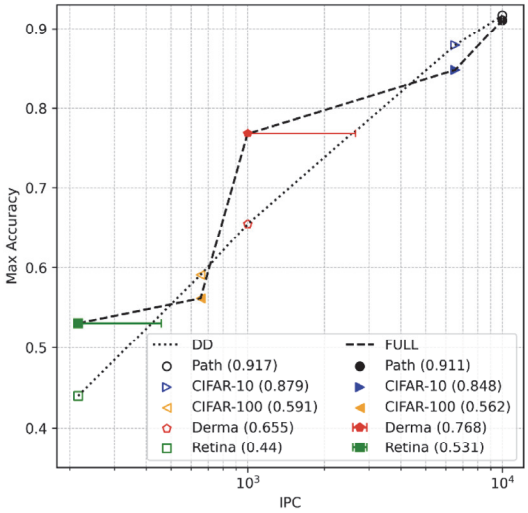


Fig. 8. Comparison of the overall maximal validation accuracy (determined over all IPC values) vs. the IPC for which this overall maximal validation accuracy was observed (right plot) where the legend contains the reference values from [2] and [26,27]. The error bars denote the standard deviations of IPC distribution for the original datasets after 5 trials. (Color figure online)

In general, GDADD presents an innovative method for augmenting datasets by leveraging expert trajectories obtained from training on real data. This approach allows for trajectory-guided augmentation where the expert trajectory guides the synthetic dataset's evolution, ensuring it captures key data characteristics. Even with a smaller number of distilled IPC, GDADD successfully encodes essential features, making it an efficient data augmentation method with the better performance. While effective for standard datasets like CIFAR-10 and CIFAR-100, performance on medical datasets like MedMNIST suggests the need for domain-specific enhancements. The potential of GDADD in healthcare applications, particularly in federated learning, is significant because synthetic datasets can replace real patient data, mitigating privacy concerns while enabling collaborative learning across institutions. Moreover, while medical imaging datasets often suffer from class imbalances and scarcity, GDADD can generate diverse synthetic samples, improving model robustness. By aligning the distilled datasets with training trajectories from multiple institutions, models can generalize better without direct access to real patient data.

GDADD offers several benefits over traditional dataset distillation and data augmentation techniques including reduce of computational overhead by utilizing previously trained models to guide distillation, enabling rapid experimen-

tation; strong generalization with fewer samples, significantly reducing storage and computational requirements; preserved model performance by consequence that models trained on synthetic data closely follow performance trends of those trained on real data. The improved diversity with increased IPC values lead to more detailed synthetic datasets that better approximate real-world variability.

Despite its advantages, GDADD has some limitations related with dependency on initial expert trajectories, high computational cost of initial expert training, potential for bias propagation. The matter is the quality of distilled datasets is highly dependent on the effectiveness of the expert trajectory. Poorly trained expert models could lead to suboptimal synthetic data. Although distillation is efficient, generating expert trajectories requires significant computational resources upfront. Any biases present in the original dataset or training trajectory can be transferred to the distilled dataset, impacting generalization.

Beyond the stated limitations, our findings suggest that the DDMTT dependency on expert trajectories makes it inherently susceptible to the quality and bias of the initial training process. The poorer performance on imbalanced datasets indicates that the trajectory matching objective can prioritize majority classes. Furthermore, the significant computational cost of generating expert models presents a practical barrier to cheap entry for various applications. Future work will need to explore more efficient trajectory sampling methods and investigate regularization techniques to prevent bias amplification during distillation, moving beyond the current identification of limitations to their active mitigation.

To bridge the gap toward clinical applicability, future work must involve close collaboration with medical experts to validate the diagnostic relevance of synthesized images and to integrate domain-specific constraints directly into the distillation loss function. Furthermore, the varied performance on imbalanced medical datasets underscores the need for the specific domain adaptation. A key next step will be to tailor the GDADD framework specifically for healthcare, perhaps by incorporating class-balanced weighting or leveraging trajectories from models trained on rare classes to ensure the synthetic data effectively addresses clinical scarcity and imbalance.

6 Conclusions

The study investigates the impact of generative data augmentation by dataset distillation (GDADD) using the dataset distillation by matching training trajectories (DDMTT) technique. This method leverages expert trajectories to guide the learning process, improving the efficiency and quality of synthetic datasets for training neural networks. Results on the well balanced CIFAR-10 and CIFAR-100 datasets demonstrate that distilled images successfully capture class-specific information, even at a low image-per-class (IPC) setting. When IPC is increased, synthetic images better approximate real-world diversity, enhancing model performance. Specifically, for CIFAR-10, increasing IPC to 4000 improves validation accuracy significantly by $3.1 \pm 0.1\%$, for CIFAR-10, increasing IPC to 400 improves validation accuracy by $2.9 \pm 0.3\%$, and for PathMNIST the overall

improvement of the validation accuracy by DD is better than $0.6 \pm 0.3\%$. For MedMNIST datasets, slightly imbalanced PathMNIST and strongly imbalanced DermaMNIST and RetinaMNIST, the synthesized images are less distinguishable compared to CIFAR datasets due to the IPC imbalances. However, increasing IPC still enhances performance, albeit with diminishing returns beyond a certain IPC threshold. Overall, GDADD enables models trained on synthetic datasets to achieve accuracy levels close and higher to those trained on the real data with a potetnial to generate compact yet informative datasets, reducing data storage needs while maintaining model generalization.

Acknowledgments. This research was supported by the National Research Fund of Ukraine under grant 2025.06/0100 in the part of the development of exploratory research on the new robust machine learning approaches for object detection and classification and by the Ministry of Education and Sciences of Ukraine (MESU), grant 2715r, as part of the search for advanced deep learning methods based on the requirements posed by devices on Edge Computing and Edge Intelligence layers.

References

1. Bengio, Y., Courville, A., Vincent, P.: Representation learning: a review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 1798–1828 (2013)
2. Cazenavette, G., Wang, T., Torralba, A., Efros, A.A., Zhu, J.Y.: Dataset distillation by matching training trajectories. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4750–4759 (2022)
3. Cazenavette, G., Wang, T., Torralba, A., Efros, A.A., Zhu, J.Y.: Wearable imagenet: synthesizing tileable textures via dataset distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2278–2282 (2022)
4. Chen, Y., Wu, Z., Shen, Z., Jia, J.: Learning from designers: fashion compatibility analysis via dataset distillation. In: 2022 IEEE International Conference on Image Processing (ICIP), pp. 856–860. IEEE (2022)
5. Fukushima, K.: Neural network model for a mechanism of pattern recognition unaffected by shift in position-neocognitron. IEICE Tech. Rep. A **62**(10), 658–665 (1979)
6. Goetz, J., Tewari, A.: Federated learning via synthetic data. arXiv preprint arXiv:2008.04489 (2020)
7. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
8. Ivakhnenko, A., Lapa, V.: Cybernetic predicting devices (1966). https://apps.dtic.mil/sti/citations/AD0654237. Accessed 24 Oct 2022
9. Kelley, H.J.: Gradient theory of optimal flight paths. ARS J. **30**(10), 947–954 (1960)
10. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
11. Krizhevsky, A., Nair, V., Hinton, G.: The CIFAR datasets (2009). https://www.cs.toronto.edu/~kriz/cifar.html. Accessed 10 Oct 2024
12. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)

13. Lei, S., Tao, D.: A comprehensive survey of dataset distillation. IEEE Trans. Pattern Anal. Mach. Intell. **46**(1), 17–32 (2023)
14. Li, G., Togo, R., Ogawa, T., Haseyama, M.: Dataset distillation for medical dataset sharing. arXiv preprint arXiv:2209.14603 (2022)
15. Li, G., Qian, G., Delgadillo, I.C., Muller, M., Thabet, A., Ghanem, B.: SGAS: sequential greedy architecture search. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1620–1630 (2020)
16. Linnainmaa, S.: Taylor expansion of the accumulated rounding error. BIT Numer. Math. **16**(2), 146–160 (1976)
17. Loo, N., Hasani, R., Amini, A., Rus, D.: Efficient dataset distillation using random feature approximation. Adv. Neural. Inf. Process. Syst. **35**, 13877–13891 (2022)
18. Rosasco, A., Carta, A., Cossu, A., Lomonaco, V., Bacciu, D.: Distilled replay: overcoming forgetting through synthetic samples. In: International Workshop on Continual Semi-Supervised Learning, pp. 104–117. Springer (2021)
19. Schmidhuber, J.: Deep learning in neural networks: an overview. Neural Netw. **61**, 85–117 (2015)
20. Schmidhuber, J., Blog, A.: The 2010s: our decade of deep learning/outlook on the 2020s. The recent decade's most important developments and industrial applications based on our AI, with an outlook on the 2020s, also addressing privacy and data markets (2020)
21. Wang, H.P., Chen, D., Kerkouche, R., Fritz, M.: Fed-GLOSS-DP: federated, global learning using synthetic sets with record level differential privacy. CoRR (2023)
22. Wang, T., Zhu, J.Y., Torralba, A., Efros, A.A.: Dataset distillation. arXiv preprint arXiv:1811.10959 (2018)
23. White, C., Jain, P., Nayak, S., Ramakrishnan, G., et al.: Speeding up nas with adaptive subset selection. arXiv preprint arXiv:2211.01454 (2022)
24. Wiewel, F., Yang, B.: Condensed composite memory continual learning. In: 2021 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2021)
25. Williams, R.: Complexity of exact gradient computation algorithms for recurrent neural networks (technical report nu-ccs-89-27). Northeastern University, College of Computer Science, Boston (1989)
26. Yang, J., Shi, R., Ni, B.: Medmnist classification decathlon: a lightweight automl benchmark for medical image analysis. In: IEEE 18th International Symposium on Biomedical Imaging (ISBI), pp. 191–195 (2021)
27. Yang, J., et al.: Medmnist v2: a large-scale lightweight benchmark for 2D and 3D biomedical image classification. arXiv preprint arXiv:2110.14795 (2021)
28. Zhang, D., Yin, J., Zhu, X., Zhang, C.: Network representation learning: a survey. IEEE Trans. Big Data **6**(1), 3 28 (2018)
29. Zhao, B., Bilen, H.: Dataset condensation with differentiable siamese augmentation. In: International Conference on Machine Learning, pp. 12674–12685. PMLR (2021)
30. Zhao, B., Bilen, H.: Dataset condensation with distribution matching. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 6514–6523 (2023)
31. Zhao, B., Mopuri, K.R., Bilen, H.: Dataset condensation with gradient matching. arXiv preprint arXiv:2006.05929 (2020)

5. Podsumowanie

Celem rozprawy „Formalne i algorytmiczne metody przetwarzania informacji nieprecyzyjnej w systemach informatycznych” było opracowanie oraz weryfikacja rozwiązań umożliwiających przejście od informacji nieprecyzyjnej, rozumianej szeroko jako nieostrość, niepełność, zaszumienie lub kontekstowość, do postaci formalnie określonej i algorytmicznie przetwarzalnej w realnych środowiskach obliczeniowych. Przedstawiony cykl publikacji potwierdza, że dla różnych postaci nieprecyzyjności możliwe jest budowanie spójnych rozwiązań spełniających trzy warunki: (I) jawna formalizacja sposobu interpretacji informacji rozumiana jako określenie modelu i semantyki przetwarzania albo jednoznacznej funkcji celu, (II) algorytmiczna realizacja formalizacji w postaci procedury możliwej do uruchomienia w systemie informatycznym (np. transformacja, wnioskowanie, rekonstrukcja lub uczenie) oraz (III) empiryczna weryfikacja jakości i kosztu przetwarzania w warunkach systemowych i/lub eksperymentalnych.

Weryfikację tezy sformułowanej w podrozdziale 2.2 przeprowadzono w czterech obszarach badawczych I–IV. W obszarze I przedstawiono wykonalność zapytań rozmytych w relacyjnych bazach danych poprzez formalizację semantyki stopniowej oraz transformację do standardowego SQL. W obszarze II ustalono, że niepełna specyfikacja przebiegu procesu przetwarzania może zostać uzupełniona przez formalny opis usług, wnioskowanie i rekonstrukcję planu wykonania. W obszarze III zaprezentowano możliwość redukcji nieprecyzyjności wymagań projektowych przez formalizację opisu komponentów i reguł ich łączenia oraz dedukcyjną syntezę struktury aplikacji. W obszarze IV potwierdzono, że ten sam schemat badawczy: formalizacja → algorytm → weryfikacja pozostaje spójny również dla innych nośników nieprecyzyjności, w tym sygnałów zewnętrznych, braków struktury w tekście oraz formalizacji optymalizacyjnej prowadzącej do kompresji informacji w danych syntetycznych. W każdym z obszarów ocena obejmowała weryfikację jakości oraz, tam gdzie było to ważne, analizę kosztu obliczeniowego z użyciem miar adekwatnych do danej domeny.

Wyniki niniejszej rozprawy potwierdzają postawioną tezę, zgodnie z którą formalne modele oraz algorytmiczne metody przetwarzania informacji nieprecyzyjnej umożliwiają skuteczne i obliczeniowo efektywne wykorzystanie informacji niejednoznacznych, niepełnych, zaszumionych i kontekstowych w systemach informatycznych, pod warunkiem jawnego zdefiniowania semantyki przetwarzania lub funkcji celu oraz empirycznej weryfikacji jakości i kosztu.

5.1. Oryginalne elementy rozprawy

Oryginalne elementy rozprawy mają charakter przekrojowy i wynikają z konsekwentnego domknięcia ścieżki: formalizacja → algorytm → weryfikacja w czterech obszarach badawczych. Szczegółowe wyniki i eksperymenty przedstawiono w Rozdziale 3, natomiast poniżej wskazano elementy stanowiące wkład autora rozprawy.

Obszar I: Nieprecyzyjność w danych i w zapytaniach relacyjnych – zapytania rozmyte oraz ich wykonanie i transformacja do SQL

Oryginalność wkładu autora polega na dostarczeniu kompletnej i odtwarzalnej metodyki przejścia od formalizmu do wykonania w standardowym SQL, bez modyfikacji rdzenia DBMS. Wkład obejmuje:

1) formalizację semantyki stopniowej zapytań rozmytych oraz miary dopasowania krotek, a następnie metodykę transformacji zapytań do postaci wykonywalnej w SQL, z jawnym wyznaczaniem stopnia spełnienia i możliwością filtrowania oraz porządkowania wyników,
2) algorytmiczne ujęcie analizy warunków zapytania (w tym klauzuli WHERE) poprzez konstrukcję drzewa wyrażeń oraz rekurencyjne wyznaczanie stopni spełnienia dla warunków złożonych zgodnie z semantyką operatorów łączenia ($\mathrm{T-norma/S-norma}$),
3) jednoznaczne ujęcie i realizację agregatów rozmytych w SQL, w tym FUZZY_COUNT oraz FUZZY_AVG,
4) rozdzielenie dwóch wariantów realizacyjnych (po stronie DBMS i po stronie aplikacji), co porządkuje konsekwencje systemowe i zwiększa przenośność rozwiązania,
5) weryfikację wykonalności i narzutu obliczeniowego na zestawie reprezentatywnych klas zapytań, wraz z opracowaniem i interpretacją wyników,
6) pozycjonowanie wyników obszaru I na tle literatury w postaci zestawień porównawczych.

Obszar II: Nieprecyzyjność na poziomie celu użytkownika – integracja heterogenicznych usług

Oryginalność wkładu autora polega na sprowadzeniu niepełności specyfikacji przebiegu procesu przetwarzania (przepływu pracy) do formalnego opisu i algorytmicznie wyznaczalnego planu wykonania, a następnie na osadzeniu tego planu w warstwie wykonawczej. Wkład obejmuje:

1) formalizację celu i usług przez warunki wstępne i końcowe oraz aksjomaty w bazie wiedzy, stanowiące podstawę dowodzenia wykonalności celu,

2) ujęcie kompozycji usług jako konstrukcji dowodu oraz mechanizm wnioskowania, w którym dowód stanowi formalny nośnik kompozycji,
3) mechanizm rekonstrukcji struktury rozwiązania prowadzący od dowodu do planu wykonania w postaci drzewa wykonania, z uwzględnieniem rozgałęzień i fragmentów możliwych do uruchomienia równolegle,
4) kontrolę złożoności wyszukiwania przez abstrakcję typów oraz selekcję dowodów,
5) domknięcie ścieżki cel użytkownika → wnioskowanie → plan wykonania → wykonanie w architekturze integracyjnej (WDS/WDC), obejmującej warstwę usług oraz warstwę orkiestracji agentowej.

Obszar III: Nieprecyzyjność specyfikacji na etapie projektowania systemów

Oryginalność wkładu autora polega na wskazaniu, że niepełne i ewoluujące wymagania można ująć formalnie w sytuacji, gdy nie wyznaczają one jednoznacznie struktury aplikacji. Wkład obejmuje:

1) formalne ujęcie specyfikacji projektowej, które pozwala przełożyć niepełne lub niejednoznaczne wymagania na przesłanki do wnioskowania prowadzącego do spójnej struktury rozwiązania, zamiast pozostawiać ten etap wyłącznie decyzjom projektowym,
2) semantyczną bazę wiedzy komponentów jako mechanizm operacyjny łączący wymagania z doborem komponentów i strukturą rozwiązania,
3) wnioskowanie prowadzące do syntezy struktury aplikacji jako konstrukcji (kompozycji modułów, klas i metod), a nie jedynie rozstrzygnięcia binarnego, z kontrolą przeszukiwania przez wstępny dobór elementów powiązanych wejściami/wyjściami i późniejszą weryfikację warunków wstępnych i końcowych,
4) domknięcie procesu generowania aplikacji: od modelu danych do szkieletu funkcjonalnego (CRUD) i interfejsu REST API, a następnie integracji szablonów na różnych poziomach architektury (od metod do komponentów złożonych z klas) z możliwością tworzenia oraz ponownego użycia szablonów użytkownika.

Obszar IV: Rozszerzenia – inne nośniki nieprecyzyjności

Oryginalność wkładu autora polega na wykazaniu, że schemat: formalizacja → algorytm → weryfikacja może być zastosowany także wtedy, gdy nieprecyzyjność nie dotyczy ani warunków zapytania, ani specyfikacji przepływu pracy, lecz ma inny nośnik. Wkład obejmuje:

1) sygnał zewnętrzny [7]: formalizację wpływu postu jako współczynnika korekty prognozy oraz zaprojektowanie porównania metod predykcyjnych (z użyciem ATAPSN oraz metod

klasycznych: ARIMA i wygładzania wykładniczego) wraz z protokołem weryfikacji statystycznej obejmującym miary błędu prognozy, korelacje Spearmana i Pearsona oraz test $t-$ Studenta,

2) braki struktury w tekście [8]: formalizację rekonstrukcji interpunkcji i kapitalizacji jako klasyfikacji tokenów w zdefiniowanej przestrzeni klas, algorytm rekonstrukcji w postaci modelu $\mathrm{XLM}-\mathrm{RoBERTa}-\mathrm{LSTM}$ z zastosowaniem masek (koniec słowa i padding) oraz metodykę oceny odporną na niezbalansowanie (wyniki z/bez klasy dominującej, analiza w podziale na klasy), uzupełnioną o porównanie z literaturą,
3) wiedzę empiryczną w danych uczących [9]: formalizację w postaci funkcji celu dla destylacji zbioru danych oraz algorytm destylacji GDADD oparty na DDMTT wraz z weryfikacją zależności jakości od liczby IPC i stopnia niezbalansowania klas na zestawach referencyjnych (CIFAR, MedMNIST). Uwzględniono koszt obliczeniowy destylacji, zależność od jakości trajektorii uczenia oraz porównanie z wynikami uczenia na zbiorach pełnych i metodami odniesienia.

Dodatkowo wkład obszaru IV obejmuje konsekwentne stosowanie procedur oceny ograniczających ryzyko zniekształceń w interpretacji jakości (m.in. przypadkowe korelacje, dominacja klasy większościowej, wpływ niezbalansowania danych) oraz wskazanie empirycznych warunków stosowalności i ograniczeń metod.

5.2. Ograniczenia przedstawionych rozwiązań

Przedstawione rozwiązania mają charakter formalno-algorytmiczny i są osadzone w konkretnych modelach oraz środowiskach wykonawczych. Zestawione poniżej ograniczenia (w podziale na obszary badawcze I–IV) stanowią naturalną konsekwencję przyjętych założeń badawczych i są związane z oryginalnym wkładem autora. Wynikają one przede wszystkim z konieczności domknięcia ścieżki: formalizacja → algorytm → weryfikacja w warunkach rzeczywistych systemów informatycznych, gdzie dążenie do wysokiej uniwersalności i wdrażalności metod wyznacza granice ich stosowalności. W konsekwencji zastosowanie rozwiązań jest determinowane przez przyjęte założenia formalne, charakter danych wejściowych oraz uwarunkowania implementacyjne.

Obszar I: Nieprecyzyjność w danych i w zapytaniach relacyjnych – zapytania rozmyte oraz ich wykonanie i transformacja do SQL

1) Zależność od definicji terminów lingwistycznych − jakość odpowiedzi zależy od postaci i parametrów funkcji przynależności, co wymaga wiedzy dziedzinowej oraz doprecyzowania znaczeń terminów w kontekście potrzeb użytkownika.
2) Koszt obliczeniowy przetwarzania stopniowego − wyznaczanie stopni spełnienia, agregacja warunków i porządkowanie wyników zwiększają koszt wykonania względem selekcji dwuwartościowej; wydajność zależy od złożoności zapytania, rozmiaru danych oraz możliwości optymalizacji w DBMS.
3) Zakres transformacji i kompromisy implementacyjne − transformacja obejmuje szeroką klasę zapytań, ale nie pokrywa w pełni wszystkich rozszerzeń spotykanych w literaturze (np. złożonych kwantyfikatorów czy rozmytych złączeń w pełnej ogólności). Wariant realizowany po stronie DBMS umożliwia wykonywanie obliczeń bezpośrednio w warstwie bazy danych i wykorzystanie natywnych mechanizmów wykonania zapytań, jednak jest zależny od możliwości rozszerzeń konkretnego silnika. Wariant aplikacyjny zwiększa niezależność od DBMS, lecz może podnieść koszty przetwarzania i komunikacji oraz wymaga konsekwentnego utrzymania zgodności semantyki obliczeń.

Obszar II: Nieprecyzyjność na poziomie celu użytkownika – integracja heterogenicznych usług

1) Zależność od jakości bazy wiedzy − skuteczność rozwiązania zależy od kompletności i spójności opisów usług obejmujących warunki wstępne i końcowe oraz aksjomaty. Braki w opisach mogą prowadzić do braku rozwiązania lub do nadmiaru wariantów.
2) Skalowalność wnioskowania − wraz ze wzrostem liczby usług i możliwych dopasowań rośnie liczba alternatywnych ścieżek dowodzenia; nawet przy mechanizmach sterowania przeszukiwaniem istnieje ryzyko znaczącego wzrostu kosztu obliczeń w środowiskach z dużą liczbą usług.
3) Ograniczenia wdrożeniowe i zmienność środowiska − formalny opis usług porządkuje wymagania i rezultaty ich działania, ale nie eliminuje problemów współdziałania na poziomie wdrożeniowym, takich jak zgodność formatów i jakość danych, bezpieczeństwo, awarie, monitoring oraz zmienność usług w czasie (zmiany interfejsu, czasowa niedostępność), które wymagają dodatkowych mechanizmów odpornościowych.

Obszar III: Nieprecyzyjność specyfikacji na etapie projektowania systemów

1) Koszt formalizacji i utrzymania bazy wiedzy − synteza struktury aplikacji zakłada przedstawienie wymagań w postaci umożliwiającej wnioskowanie oraz utrzymanie kompletnej i aktualnej bazy wiedzy komponentów (własności i reguł łączenia). Rozszerzanie rozwiązania o nowe technologie, frameworki i wzorce wiąże się z koniecznością aktualizacji tej bazy.
2) Ograniczenie do określonej klasy aplikacji − automatyzacja jest powiązana z przyjętymi założeniami architektonicznymi i dotyczy wybranej klasy aplikacji webowych; pełna automatyzacja wytwarzania dowolnych aplikacji nie jest przedmiotem pracy.

Obszar IV: Rozszerzenia – inne nośniki nieprecyzyjności

1) Sygnały zewnętrzne są zależne od kontekstu − wpływ informacji z mediów społecznościowych jest krótkotrwały i trudny do izolacji od innych czynników; formalizacja ułatwia algorytmiczne uwzględnienie sygnału, ale nie eliminuje nakładania się wielu czynników.
2) Rekonstrukcja struktury tekstu zależy od danych wejściowych − skuteczność odtwarzania interpunkcji i kapitalizacji jest wrażliwa na styl językowy, długość sekwencji, jakość transkrypcji oraz różnice między źródłami danych; w zastosowaniach wdrożeniowych może być konieczna adaptacja modelu do danych docelowych.
3) Destylacja danych ogranicza interpretowalność − wiedza jest tu reprezentowana niejawnie w danych syntetycznych, co ogranicza możliwość wyjaśniania w kategoriach reguł; skuteczność zależy od hiperparametrów, niezbalansowania i zasobów obliczeniowych, a zastosowanie metody dla innego zbioru danych lub innej domeny zwykle wymaga ponownego przeprowadzenia destylacji.

Powyższe ograniczenia wyznaczają zakres bezpośredniej stosowalności wyników oraz wskazują konkretne kierunki dalszego rozwoju metodyki. Obejmują one adaptację semantyki terminów lingwistycznych (obszar I), budowę mechanizmów odporności i monitoringu wykonania (obszar II), utrzymanie i systematyczne rozszerzanie bazy wiedzy (obszar III) oraz analizę stabilności metod przy przenoszeniu między różnorodnymi źródłami danych (obszar IV). Dzięki takiemu ujęciu zidentyfikowane ograniczenia przestają być jedynie barierami wdrożeniowymi, a stają się punktami wyjścia do dalszych prac badawczych podkreślając otwarty i rozwojowy charakter zaproponowanej w rozprawie metodyki przetwarzania informacji nieprecyzyjnej.

5.3. Końcowe wnioski

W rozprawie przyjęto tezę, że formalne modele oraz algorytmiczne metody przetwarzania informacji nieprecyzyjnej umożliwiają skuteczną i kontrolowaną, a zarazem wykonalną obliczeniowo integrację oraz analizę danych niejednoznacznych, niepełnych, zaszumionych i kontekstowych w systemach informatycznych pod warunkiem jawnego określenia semantyki przetwarzania oraz empirycznej weryfikacji jakości i kosztu. Przedstawiony cykl publikacji potwierdza tę tezę w ujęciu przekrojowym, ponieważ w czterech obszarach badawczych wykazano możliwość przejścia od nieprecyzyjnej reprezentacji problemu do rozwiązania uruchamialnego w środowisku obliczeniowym oraz weryfikowalnego empirycznie lub systemowo.

Wynik przekrojowy można ująć jako konsekwentne domknięcie schematu: formalizacja → algorytm → weryfikacja. Formalizacja przyjmuje różne postacie w zależności od obszaru badawczego − od semantyki stopniowej i funkcji przynależności, przez formalny opis usług i warunków ich zastosowania, po semantyczny opis komponentów projektowych lub funkcję celu w ujęciu optymalizacyjnym. Niezmienny pozostaje wymóg jednoznacznej interpretacji rezultatów formalizacji w systemie informatycznym oraz ich weryfikowalności w zakresie jakości i narzutu obliczeniowego.

W każdym z obszarów formalizacja i algorytmiczna realizacja zostały uzupełnione o weryfikację jakości i kosztu w formie adekwatnej do charakteru zadania. W obszarze I oceniano wykonalność transformacji w standardowym SQL oraz narzut obliczeniowy dla klas zapytań o rosnącej złożoności i skali danych, a dodatkowo pozycjonowano zakres funkcjonalny rozwiązań w zestawieniach porównawczych na tle literatury. W obszarze II poprawność rozumowania potwierdzono przez konstrukcję dowodu i rekonstrukcję planu wykonania, a koszt wyszukiwania ograniczano mechanizmami sterowania przeszukiwaniem (m.in. kryteria selekcji wariantów dowodu oraz ograniczenia liczby analizowanych rozwiązań). W obszarze III weryfikacja miała charakter systemowy i obejmowała demonstrację domknięcia procesu od formalizacji wymagań do syntezy struktury aplikacji oraz generowania artefaktów (szkielet CRUD/REST i integracja szablonów), przy zachowaniu spójności i zgodności wynikowej struktury z wymaganiami. W obszarze IV zastosowano protokoły empiryczne właściwe dla danej domeny: testy istotności i miary błędu dla sygnałów zewnętrznych oraz metryki klasyfikacji z analizą niezbalansowania dla rekonstrukcji tekstu. W przypadku destylacji danych oceniano maksymalną dokładność walidacyjną (średnia ± odchylenie standardowe) analizowaną w funkcji IPC oraz dla zbiorów

o różnej nierównowadze klas. Dodatkowo wykonywano porównania z wartościami referencyjnymi − w [8] z wynikami raportowanymi w literaturze, a w [9] z wynikami uczenia na zbiorach pełnych oraz metodami referencyjnymi.

Realizacja celów szczegółowych C1–C5 przebiegała następująco:

- **C1–C2** zrealizowano w obszarze I przez formalizację stopnia spełnienia warunków zapytania i sprowadzenie tej semantyki do postaci wykonywalnej w standardowym SQL, z zachowaniem interpretowalności wyniku oraz kontrolą kosztu obliczeniowego.
- **C3** zrealizowano w obszarze II przez formalny opis usług i celu, konstrukcję dowodu w procesie wnioskowania oraz rekonstrukcję planu wykonania, co pozwala przejść od celu użytkownika do wykonalnego przebiegu procesu przetwarzania (przepływu pracy) w środowisku heterogenicznym.
- **C4** zrealizowano w obszarze III przez formalizację wymagań i komponentów w bazie wiedzy oraz wnioskowanie prowadzące do syntezy spójnej struktury aplikacji, co skraca przejście od specyfikacji do konstrukcji możliwej do implementacji przy zachowaniu kontroli semantycznej.
- **C5** zrealizowano w obszarze IV, wykazując, że schemat: formalizacja → algorytm → weryfikacja jest zasadny również wtedy, gdy nośnikiem nieprecyzyjności jest sygnał zewnętrzny, braki struktury w tekście lub pośrednia reprezentacja wiedzy w danych uczących, a formalizacja ma postać celu optymalizacyjnego.

W świetle uzyskanych wyników daje się sformułować wniosek ogólny, który wskazuje na to, że skuteczne przetwarzanie informacji nieprecyzyjnej wymaga konsekwentnego domknięcia trzech elementów. Zalicza się do nich: (I) jawną formalizację sposobu interpretacji informacji, (II) algorytmiczną realizację tej formalizacji w postaci procedury uruchamialnej w systemie informatycznym oraz (III) empiryczną weryfikację jakości i kosztu przetwarzania. Cykl publikacji potwierdza, że zaproponowana metodyka pozwala projektować rozwiązania formalnie spójne, uruchamialne w realnych środowiskach obliczeniowych oraz weryfikowalne empirycznie.

5.4. Kierunki dalszych badań

Wyniki rozprawy potwierdzają zasadność schematu: formalizacja → algorytm → weryfikacja, a jednocześnie wskazują kierunki, w których można wzmocnić wykonalność, interpretowalność i jakość rozwiązań. Poniższe propozycje wynikają bezpośrednio z ograniczeń omówionych w 5.2. Przy każdym z nich w nawiasie wskazano obszar badawczy (I–IV), którego dany kierunek dotyczy.

1) Rozwój metod automatycznej lub półautomatycznej adaptacji parametrów funkcji przynależności na podstawie danych, informacji zwrotnej użytkownika lub kontekstu, przy zachowaniu jednoznacznej interpretacji semantyki (obszar I).
2) Lepsza integracja obliczeń stopnia spełnienia z planowaniem i optymalizacją zapytań (selektywność, statystyki, indeksowanie), aby ograniczać narzut obliczeniowy i koszt porządkowania wyników (obszar I).
3) Rozwój mechanizmów odporności wykonania na zmienność usług, obejmujących adaptację planu w trakcie wykonania, ponowne planowanie, wybór alternatyw, obsługę awarii i monitoring (obszar II).
4) Rozszerzenie formalnego opisu o własności niefunkcjonalne (czas, koszt, bezpieczeństwo, jakość danych), tak aby konstrukcja planu uwzględniała nie tylko poprawność, ale również kryteria optymalizacyjne (obszar II).
5) Rozwój metod przejścia od wymagań opisowych do formalizacji, obejmujących pozyskiwanie i doprecyzowywanie wymagań, mapowanie do pojęć ontologicznych oraz zapewnianie spójności reprezentacji formalnej (obszar III).
6) Rozbudowa i utrzymanie bazy wiedzy komponentów poprzez rozszerzanie jej o nowe technologie i wzorce oraz mechanizmy aktualizacji i walidacji reguł łączenia (obszar III).
7) Dokładniejsze modelowanie sygnałów zewnętrznych z uwzględnieniem czasu życia sygnału, konkurencji informacyjnej i wiarygodności źródeł, a także integracja wielu kanałów danych w jednym protokole oceny (obszar IV).
8) Wzmocnienie odporności metod na własności danych poprzez stabilizację jakości dla klas rzadkich w rekonstrukcji struktury tekstu oraz wprowadzanie mechanizmów uwzględniających niezbalansowanie i stabilność w destylacji danych (np. w funkcji celu lub procedurze optymalizacji) (obszar IV).

Literatura

[1] Nowakowski G., *Fuzzy queries on relational databases*, in 2018 International Interdisciplinary PhD Workshop (IIPhDW), 2018, Institute of Electrical and Electronics Engineers, IEEE, pp. 293–299, ISBN 978-1-5386-6143-7. DOI: 10.1109/IIPHDW.2018.8388376.

[2] Nowakowski G., *Methodology of Transformation of Fuzzy Queries into Queries in the SQL Standard*, in IDAACS 2019: proceedings of the 2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), 2019, vol. 2, Institute of Electrical and Electronics Engineers, IEEE, pp. 674–679, ISBN 978-1-7281-4069-8. DOI: 10.1109/IDAACS.2019.8924312.

[3] Telenyk S., Nowakowski G., Yefremov K. and Khmeliuk V., *Logics based application integration for interdisciplinary scientific investigations*, in IDAACS 2017: proceedings of the 2017 IEEE 9th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), 2017, vol. 2, Institute of Electrical and Electronics Engineers, IEEE, pp. 1026–1031, ISBN 978-1-5386-0697-1. DOI: 10.1109/IDAACS.2017.8095241.

[4] Nowakowski G., Telenyk S., Yefremov K. and Khmeliuk V., *The Approach to Applications Integration for World Data Center Interdisciplinary Scientific Investigations*, in Proceedings of the 2019 Federated Conference on Computer Science and Information Systems (FedCSIS), Annals of Computer Science and Information Systems, 2019, vol. 18, New York, Institute of Electrical and Electronics Engineers, pp. 539–545, ISBN 978-83-955416-0-5. DOI: 10.15439/2019F71

[5] Nowakowski G., Telenyk S., Yefremov K. and Khmeliuk V., *Simple and Flexible Way to Integrate Heterogeneous Information Systems and Their Services into the World Data System*, Journal of Automation, Mobile Robotics and Intelligent Systems, 2021, vol. 15, no. 4, pp. 76–90. DOI: 10.14313/JAMRIS/4-2021/29.

[6] Telenyk S., Nowakowski G., Zharikov E. and Vovk J., *Conceptual Foundations of the Use of Formal Models and Methods for the Rapid Creation of Web Applications*, in IDAACS 2019: proceedings of the 2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), 2019, vol. 1, Institute of Electrical and Electronics Engineers, IEEE, pp. 512–518. DOI: 10.1109/IDAACS.2019.8924416.

[7] Telenyk S., Nowakowski G., Gavrilenko O., Miahkyi M. and Khalus O., *An analysis of the influence of famous people's posts on social networks on the cryptocurrency exchange rate*, Bulletin of the Polish Academy of Sciences: Technical Sciences, 2024, vol. 72, no. 4. DOI: 10.24425/bpasts.2024.150117.

[8] Shymkovych V., Nowakowski G. and Telenyk S., *Joint Punctuation Restoration and Text* Capitalisation *with a Hybrid XLM-RoBERTa–LSTM Model*, in IDAACS 2025: proceedings of the 13th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), 2025, vol. 1, Institute of Electrical and Electronics Engineers, IEEE, pp. 990–997. DOI: 10.1109/IDAACS68557.2025.11322226.

[9] Gordienko, Y., Nowakowski, G., Kochura, Y., Taran, V., Stirenko, S., *Generative Data Augmentation by Dataset Distillation*, in Artificial Intelligence for Biomedical Data: First International Workshop, AIBio 2025, Held in Conjunction with the European Conference on Artificial Intelligence, ECAI 2025, Communications in Computer and Information Science, 2026, vol. 2696, Cham, Springer, pp. 105-118, ISBN 978-3-032-17215-0. DOI: 10.1007/978-3-032-17216-7_9.

[10] Codd E. F., *A relational model of data for large shared data banks*, Communications of the ACM, vol. 13, no. 6, Jun. 1970, pp. 377–387. DOI: 10.1145/362384.362685.

[11] Codd E. F., *Relational Completeness of Data Base Sublanguages*, Research Report / RJ / IBM / San Jose, California, vol. RJ987, 1972.

[12] Aho A. V, Sagiv Y. and Ullman J. D., *Equivalences among Relational Expressions*, SIAM Journal on Computing, vol. 8, 1979, pp. 218–246. DOI: 10.1137/0208017.

[13] Ullman J. D. and Widom J., *A first course in database systems*. USA: Prentice-Hall, Inc., 1997.

[14] DeBarros A., *SQL w praktyce. Jak dzięki danym uzyskiwać cenne informacje*, 2nded. Gliwice: Helion, 2024.

[15] Information *Technology - Database Languages SQL - Part 1: Framework (SQL/Framework)*, Available: https://www.iso.org/standard/76583.html [Accessed: 5 January 2026].

[16] Dubois D. and Prade H., *Fuzzy cardinality and the modeling of imprecise quantification*, Fuzzy Sets and Systems, vol. 16, no. 3, 1985, pp. 199–230. DOI: 10.1016/0165-0114(85)90025-9.

[17] Ichikawa T. and Hirakawa M., *ARES: a relational database with the capability of performing flexible interpretation of queries*, IEEE Transactions on Software Engineering, vol. 12, no. 5, May 1986, pp. 624–634. DOI: 10.1109/TSE.1986.6312958.

[18] Cox E., *Relational database queries using fuzzy logic*, AI Expert: The Magazine of Artificial Intelligence in Practice, vol. 10, no. 1, Jan. 1995, pp. 22–29.

[19] George R., Petry F. E., Buckles B. P. and Srikanth R., *Fuzzy database systems - challenges and opportunities of a new era*, International Journal of Intelligent Systems, vol. 11, no. 9, 1996, pp. 649–659. DOI: 10.1002/(SICI)1098-111X(199609)11:9<649::AID-INT4>3.0.CO;2-K.

[20] Dubois Didier and Prade H., *Using Fuzzy Sets in Flexible Querying: Why and How?*, in Flexible Query Answering Systems, Andreasen Troels and Christiansen H. and L. H. L. Eds. Boston, MA: Springer US, 1997, pp. 45–60. DOI: 10.1007/978-1-4615-6075-3_3.

[21] Bosc Patrick and Pivert O., *SQLf Query Functionality on Top of a Regular Relational Database Management System*, in Knowledge Management in Fuzzy Databases, Pons Olga and Vila M. A. and K. J. Eds. Heidelberg: Physica-Verlag HD, 2000, pp. 171–190. DOI: 10.1007/978-3-7908-1865-9_11.

[22] Bosc P. and Pivert O., *Fuzzy querying in conventional databases*, in Fuzzy Logic for the Management of Uncertainty, USA: John Wiley & Sons, Inc., 1992, pp. 645–671.

[23] Bosc P., Galibourg M. and Hamon G., *Fuzzy querying with SQL: Extensions and implementation aspects*, in Readings in Fuzzy Sets for Intelligent Systems, Dubois D. et al. Eds. Morgan Kaufmann, 1993, pp. 686–694. DOI: 10.1016/B978-1-4832-1450-4.50074-2.

[24] Loo Grace Saulan and Lee K.-H., *An Interface to Databases for Flexible Query Answering: A Fuzzy-Set Approach*, in Database and Expert Systems Applications, 2000, pp. 654–663. DOI: 10.1007/3-540-44469-6_61.

[25] Nowakowski G., *Conversion of fuzzy queries into standard SQL queries using Oracle 11G XE*, Czasopismo Techniczne, vol. 2016, no. Elektrotechnika Zeszyt 3-E 2016, 2016, pp. 198–213.

[26] Zadeh L. A., *Fuzzy sets*, *Information and Control*, vol. 8, no. 3, 1965, pp. 338–353. DOI: 10.1016/S0019-9958(65)90241-X.

[27] Yager R. R., *On ordered weighted averaging aggregation operators in multicriteria decisionmaking*, IEEE Transactions on Systems, Man, and Cybernetics, vol. 18, no. 1, 1988, pp. 183–190. DOI: 10.1109/21.87068.

[28] Yager R. R., *Connectives and quantifiers in fuzzy sets*, Fuzzy Sets and Systems, vol. 40, no. 1, 1991, pp. 39–75. DOI: 10.1016/0165-0114(91)90046-S.

[29] Bouchon-Meunier B. and Jia Y., *Linguistic modifiers and imprecise categories*, International Journal of Intelligent Systems, vol. 7, no. 1, 1992, pp. 25–36. DOI: 10.1002/int.4550070105.

[30] Zadrożny S., *Zapytania nieprecyzyjne i lingwistyczne podsumowania baz danych*. Akademicka Oficyna Wydawnicza EXIT, 2006.

[31] Bosc P. and Pivert O., *Some approaches for relational databases flexible querying*, Journal of Intelligent Information Systems, vol. 1, no. 3, 1992, pp. 323–354. DOI: 10.1007/BF00962923.

[32] Buckles B. P. and Petry F. E., *A fuzzy representation of data for relational databases*, Fuzzy Sets and Systems, vol. 7, no. 3, 1982, pp. 213–226. DOI: 10.1016/0165-0114(82)90052-5.

[33] Fagin R., *Fuzzy queries in multimedia database systems*, in Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, 1998, pp. 1–10. DOI: 10.1145/275487.275488.

[34] Fukami S., Umano M., Muzimoto M. and Tanaka H., *Fuzzy database retrieval and manipulation language*, IEICE Technical Reports, AL-78-85 (Automata and Language), vol. 78, no. 233, 1979, pp. 65–72.

[35] Kacprzyk J. and Zadrozny S., *Fuzzy querying for Microsoft Access*, in Proceedings of 1994 IEEE 3rd International Fuzzy Systems Conference, 1994, pp. 167–171 vol.1. DOI: 10.1109/FUZZY.1994.343698.

[36] Mansfield W. H. and Fleischman R. M., *A high-performance, ad hoc, fuzzy query processing system*, Journal of Intelligent Information Systems, vol. 2, no. 4, 1993, pp. 397–419. DOI: 10.1007/BF00961661.

[37] Medina J. M., Pons O. and Vila M. A., *Gefred: A generalized model of Fuzzy Relational Databases*, Information Sciences, vol. 76, no. 1, 1994, pp. 87–109. DOI: 10.1016/0020-0255(94)90069-8.

[38] Nakajima H., Sogoh T. and Arao M., *Fuzzy database language and library-fuzzy extension to SQL*, in [Proceedings 1993] Second IEEE International Conference on Fuzzy Systems, 1993, pp. 477–482 vol.1. DOI: 10.1109/FUZZY.1993.327514.

[39] Umano M., *Freedom-0: A fuzzy database system*, in Readings in Fuzzy Sets for Intelligent Systems, Dubois D. et al. Eds. Morgan Kaufmann, 1993, pp. 667–675. DOI: 10.1016/B978-1-4832-1450-4.50072-9.

[40] Yager R. R., *Database discovery using fuzzy sets*, International Journal of Intelligent Systems, vol. 11, no. 9, 1996, pp. 691–712. DOI: 10.1002/(SICI)1098-111X(199609)11:9<691::AID-INT7>3.0.CO;2-F.

[41] Bosc P. and Pivert O., *SQLf: a relational database language for fuzzy querying*, IEEE Transactions on Fuzzy Systems, vol. 3, no. 1, 1995, pp. 1–17. DOI: 10.1109/91.366566.

[42] Ma Z. and Yan L., *Data modeling and querying with fuzzy sets: A systematic survey*, Fuzzy Sets and Systems, vol. 445, 2022, pp. 147–183. DOI: 10.1016/j.fss.2022.01.006.

[43] Mama R. and Machkour M., *Fuzzy querying with SQL: Fuzzy view-based approach*, Journal of Intelligent & Fuzzy Systems, vol. 40, no. 5, 2021, pp. 9937–9948. DOI: 10.3233/JIFS-202551.

[44] Andreasen T., Bordogna G., Tré G. De, Kacprzyk J., Larsen H. L. and Zadrożny S., *The power and potentials of Flexible Query Answering Systems: A critical and comprehensive analysis*, Data & Knowledge Engineering, vol. 149, 2024, p. 102246. DOI: 10.1016/j.datak.2023.102246.

[45] Fosci P. and Psaila G., *Evolving J-CO-QL+ with fuzzy evaluators for flexible querying of JSON data sets*, Neurocomputing, vol. 633, 2025, p. 129621. DOI: 10.1016/j.neucom.2025.129621.

[46] Levy A. Y., *Logic-Based Techniques in Data Integration*, in Logic-Based Artificial Intelligence, Minker J. Eds. Boston, MA: Springer US, 2000, pp. 575–595. DOI: 10.1007/978-1-4615-1567-8_24.

[47] Kavis M. J., *Architecting the Cloud: Design Decisions for Cloud Computing Service Models (SaaS, PaaS, and IaaS)*. John Wiley & Sons, Inc., 2014. DOI: 10.1002/9781118691779.

[48] Shakhovska N., Veres O., Bolubash Y. and Bychkovska-Lipinska L., *Data space architecture for Big Data managering*, in Xth International Scientific and Technical Conference „Computer Sciences and Information Technologies", CSIT 2015, Lviv, Ukraine, September 14-17, 2015, 2015, pp. 184–187. DOI: 10.1109/STC-CSIT.2015.7325461.

[49] Guler A. T., Waaijer C. J. F., Mohammed Y. and Palmblad M., *Automating bibliometric analyses using Taverna scientific workflows: A tutorial on integrating Web Services*, Journal of Informetrics, vol. 10, no. 3, 2016, pp. 830–841. DOI: 10.1016/j.joi.2016.05.002.

[50] Lamprecht A. L., Palmblad M., Ison J., Schwammle V., Al Manir M. S., Altintas I., Baker C. J. O., Ben Hadj Amor A., Capella-Gutierrez S., Charonyktakis P. et al., *Perspectives on automated composition of workflows in the life sciences*, F1000Research, vol. 10, no. 897, 2021. DOI: 10.12688/f1000research.54159.1.

[51] Schultes E., *The FAIR hourglass: A framework for FAIR implementation*, FAIR Connect, vol. 1, no. 1, 2023, pp. 13–17. DOI: 10.3233/FC-221514.

[52] Welter D., Juty N., Rocca-Serra P., Xu F., Henderson D., Gu W., Strubel J., Giessmann R. T., Emam I., Gadiya Y. et al., *FAIR in action - a flexible framework to guide FAIRification*, Scientific Data, vol. 10, no. 1, 2023, p. 291. DOI: 10.1038/s41597-023-02167-2.

[53] Wilkinson S. R., Gustafsson J., Bacall F., Belhajjame K., Capella S., Gonzalez J. M. F., Tande J. F., Gadelha L., Garijo D., Grubel P. et al., *An Ecosystem of Services for FAIR Computational Workflows*. 2025. DOI: 10.48550/arXiv.2505.15988.

[54] Gustafsson O. J. R., Wilkinson S. R., Bacall F., Soiland-Reyes S., Leo S., Pireddu L., Owen S., Juty N., Fernández J. M., Brown T. et al., *WorkflowHub: a registry for computational workflows*, Scientific Data, vol. 12, no. 1, 2025, p. 837. DOI: 10.1038/s41597-025-04786-3.

[55] Visser C., Johansson L. F., Kulkarni P., Mei H., Neerincx P., Joeri van der Velde K., Horvatovich P., Gool A. J., Swertz M. A., Hoen P. A. C. et al., *Ten quick tips for building FAIR workflows*, PLOS Computational Biology, vol. 19, no. 9, 2023. DOI: 10.1371/journal.pcbi.1011369.

[56] Shin W., Souza R., Rosendo D., Suter F., Wang F., Balaprakash P. and da Silva R., *The (R)evolution of Scientific Workflows in the Agentic AI Era: Towards Autonomous Science*, in Proceedings of the SC '25 Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis, 2025, pp. 2305–2316. DOI: 10.1145/3731599.3767580.

[57] Balaprakash P., Raghavan K., Cappello F., Deelman E., Mandal A., Hongwei J., Mahmud I., Komal T., Shixun W., Zuk P. et al., *SWARM: Reimagining scientific workflow management systems in a distributed world*, The International Journal of High Performance Computing Applications, vol. 39, no. 5, 2025, pp. 692–712. DOI: 10.1177/10943420251339317.

[58] Silva R., Badia R. M., Bard D., Foster I. T., Jha S. and Suter F., *Frontiers in Scientific Workflows: Pervasive Integration With High-Performance Computing*, Computer, vol. 57, no. 08, 2024, pp. 36–44. DOI: 10.1109/MC.2024.3401542.

[59] Wakil K. and Jawawi D. N. A., *A New Adaptive Model for Web Engineering Methods to Develop Modern Web Applications*, in Proceedings of the 2018 International Conference on Software Engineering and Information Management, 2018, pp. 32–39. DOI: 10.1145/3178461.3178468.

[60] Panetti T. and D'Ambrogio A., *A Complexity-less Approach for Automated Development of Data-intensive Web Applications*, in 2018 International Symposium on Networks, Computers and Communications, ISNCC 2018, Rome, Italy, June 19-21, 2018, 2018, pp. 1–6. DOI: 10.1109/ISNCC.2018.8531027.

[61] Sferruzza D., Rocheteau J., Attiogbé C. and Lanoix A., *A Model-Driven Method for Fast Building Consistent Web Services in Practice*, in Proceedings of the 6th International Conference on Model-Driven Engineering and Software Development - MODELSWARD, 2018, pp. 15–24. DOI: 10.5220/0006531900150024.

[62] Falzone Emanuele and Bernaschina C., *Model Based Rapid Prototyping and Evolution of Web Application*, in Web Engineering (ICWE), 2018, pp. 496–500. DOI: 10.1007/978-3-319-91662-0_43.

[63] Lorch R., Meng B., Siu K., Moitra A., Durling M., Paul S., Varanasi S. C. and Mcmillan C., *Formal Methods in Requirements Engineering: Survey and Future Directions*, in Proceedings of the 2024 IEEE/ACM 12th International Conference on Formal Methods in Software Engineering (FormaliSE), 2024, pp. 88–99. DOI: 10.1145/3644033.3644373.

[64] Giannakopoulou D., Pressburger T., Mavridou A. and Schumann J., *Automated formalization of structured natural language requirements*, Information and Software Technology, vol. 137, 2021. DOI: 10.1016/j.infsof.2021.106590.

[65] Mokos K., Nestoridis T., Katsaros P. and Bassiliades N., *Semantic Modeling and Analysis of Natural Language System Requirements*, IEEE Access, vol. 10, 2022, pp. 84094–84119. DOI: 10.1109/ACCESS.2022.3197281.

[66] Castillo J. and Dávila A., *Tools in Model-Driven Web Development: A Systematic Mapping Study*, in Proceedings of the 22nd LACCEI International Multi-Conference for Engineering, Education and Technology, 2024. DOI: 10.18687/LACCEI2024.1.1.1336.

[67] David I., Latifaj M., Pietron J., Zhang W., Ciccozzi F., Malavolta I., Raschke A., Steghöfer J. P. and Hebig R., *Blended modeling in commercial and open-source model-driven software engineering tools: A systematic study*, Software and Systems Modeling, vol. 22, no. 1, 2023, pp. 415–447. DOI: 10.1007/s10270-022-01010-3.

[68] Shamsujjoha M., Grundy J., Li L., Khalajzadeh H. and Lu Q., *Developing Mobile Applications Via Model Driven Development: A Systematic Literature Review*, Information and Software Technology, vol. 140, 2021. DOI: 10.1016/j.infsof.2021.106693.

[69] Welter C. and Farias K., *MoT: A Model-Driven Low-Code Approach for Simplifying Cloud-of-Things Application Development*. 2025. DOI: 10.48550/arXiv.2512.14613.

[70] Kirchhof J. C., Rumpe B., Schmalzing D. and Wortmann A., *MontiThings: Model-Driven Development and Deployment of Reliable IoT Applications*, Journal of Systems and Software, vol. 183, 2022. DOI: 10.1016/j.jss.2021.111087.

[71] Corradini F., Fedeli A., Fornari F., Polini A., Re B. and Ruschioni L., *X-IoT: a model-driven approach to support IoT application portability across IoT platforms*, Computing, vol. 105, no. 9, 2023, pp. 1981–2005. DOI: 10.1007/s00607-023-01155-z.

[72] Ante L., *How Elon Musk's Twitter activity moves cryptocurrency markets*, Technological Forecasting and Social Change, vol. 186, 2023, p. 122112. DOI: 10.1016/j.techfore.2022.122112.

[73] Huynh T. L. D., *When Elon Musk Changes his Tone, Does Bitcoin Adjust Its Tune?*, Computational Economics, vol. 62, no. 2, 2023, pp. 639–661. DOI: 10.1007/s10614-021-10230-6.

[74] Pandey T. D., *Impact of Musk's remarks on volatility of Bitcoin and Dogecoin amid COVID-19 pandemic*, Journal of Digital Economy, vol. 3, 2024, pp. 85–102. DOI: 10.1016/j.jdec.2024.12.002.

[75] Subramanian H., Angle P., Rouxelin F. and Zhang Z., *A decision support system using signals from social media and news to predict cryptocurrency prices*, Decision Support Systems, vol. 178, 2024, p. 114129. DOI: 10.1016/j.dss.2023.114129.

[76] Zhou Z., Song Z., Xiao H. and Ren T., *Multi-source data driven cryptocurrency price movement prediction and portfolio optimization*, Expert Systems with Applications, vol. 219, 2023. DOI: 10.1016/j.eswa.2023.119600.

[77] Naifar N., Altamimi S., Alshahrani F. and Alhashim M., *How media coverage news and global uncertainties drive forecast of cryptocurrencies returns?*, Heliyon, vol. 9, no. 6, 2023. DOI: 10.1016/j.heliyon.2023.e16502.

[78] Gurgul V., Lessmann S. and Härdle W. K., *Deep Learning and NLP in Cryptocurrency Forecasting: Integrating Financial, Blockchain, and Social Media Data*. 2024. DOI: 10.48550/arXiv.2311.14759.

[79] Ortu M., Uras N., Conversano C., Bartolucci S. and Destefanis G., *On technical trading and social media indicators for cryptocurrency price classification through deep learning*, Expert Systems with Applications, vol. 198, 2022, p. 116804. DOI: 10.1016/j.eswa.2022.116804.

[80] M. P., Nguyen T. N., Hamdi M. and Cengiz K., *Global cryptocurrency trend prediction using social media*, *Information Processing & Management*, vol. 58(6), 2021. DOI: 10.1016/j.ipm.2021.102708

[81] Păiş V. and Tufiş D., *Capitalization and punctuation restoration: a survey*, *Artificial Intelligence Review*, vol. 55, no. 3, 2022, pp. 1681–1722. DOI: 10.1007/s10462-021-10051-x.

[82] Pan R., Garcia-Diaz J. A., Vicente P. J. V. and Valencia-Garcia R., *Evaluation of transformer-based models for punctuation and capitalization restoration in Catalan and Galician*, Procesamiento del Lenguaje Natural, vol. 70, 2023, pp. 27–38. DOI: 10.26342/2023-70-2.

[83] Yi J., Tao J., Bai Y., Tian Z. and Fan C., *Transfer knowledge for punctuation prediction via adversarial training*, Speech Communication, vol. 149, no. C, 2023, pp. 1–10. DOI: 10.1016/ j.specom.2023.03.003.

[84] Bakare A. M., Anbananthen K. S. M., Muthaiyah S., Krishnan J. and Kannan S., *Punctuation Restoration with Transformer Model on Social Media Data*, Applied Sciences, vol. 13, no. 3, 2023. DOI: 10.3390/app13031685.

[85] Lima T. B., Rolim V., Nascimento A. C. A., Miranda P., Macario V., Rodrigues L., Freitas E., Gašević D. and Mello R. F., *Towards explainable automatic punctuation restoration for Portuguese using transformers*, Expert Systems with Applications, vol. 257, 2024. DOI: 10.1016/j.eswa.2024.125097.

[86] Polacek M., Cerva P. and Zdansky J., *Lightweight online punctuation and capitalization restoration for streaming ASR systems*, Speech Communication, vol. 173, 2025, p. 103269. DOI: 10.1016/j.specom.2025.103269.

[87] Beigi H. and Liu X. Y., *Efficient Ensemble of Deep Neural Networks for Multimodal Punctuation Restoration and the Spontaneous Informal Speech Dataset*, Electronics, vol. 14, no. 5, 2025. DOI: 10.3390/electronics14050973.

[88] Wang T., Zhu J.-Y., Torralba A. and Efros A. A., *Dataset Distillation*. 2020. DOI: 10.48550/arXiv.1811.10959.

[89] Li G., Togo R., Ogawa T. and Haseyama M., *Importance-aware adaptive dataset distillation*, Neural Networks, vol. 172, 2024. DOI: 10.1016/j.neunet.2024.106154.

[90] Wu Z., Gao X., Qian Y., Hao Y. and Chen M., *Dynamic differential privacy-based dataset condensation*, Neurocomputing, vol. 608, 2024. DOI: 10.1016/j.neucom.2024.128394.

[91] Jin H. and Kim E., *Dataset condensation with coarse-to-fine regularization*, *Pattern Recognition Letters*, vol. 188, 2025, pp. 178–184. DOI: 10.1016/j.patrec.2024.12.018.

[92] Lei S. and Tao D., *A Comprehensive Survey of Dataset Distillation*, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 1, Jan. 2024, pp. 17–32. DOI: 10.1109/TPAMI.2023.3322540.

[93] Cazenavette G., Wang T., Torralba A., Efros A. A. and Zhu J.-Y., *Dataset Distillation by Matching Training Trajectories*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10718–10727. DOI: 10.1109/CVPR52688.2022.01045.

[94] Zhao B. and Bilen H., *Dataset Condensation with Differentiable Siamese Augmentation*, in Proceedings of the 38th International Conference on Machine Learning, 2021, vol. 139, pp. 12674−12685. DOI: 10.48550/arXiv.2102.08259.

[95] Czogała E. and Pedrycz W., *Elementy i metody teorii zbiorów rozmytych*. Warszawa: Państwowe Wydawnictwo Naukowe (PWN), 1985.

[96] Yager R. R. and Filev D. P., *Podstawy modelowania i sterowania rozmytego*. Warszawa: Wydawnictwo Naukowo-Techniczne (WNT), 1995.

[97] Baczyński D., Bielecki S., Parol M., Piotrowski P. and Wasilewski J., *Sztuczna inteligencja w praktyce. Laboratorium*. Warszawa: Oficyna Wydawnicza Politechniki Warszawskiej, 2008.

[98] Rutkowska D., Piliński M. and Rutkowski L., *Sieci neuronowe, algorytmy genetyczne i systemy rozmyte*. Warszawa and Łódź: Wydawnictwo Naukowe PWN, 1997.

[99] Łachwa A., *Rozmyty świat zbiorów, liczb, relacji, faktów, reguł i decyzji*. Warszawa: Akademicka Oficyna Wydawnicza EXIT, 2001.

[100] Piegat A., *Modelowanie i sterowanie rozmyte*. Warszawa: Akademicka Oficyna Wydawnicza EXIT, 1999.

[101] Rutkowski L., *Metody i techniki sztucznej inteligencji*. Warszawa: Wydawnictwo Naukowe PWN, 2012.

[102] Prade H. and Testemale C., *Generalizing database relational algebra for the treatment of incomplete or uncertain information and vague queries*, Information Sciences, vol. 34, no. 2, 1984, pp. 115–143. DOI: 10.1016/0020-0255(84)90020-3.

[103] Zemankova M. and Kandel A., *Implementing imprecision in information systems*, Information Sciences, vol. 37, no. 1, 1985, pp. 107–141. DOI: 10.1016/0020-0255(85)90008-8.

[104] Umano M., Hatono I. and Tamura H., *Fuzzy Database Systems*, in Proceedings of the 1995 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE/IFES'95), 1995, vol. 5, pp. 35–36. DOI: 10.1109/FUZZY.1995.410030.

[105] Kacprzyk J. and Zadrozny S., *SQLf and FQUERY for Access*, in Proceedings Joint 9th IFSA World Congress and 20th NAFIPS International Conference (Cat. No. 01TH8569), 2001, vol. 4, pp. 2464–2469 vol.4. DOI: 10.1109/NAFIPS.2001.944459.

[106] Tahani V., *A conceptual framework for fuzzy query processi - A step toward very intelligent database systems*, *Information Processing & Management*, vol. 13, no. 5, 1977, pp. 289–303. DOI: 10.1016/0306-4573(77)90018-8.

[107] Kacprzyk J., Zadrożny S. and Ziołkowski A., *FQUERY III+: A "human-consistent" database querying system based on fuzzy logic with linguistic quantifiers*, Information Systems, vol. 14, no. 6, 1989, pp. 443–453. DOI: 10.1016/0306-4379(89)90012-4.

[108] Kamel M., Hadfield B. and Ismail M., *Fuzzy query processing using clustering techniques*, *Information Processing & Management*, vol. 26, no. 2, 1990, pp. 279–293. DOI: 10.1016/0306-4573(90)90031-V.

[109] Bosc P. and Lietard L., *Fuzzy integrals and database flexible querying*, in Proceedings of IEEE 5th International Fuzzy Systems, 1996, vol. 1, pp. 100–106 vol.1. DOI: 10.1109/FUZZY.1996.551726.

[110] Takahashi Y., *A fuzzy query language for relational databases*, IEEE Transactions on Systems, Man, and Cybernetics, vol. 21, no. 6, 1991, pp. 1576–1579. DOI: 10.1109/21.135699.

[111] Kacprzyk J. and Zadrożny S., *Fquery for Access: Fuzzy Querying for a Windows-Based DBMS*, in Fuzziness in Database Management Systems, 1995, pp. 415–433. DOI: 10.1007/978-3-7908-1897-0_18.

[112] Galindo J., *New characteristics in FSQL, a fuzzy SQL for fuzzy databases*, in Proceedings of the 4th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases (AIKED), 2005, pp. 4:1-4:9.

[113] Bordogna G. and Psaila G., *Customizable Flexible Querying in Classical Relational Databases*, in Handbook of Research on Fuzzy Information Processing in Databases, Galindo J. Eds. IGI Global, 2008, pp. 191–217. DOI: 10.4018/978-1-59904-853-6.ch008.

[114] Dubois Didier and Prade H., *Bipolarity in Flexible Querying*, in Flexible Query Answering Systems, 2002, pp. 174–182. DOI: 10.1007/3-540-36109-X_14.

[115] Pivert O. and Bosc P., *Fuzzy Preference Queries to Relational Databases*. Imperial College Press, 2012. DOI: 10.1142/p840.

[116] Alam T., Khan A. and Alam F., *Punctuation Restoration using Transformer Models for High-and Low-Resource Languages*, in Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT), 2020, pp. 132–142. DOI: 10.18653/v1/2020.wnut-1.18.

[117] Guhr O., Schumann A. K., Bahrmann F. and Böhme H. J., *FullStop: Multilingual Deep Models for Punctuation Prediction*, in Proceedings of the Swiss Text Analytics Conference 2021 (SwissText), 2021, vol. 2957, pp. 1–9.

[118] Min J., Lee M., Lee W. and Lee Y., *Punctuation Restoration Improves Structure Understanding without Supervision*, 2024. DOI: 10.48550/arXiv.2402.08382.

[119] Sazhok M., Poltieva A., Robeiko V., Seliukh R. and Fedoryn D., *Punctuation Restoration for Ukrainian Broadcast Speech Recognition System based on Bidirectional Recurrent Neural Network and Word Embeddings*, in Proceedings of the 5th International Conference on Computational Linguistics and Intelligent Systems (COLINS), Volume I: Main Conference, 2021, vol. 2870, pp. 300–310.

Spis tabel

Oświadczenie autora o wkładzie merytorycznym i samodzielnym opracowaniu prac stanowiących podstawę rozprawy doktorskiej

Politechnika Krakowska
Wydział Inżynierii
Elektrycznej i Komputerowej

mgr inż. Grzegorz Nowakowski
Katedra Automatyki i Informatyki E-1
Politechnika Krakowska

Kraków, 13.01.2026r.

OŚWIADCZENIE

Ja, niżej podpisany **Grzegorz Nowakowski**, oświadczam, że:

- publikacje złożone do oceny, oznaczone w manuskrypcie jako [1]–[9], zostały przeze mnie opracowane samodzielnie w zakresie mojego wkładu autorskiego,
- wykaz publikacji złożonych do oceny wraz z opisem mojego wkładu został przedstawiony w rozdziale pierwszym: „Wykaz publikacji stanowiących podstawę rozprawy", a pełne teksty publikacji zostały zamieszczone w rozdziale czwartym: „Teksty publikacji stanowiących podstawę rozprawy",
- publikacje [1] i [2] są publikacjami jednoautorskimi i zostały przeze mnie opracowane samodzielnie,
- publikacje [3]–[9] są publikacjami współautorskimi. Oświadczam, że mój wkład w powstanie każdej z tych publikacji był znaczący, samodzielnie wykonany w zakresie przypisanych mi prac oraz możliwy do jednoznacznego wyodrębnienia,
- w odniesieniu do publikacji współautorskich [3]–[9] posiadam oświadczenia współautorów dotyczące ich wkładu w przygotowanie publikacji złożonych do oceny (w formie opisowej), zamieszczone w części: „Oświadczenia współautorów o wkładzie merytorycznym w publikacje wchodzące w skład rozprawy",
- wskazane publikacje oraz manuskrypt rozprawy zostały przygotowane z poszanowaniem zasad rzetelności naukowej.

Podpis: G. Nowakowski Data: 13.01.2026

POLITECHNIKA KRAKOWSKA
im. Tadeusza Kościuszki
WYDZIAŁ INŻYNIERII ELEKTRYCZNEJ I KOMPUTEROWEJ
Katedra Automatyki i Informatyki
31-155 Kraków, ul. Warszawska 24
tel.(+48) 12-628-26-85

Katedra Automatyki i Informatyki
Wydział Inżynierii Elektrycznej i Komputerowej
ul. Warszawska 24, 31-155 Kraków
tel. 12 628 26 85, e-1@pk.edu.pl

wieik.pk.edu.pl

Oświadczenia współautorów o wkładzie merytorycznym w publikacje wchodzące w skład rozprawy

Politechnika Krakowska
Wydział Inżynierii
Elektrycznej i Komputerowej

prof. dr hab. inż. Sergii Telenyk
Katedra Automatyki i Informatyki E-1
Politechnika Krakowska

Kraków, 13.01.2026r.

OŚWIADCZENIE

Ja, niżej podpisany **Sergii Telenyk**, oświadczam, że jestem współautorem publikacji:

Telenyk S., Nowakowski G., Yefremov K., Khmeliuk V., *Logics based application integration for interdisciplinary scientific investigations, w: 2017 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), 2017, vol. 2, s. 1026–1031, DOI: 10.1109/IDAACS.2017.8095241.*

Oświadczam, że mój wkład w powstanie powyższej publikacji wyniósł **35%** i obejmował w szczególności:

- współudział w koncepcji i uzasadnieniu podejścia oraz doprecyzowaniu problemu
- wkład w strukturę artykułu oraz opracowanie tła/related work
- udział w opisie podejścia na poziomie ogólnym (architektura i orkiestracja usług) oraz w dyskusji i wnioskach
- przegląd merytoryczny i redakcję tekstu oraz uwagi do wersji finalnej.

Podpis: /Sergii Telenyk/ Data: 13.01.2026

POLITECHNIKA KRAKOWSKA
im. Tadeusza Kościuszki
WYDZIAŁ INŻYNIERII ELEKTRYCZNEJ I KOMPUTEROWEJ
Katedra Automatyki i Informatyki
31-155 Kraków, ul. Warszawska 24
tel.(+48) 12-628-26-85

Katedra Automatyki i Informatyki
Wydział Inżynierii Elektrycznej i Komputerowej
ul. Warszawska 24, 31-155 Kraków
tel. 12 628 26 85, e-1@pk.edu.pl

wieik.pk.edu.pl

Kostiantyn Yefremov
Director at Institute for Applied System Analysis
National Technical University of Ukraine
"Igor Sikorsky Kyiv, Polytechnic Institute"
Kyiv, Ukraine

Kyiv, 15.01.2026

STATEMENT

I, the undersigned **Kostiantyn Yefremov**, hereby declare that I am a co-author of the publication:

Telenyk S., Nowakowski G., Yefremov K., Khmeliuk V., Logics based application integration for interdisciplinary scientific investigations, in: 2017 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), 2017, vol. 2, pp. 1026–1031, DOI: 10.1109/IDAACS.2017.8095241.

I declare that my contribution to the above publication amounted to 5% and included in particular:

- domain consultations on WDC/WDS (requirements and practical constraints)
- verification of the use case scenario and the assumptions adopted
- domain-specific comments on the content (substantive/technical review)

Signature Єфремов К.В. Date: 15.01.2026

Підпис гр. Єфремова К.В.
ЗАСВІДЧУЮ
Відділ кадрів
підпис пр-ще

Volodymyr Khmeliuk
National Technical University of Ukraine
"Igor Sikorsky Kyiv, Polytechnic Institute"
Kyiv, Ukraine

Kyiv, 15.01.2026

STATEMENT

I, the undersigned **Volodymyr Khmeliuk**, hereby declare that I am a co-author of the publication:

Telenyk S., Nowakowski G., Yefremov K., Khmeliuk V., Logics based application integration for interdisciplinary scientific investigations, in: 2017 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), 2017, vol. 2, pp. 1026–1031, DOI: 10.1109/IDAACS.2017.8095241.

I declare that my contribution to the above publication amounted to 5% and included in particular:

- technical consultations regarding a distributed/agent-based environment (feasibility and deployment/execution aspects)
- comments on the architecture description and implementation details (integration/communication, service registry and orchestration)
- technical corrections and proofreading

Signature V. Khmeliuk Date: 15.01.2026.

Підпис гр. Хмелюка В. С.
ЗАСВІДЧУЮ
Відділ кадрів
підпис пр-ще

Politechnika Krakowska
Wydział Inżynierii
Elektrycznej i Komputerowej

prof. dr hab. inż. Sergii Telenyk
Katedra Automatyki i Informatyki E-1
Politechnika Krakowska

Kraków, 13.01.2026r.

OŚWIADCZENIE

Ja, niżej podpisany **Sergii Telenyk**, oświadczam, że jestem współautorem publikacji:

Nowakowski G., Telenyk S., Yefremov K. and Khmeliuk V., *The approach to applications integration for world data center interdisciplinary scientific investigations*, in Proceedings of the 2019 Federated Conference on Computer Science and Information Systems, FedCSIS 2019, 2019. https://doi.org/10.15439/2019F71.

Oświadczam, że mój wkład w powstanie powyższej publikacji wyniósł **15%** i obejmował w szczególności:

- dopracowanie metodologii podejścia oraz doprecyzowanie problemu i założeń na poziomie koncepcyjnym
- uwagi do opisu formalizmu (spójność i sposób prezentacji wnioskowania oraz rekonstrukcji rozwiązania)
- redakcję merytoryczną manuskryptu, w tym uwagi do struktury i czytelności opisu.

Podpis: /Sergii Telenyk/ Data: 13.01.2026

POLITECHNIKA KRAKOWSKA
im. Tadeusza Kościuszki
WYDZIAŁ INŻYNIERII ELEKTRYCZNEJ I KOMPUTEROWEJ
Katedra Automatyki i Informatyki
31-155 Kraków, ul. Warszawska 24
tel.(+48) 12-628-26-85

Kostiantyn Yefremov
Director at Institute for Applied System Analysis
National Technical University of Ukraine
"Igor Sikorsky Kyiv, Polytechnic Institute"
Kyiv, Ukraine

Kyiv, 15.01.2026

STATEMENT

I, the undersigned **Kostiantyn Yefremov**, hereby declare that I am a co-author of the publication:

Nowakowski G., Telenyk S., Yefremov K. and Khmeliuk V., *The approach to applications integration for world data center interdisciplinary scientific investigations*, in Proceedings of the 2019 Federated Conference on Computer Science and Information Systems, FedCSIS 2019, 2019. https://doi.org/10.15439/2019F71.

I declare that my contribution to the above publication amounted to 5% and included in particular:

- domain consultations regarding the WDS/WDC environment (requirements and real-world application constraints)
- verification of the use case scenario and the practical assumptions adopted
- domain-specific comments on the content (substantive/technical review).

Signature Ефремов К.В. Date: 15.01.2026

Підпис гр. Єфремова В.В.
ЗАСВІДЧУЮ
Відділ кадрів
підпис пр-ще

Volodymyr Khmeliuk
National Technical University of Ukraine
"Igor Sikorsky Kyiv, Polytechnic Institute"
Kyiv, Ukraine

Kyiv, 15.01.2026

STATEMENT

I, the undersigned **Volodymyr Khmeliuk**, hereby declare that I am a co-author of the publication:

Nowakowski G., Telenyk S., Yefremov K. and Khmeliuk V., *The approach to applications integration for world data center interdisciplinary scientific investigations*, in Proceedings of the 2019 Federated Conference on Computer Science and Information Systems, FedCSIS 2019, 2019. https://doi.org/10.15439/2019F71.

I declare that my contribution to the above publication amounted to 5% and included in particular:

- technical consultations regarding the prototype (feasibility and operational/deployment aspects of the agent-based approach)
- comments on the technical architecture description (service communication and orchestration)
- technical corrections and terminology consistency (proofreading).

Signature V. Khmeliuk Date: 15.01.2026

Підпис гр. Хмелюка В.С.
ЗАСВІДЧУЮ
Відділ кадрів
підпис пр-ще

Politechnika Krakowska
Wydział Inżynierii
Elektrycznej i Komputerowej

prof. dr hab. inż. Sergii Telenyk
Katedra Automatyki i Informatyki E-1
Politechnika Krakowska

Kraków, 13.01.2026r.

OŚWIADCZENIE

Ja, niżej podpisany **Sergii Telenyk**, oświadczam, że jestem współautorem publikacji:

Nowakowski G., Telenyk S., Yefremov K. and Khmeliuk V., *Simple and Flexible Way to Integrate Heterogeneous Information Systems and Their Services into the World Data System*, *Journal of Automation, Mobile Robotics and Intelligent Systems*, vol. 15, no. 4, (Sep. 2022), pp. 76–90. https://doi.org/10.14313/JAMRIS/4-2021/29

Oświadczam, że mój wkład w powstanie powyższej publikacji wyniósł **15%** i obejmował w szczególności:

- dopracowanie koncepcji i metodologii podejścia oraz argumentacji na poziomie koncepcyjno-opisowym
- uwagi do narracji pracy, w tym sposobu prezentacji rozwiązania i osadzenia na tle prac powiązanych
- przegląd merytoryczny i redakcję tekstu, w tym uwagi do spójności i struktury opisu.

Podpis: /Sergii Telenyk/ Data: 13.01.2026

POLITECHNIKA KRAKOWSKA
im. Tadeusza Kościuszki
WYDZIAŁ INŻYNIERII ELEKTRYCZNEJ I KOMPUTEROWEJ
Katedra Automatyki i Informatyki
31-155 Kraków, ul. Warszawska 24
tel.(+48) 12-628-26-85

Katedra Automatyki i Informatyki
Wydział Inżynierii Elektrycznej i Komputerowej
ul. Warszawska 24, 31-155 Kraków
tel. 12 628 26 85, e-1@pk.edu.pl

Kostiantyn Yefremov
Director at Institute for Applied System Analysis
National Technical University of Ukraine
"Igor Sikorsky Kyiv, Polytechnic Institute"
Kyiv, Ukraine

Kyiv, 15.01.2026

STATEMENT

I, the undersigned **Kostiantyn Yefremov**, hereby declare that I am a co-author of the publication:

Nowakowski G., Telenyk S., Yefremov K. and Khmeliuk V., *Simple and Flexible Way to Integrate Heterogeneous Information Systems and Their Services into the World Data System, Journal of Automation, Mobile Robotics and Intelligent Systems*, vol. 15, no. 4, (Sep. 2022), pp. 76–90. https://doi.org/10.14313/JAMRIS/4-2021/29

I declare that my contribution to the above publication amounted to 5% and included in particular:

- domain consultations regarding the WDS/WDC environment (context and real-world application conditions)
- verification of the scenario/example and the practical assumptions adopted
- domain-specific comments on the content (substantive review).

Signature Єфремов К.В. Date: 15.01.2026

Підпис гр. Єфремова К.В.
ЗАСВІДЧУЮ
Відділ кадрів
підпис пр-ще

Volodymyr Khmeliuk
National Technical University of Ukraine
"Igor Sikorsky Kyiv, Polytechnic Institute"
Kyiv, Ukraine

Kyiv, 15.01.2026

STATEMENT

I, the undersigned **Volodymyr Khmeliuk**, hereby declare that I am a co-author of the publication:

Nowakowski G., Telenyk S., Yefremov K. and Khmeliuk V., *Simple and Flexible Way to Integrate Heterogeneous Information Systems and Their Services into the World Data System*, *Journal of Automation, Mobile Robotics and Intelligent Systems*, vol. 15, no. 4, (Sep. 2022), pp. 76–90. https://doi.org/10.14313/JAMRIS/4-2021/29.

I declare that my contribution to the above publication amounted to 5% and included in particular:

- technical consultations on feasibility and implementation/execution elements (agent-based approach and service communication)
- comments on the technical description of deployment/execution, including protocols and service orchestration
- technical and terminological corrections (proofreading).

Signature Khmeliuk Date: 15.01.2026

Підпис гр. Хмелюка В.С.
ЗАСВІДЧУЮ
Відділ кадрів
підпис пр-ще

Politechnika Krakowska
Wydział Inżynierii
Elektrycznej i Komputerowej

prof. dr hab. inż. Sergii Telenyk
Katedra Automatyki i Informatyki E-1
Politechnika Krakowska

Kraków, 13.01.2026r.

OŚWIADCZENIE

Ja, niżej podpisany **Sergii Telenyk**, oświadczam, że jestem współautorem publikacji:

Telenyk S., Nowakowski G., Zharikov E. and Vovk J., *Conceptual foundations of the use of formal models and methods for the rapid creation of web applications*, in Proceedings of the 2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS 2019, 2019, vol. 1. https://doi.org/10.1109/IDAACS.2019.8924416

Oświadczam, że mój wkład w powstanie powyższej publikacji wyniósł **15%** i obejmował w szczególności:

- współudział w koncepcji i określeniu zakresu pracy oraz doprecyzowaniu założeń podejścia
- wkład do części opisowej dotyczącej architektury rozwiązania oraz osadzenia podejścia w kontekście formalnych metod projektowania (tło i odniesienia)
- dopracowanie argumentacji oraz przegląd merytoryczny i redakcję manuskryptu.

Podpis: /Sergii Telenyk/ Data: 13.01.2026

POLITECHNIKA KRAKOWSKA
im. Tadeusza Kościuszki
WYDZIAŁ INŻYNIERII ELEKTRYCZNEJ I KOMPUTEROWEJ
Katedra Automatyki i Informatyki
31-155 Kraków, ul. Warszawska 24
tel.(+48) 12-628-26-85

Katedra Automatyki i Informatyki
Wydział Inżynierii Elektrycznej i Komputerowej
ul. Warszawska 24, 31-155 Kraków
tel. 12 628 26 85, e-1@pk.edu.pl

wieik.pk.edu.pl

Eduard Zharikov
Head of the Department of Computer Science
and Software Engineering
National Technical University of Ukraine
"Igor Sikorsky Kyiv, Polytechnic Institute"
Kyiv, Ukraine

Kyiv, 15.01.2026

STATEMENT

I, the undersigned **Eduard Zharikov**, hereby declare that I am a co-author of the publication:

Telenyk S., Nowakowski G., Zharikov E. and Vovk J., *Conceptual foundations of the use of formal models and methods for the rapid creation of web applications*, in Proceedings of the 2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS 2019, 2019, vol. 1. https://doi.org/10.1109/IDAACS.2019.8924416

I declare that my contribution to the above publication amounted to 5% and included in particular:

- substantive consultations on the concept and the way selected parts of the work are presented
- comments on the text and refinement of selected passages, including clarifying wording and terminology

Signature Жаріков Е.В. Date: 15.01.2026

Підпис гр. Жарікова Е.В.
ЗАСВІДЧУЮ
Відділ кадрів
підпис пр-ще

Jewhenii Vovk
National Technical University of Ukraine
"Igor Sikorsky Kyiv, Polytechnic Institute"
Kyiv, Ukraine

Kyiv, 15.01.2026

STATEMENT

I, the undersigned **Jewhenii Vovk**, hereby declare that I am a co-author of the publication:

Telenyk S., Nowakowski G., Zharikov E. and Vovk J., *Conceptual foundations of the use of formal models and methods for the rapid creation of web applications*, in Proceedings of the 2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS 2019, 2019, vol. 1. https://doi.org/10.1109/IDAACS.2019.8924416

I declare that my contribution to the above publication amounted to 5% and included in particular:

- preparation of demonstration/technical components supporting the presented concept
- support in refining the technical descriptions related to the solution demonstration (at the implementation/execution level).

Signature _____ / Вовк Є.А. Date: 15.01.2026

Підпис гр. Вовк Є.А.
ЗАСВІДЧУЮ
Відділ кадрів
підпис пр-ще

Politechnika Krakowska
Wydział Inżynierii
Elektrycznej i Komputerowej

prof. dr hab. inż. Sergii Telenyk
Katedra Automatyki i Informatyki E-1
Politechnika Krakowska

Kraków, 13.01.2026r.

OŚWIADCZENIE

Ja, niżej podpisany **Sergii Telenyk**, oświadczam, że jestem współautorem publikacji:

Telenyk S., Nowakowski G., Gavrilenko O., Miahkyi M. and Khalus O., *An analysis of the influence of famous people's posts on social networks on the cryptocurrency exchange rate*, *Bulletin of the Polish Academy of Sciences: Technical Sciences*, vol. 72, no. 4, (Apr. 2024), p. e150117. https://doi.org/10.24425/BPASTS.2024.150117

Oświadczam, że mój wkład w powstanie powyższej publikacji wyniósł **30%** i obejmował w szczególności:

- udział w dopracowaniu opisu badań i procedury eksperymentalnej
- udział w interpretacji wyników, wnioskach oraz omówieniu ograniczeń
- redakcję merytoryczną manuskryptu.

Podpis: [Sergii Telenyk]

Data: 13.01.2026

POLITECHNIKA KRAKOWSKA
im. Tadeusza Kościuszki
WYDZIAŁ INŻYNIERII ELEKTRYCZNEJ I KOMPUTEROWEJ
Katedra Automatyki i Informatyki
31-155 Kraków, ul. Warszawska 24
tel.(+48) 12-628-26-85

Katedra Automatyki i Informatyki
Wydział Inżynierii Elektrycznej i Komputerowej
ul. Warszawska 24, 31-155 Kraków
tel. 12 628 26 85, e-1@pk.edu.pl

Olena Gavrilenko
National Technical University of Ukraine
"Igor Sikorsky Kyiv, Polytechnic Institute"
Kyiv, Ukraine

Kyiv, 15.01.2026

STATEMENT

I, the undersigned **Olena Gavrilenko**, hereby declare that I am a co-author of the publication:

Telenyk S., Nowakowski G., Gavrilenko O., Miahkyi M. and Khalus O., *An analysis of the influence of famous people's posts on social networks on the cryptocurrency exchange rate*, *Bulletin of the Polish Academy of Sciences: Technical Sciences*, vol. 72, no. 4, (Apr. 2024), p. e150117. https://doi.org/10.24425/BPASTS.2024.150117

I declare that my contribution to the above publication amounted to 5% and included in particular:

- consultations on data selection, the scope of analysis, and results reporting
- substantive comments on the analytical and descriptive sections
- comments on the final version of the manuscript

Signature Гавриленко О.В. Date: 15.01.2026

Підпис гр. Гавриленко О.В.
ЗАСВІДЧУЮ
Відділ кадрів
підпис пр-ще

Mykhailo Miahkyi
National Technical University of Ukraine
"Igor Sikorsky Kyiv, Polytechnic Institute"
Kyiv, Ukraine

Kyiv, 15.01.2026

STATEMENT

I, the undersigned **Mykhailo Miahkyi**, hereby declare that I am a co-author of the publication:

Telenyk S., Nowakowski G., Gavrilenko O., Miahkyi M. and Khalus O., *An analysis of the influence of famous people's posts on social networks on the cryptocurrency exchange rate*, *Bulletin of the Polish Academy of Sciences: Technical Sciences*, vol. 72, no. 4, (Apr. 2024), p. e150117. https://doi.org/10.24425/BPASTS.2024.150117

I declare that my contribution to the above publication amounted to 5% and included in particular:

- implementation of selected technical components of the computational pipeline (data and runs/executions)
- technical support for the experiments (configuration, debugging, reproducibility).

Signature Тягкий М.Ю. Date: 15.01.2026

Підпис гр. Тягкого М.Ю.
ЗАСВІДЧУЮ
Відділ кадрів
підпис пр-ще

Olena Khalus
National Technical University of Ukraine
"Igor Sikorsky Kyiv, Polytechnic Institute"
Kyiv, Ukraine

Kyiv, 15.01.2026

STATEMENT

I, the undersigned **Olena Khalus**, hereby declare that I am a co-author of the publication:

Telenyk S., Nowakowski G., Gavrilenko O., Miahkyi M. and Khalus O., *An analysis of the influence of famous people's posts on social networks on the cryptocurrency exchange rate*, *Bulletin of the Polish Academy of Sciences: Technical Sciences*, vol. 72, no. 4, (Apr. 2024), p. e150117. https://doi.org/10.24425/BPASTS.2024.150117

I declare that my contribution to the above publication amounted to 5% and included in particular:

- consultations and support in organizing the experimental part, including verification of run correctness
- comments on the compilation and presentation of results, as well as on the final version of the paper.

Signature Халус О.А. Date: 15.01.2026

Підпис гр. Халус О.А.
ЗАСВІДЧУЮ
Відділ кадрів
підпис пр-ще

Volodymyr Shymkovych
National Technical University of Ukraine
"Igor Sikorsky Kyiv, Polytechnic Institute"
Kyiv, Ukraine

Kyiv, 15.01.2026

STATEMENT

I, the undersigned **Volodymyr Shymkovych**, hereby declare that I am a co-author of the publication:

Shymkovych V., Nowakowski G. and Telenyk S., *Joint Punctuation Restoration and Text* Capitalisation *with a Hybrid XLM-RoBERTa–LSTM Model*, in IDAACS 2025: proceedings of the 13th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), 2025, vol. 1, IEEE, pp. 990–997. DOI: 10.1109/IDAACS68557.2025.11322226.

I declare that my contribution to the above publication amounted to 40% and included in particular:

- implementation and training of the hybrid model (XLM-RoBERTa + BiLSTM + classifier)
- conducting experiments on IWSLT 2012 and compiling the results (tables/figures, per-class) in line with the adopted reporting procedure
- refining the description of the implementation and experimental section in the manuscript.

Signature Шимковач В.М. Date: 15.01.2026

Підпис гр. Шимковича В.М.
ЗАСВІДЧУЮ
Відділ кадрів
підпис пр-ще

Politechnika Krakowska
Wydział Inżynierii
Elektrycznej i Komputerowej

prof. dr hab. inż. Sergii Telenyk
Katedra Automatyki i Informatyki E-1
Politechnika Krakowska

Kraków, 13.01.2026r.

OŚWIADCZENIE

Ja, niżej podpisany **Sergii Telenyk**, oświadczam, że jestem współautorem publikacji:

Shymkovych V., Nowakowski G. and Telenyk S., *Joint Punctuation Restoration and Text* Capitalisation *with a Hybrid XLM-RoBERTa–LSTM Model*, in IDAACS 2025: proceedings of the 13th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), 2025, vol. 1, IEEE, pp. 990–997. DOI: 10.1109/IDAACS68557.2025.11322226.

Oświadczam, że mój wkład w powstanie powyższej publikacji wyniósł **5%** i obejmował w szczególności:

- konsultacje merytoryczne oraz przegląd manuskryptu pod kątem spójności metodologii i czytelności opisu
- uwagi redakcyjne do wybranych fragmentów i doprecyzowanie terminologii.

Podpis: /Sergii Telenyk/ Data: 13.01.2026

POLITECHNIKA KRAKOWSKA
im. Tadeusza Kościuszki
WYDZIAŁ INŻYNIERII ELEKTRYCZNEJ I KOMPUTEROWEJ
Katedra Automatyki i Informatyki
31-155 Kraków, ul. Warszawska 24
tel.(+48) 12-628-26-85

Katedra Automatyki i Informatyki
Wydział Inżynierii Elektrycznej i Komputerowej
ul. Warszawska 24, 31-155 Kraków
tel. 12 628 26 85, e-1@pk.edu.pl

Yuri Gordienko
National Technical University of Ukraine
"Igor Sikorsky Kyiv, Polytechnic Institute"
Kyiv, Ukraine

Kyiv, 15.01.2026

STATEMENT

I, the undersigned **Yuri Gordienko**, hereby declare that I am a co-author of the publication:

Gordienko, Y., Nowakowski, G., Kochura, Y., Taran, V., Stirenko, S., *Generative Data Augmentation by Dataset Distillation*, in Artificial Intelligence for Biomedical Data: First International Workshop, AIBio 2025, Held in Conjunction with the European Conference on Artificial Intelligence, ECAI 2025, Communications in Computer and Information Science, 2026, vol. 2696, Cham, Springer, pp. 105-118, ISBN 978-3-032-17215-0. DOI: 10.1007/978-3-032-17216-7_9.

I declare that my contribution to the above publication amounted to 40% and included in particular:

- development and refinement of the algorithmic part (distillation and synthetic data generation) in accordance with the adopted formalization of the objective function
- running a substantial portion of the experiments and compiling the results and comparisons
- co-authorship of the methods and results sections and comments on the final version.

Signature Гордієнко Ю.Г. Date: 15.01.2026

Підпис гр. Гордієнко Ю.Г.
ЗАСВІДЧУЮ
Відділ кадрів
підпис пр-ще

Yuriy Kochura
National Technical University of Ukraine
"Igor Sikorsky Kyiv, Polytechnic Institute"
Kyiv, Ukraine

Kyiv, 15.01.2026

STATEMENT

I, the undersigned **Yuriy Kochura**, hereby declare that I am a co-author of the publication:

Gordienko, Y., Nowakowski, G., Kochura, Y., Taran, V., Stirenko, S., *Generative Data Augmentation by Dataset Distillation*, in Artificial Intelligence for Biomedical Data: First International Workshop, AIBio 2025, Held in Conjunction with the European Conference on Artificial Intelligence, ECAI 2025, Communications in Computer and Information Science, 2026, vol. 2696, Cham, Springer, pp. 105-118, ISBN 978-3-032-17215-0. DOI: 10.1007/978-3-032-17216-7_9.

I declare that my contribution to the above publication amounted to 10% and included in particular:

- methodological support: refining the experimental setup and analyzing the results
- substantive consultations and co-contribution to drafting parts of the discussion/related work section
- substantive review of the text and comments on the coherence of the argumentation.

Signature Когура Ю.І. Date: 15.01.2026 г.

Підпис гр. Когури Ю. І.
ЗАСВІДЧУЮ
Відділ кадрів
підпис пр-ще

Vladyslav Taran
National Technical University of Ukraine
"Igor Sikorsky Kyiv, Polytechnic Institute"
Kyiv, Ukraine

Kyiv, 15.01.2026

STATEMENT

I, the undersigned **Vladyslav Taran**, hereby declare that I am a co-author of the publication:

Gordienko, Y., Nowakowski, G., Kochura, Y., Taran, V., Stirenko, S., *Generative Data Augmentation by Dataset Distillation*, in Artificial Intelligence for Biomedical Data: First International Workshop, AIBio 2025, Held in Conjunction with the European Conference on Artificial Intelligence, ECAI 2025, Communications in Computer and Information Science, 2026, vol. 2696, Cham, Springer, pp. 105-118, ISBN 978-3-032-17215-0. DOI: 10.1007/978-3-032-17216-7_9.

I declare that my contribution to the above publication amounted to 5% and included in particular:

- implementation of components of the experimental code and computation pipeline (runs/executions, data generation, result artifacts)
- technical support for the experiments (configurations, correctness checks, and reproducibility).

Signature Таран В.І. Date: 15.01.2026 р.

Підпис гр. Таран В.І.
ЗАСВІДЧУЮ
Відділ кадрів
підпис пр-ще

Sergii Stirenko
National Technical University of Ukraine
"Igor Sikorsky Kyiv, Polytechnic Institute"
Vice-Rector for Research
Kyiv, Ukraine

Kyiv, 15.01.2026

STATEMENT

I, the undersigned **Sergii Stirenko**, hereby declare that I am a co-author of the publication:

Gordienko, Y., Nowakowski, G., Kochura, Y., Taran, V., Stirenko, S., *Generative Data Augmentation by Dataset Distillation*, in Artificial Intelligence for Biomedical Data: First International Workshop, AIBio 2025, Held in Conjunction with the European Conference on Artificial Intelligence, ECAI 2025, Communications in Computer and Information Science, 2026, vol. 2696, Cham, Springer, pp. 105-118, ISBN 978-3-032-17215-0. DOI: 10.1007/978-3-032-17216-7_9.

I declare that my contribution to the above publication amounted to 5% and included in particular:

- substantive consultations and manuscript review
- editorial comments and refinement of wording

Signature Стіренко С.Г. Date: 15.01.2026

Відділ кадрів і архівної справи
Підпис гр. Стіренко С.Г.
ЗАСВІДЧУЮ
Відділ кадрів
підпис пр-ще