Załącznik nr 3
Autoreferat

Dr inż. Adrianna Kozierkiewicz
Katedra Informatyki Stosowanej
Wydział Informatyki i Zarządzania
Politechnika Wrocławska

dr inż. Adrianna Kozierkiewicz
Katedra Informatyki Stosowanej
Wydział Informatyki i Zarządzania
Politechnika Wrocławska

Wrocław, 7.05.2021

AUTOREFERAT

1. Imię i nazwisko: Adrianna Kozierkiewicz

2. Posiadane dyplomy i stopnie naukowe

- 03.07.2007- uzyskanie tytułu zawodowego magistra inżyniera, Politechnika Wrocławska, temat pracy: „*Zastosowanie metod klasyfikacji przybliżonej w systemach E-learning*"
- 29.03.2011- uzyskanie stopnia naukowego doktora, Politechnika Wrocławska, dyscyplina naukowa: informatyka, temat pracy: „*Metoda wyznaczania personalizowanego scenariusza uczenia w systemach e-kształcenia*", obrona z wyróżnieniem, promotor prof. dr hab. inż. Ngoc Thanh Nguyen, recenzenci: prof. dr hab. Piotr Jędrzejowicz, prof. dr hab. inż. Marek Krętowski

3. Informacje o dotychczasowym zatrudnieniu i pełnionych funkcjach

Zatrudnienie:

- 2020- (nadal trwa) - Politechnika Wrocławska, Wydział Informatyki i Zarządzania, Katedra Informatyki Stosowanej, stanowisko: adiunkt badawczo-dydaktyczny, Senator Politechniki Wrocławskiej kadencji 2020-2024
- 01.12.2014 - 31.12.2019- Politechnika Wrocławska, Wydział Informatyki i Zarządzania, Katedra Systemów Informatycznych, stanowisko: adiunkt badawczo-dydaktyczny
- 01.02. 2013 - 30.11.2014- Politechnika Wrocławska, Wydział Informatyki i Zarządzania, Instytut Informatyki, stanowisko: adiunkt badawczo-dydaktyczny
- 01.02.2011 - 31.01.2013- Politechnika Wrocławska, Wydział Informatyki i Zarządzania, Instytut Informatyki, stanowisko: asystent

4. Omówienie osiągnięć, o których mowa w art. 219 ust. 1 pkt. 2 Ustawy

4.1 Tytuł osiągnięcia

Rozprawa habilitacyjna pt. ***Metody oceny efektywności procesu integracji informacji i wiedzy pochodzących z różnych źródeł z wykorzystaniem teorii consensusu*** stanowi cykl publikacji naukowych powiązanych tematycznie. W ramach podjętych badań osiągnięto następujące rezultaty:

- dla wybranych reprezentacji wiedzy (zbiorów z powtórzeniami, wektorów binarnych, pokryć oraz podziałów uporządkowanych) wyznaczono kryteria jakie muszą spełniać źródła wejściowe, aby być podatne na consensus. Sformułowane kryteria są warunkami wystarczającymi podatności źródeł wejściowych na consensus.
- dla wybranych reprezentacji wiedzy (m.in.: wektorów binarnych, pokryć oraz podziałów uporządkowanych) zaprezentowano kryteria, które pozwalają na stwierdzenie jakie własności muszą spełniać źródła wejściowe, aby wynik ich integracji charakteryzował się najwyższą lub zadaną jakością (według przyjętej w cyklu definicji). Sformułowane kryteria są warunkami wystarczającymi maksymalizacji jakości wyniku integracji źródeł wejściowych.
- opracowano zestaw metod integracji i oceny efektywności integracji ontologii zwanego FOKI (Framework for Ontological Knowledge Integration) w ramach, którego:
  - opracowano zestaw metod, które pozwalają na ocenę potencjalnego przyrostu wiedzy w procesie integracji ontologii na poziomie: konceptów, relacji, instancji, relacji instancji,
  - opracowano metodę półautomatycznego wyznaczania wartości semantyki atrybutów,
  - opracowano metodę oceny odwzorowań ontologii bez konieczności porównywania go z odwzorowaniem wzorcowym wyznaczonym przez eksperta.
- opracowano zestaw metod oceny ekspresywności informacji w federacyjnych hurtowniach danych jako kryterium efektywności procesu integracji cząstkowych hurtowni.

Opracowane rozwiązania były prezentowane na renomowanych konferencjach i publikowane w prestiżowych czasopismach.

4.2 Cykl publikacji powiązanych tematycznie

Monotematyczny cykl publikacji składa się z 10 prac:

[P1] Kozierkiewicz-Hetmańska A., 2017, *The analysis of expert opinions' consensus quality,* Information Fusion. 2017, vol. 34, s. 80-86
IF: 13.669
Punkty MNiSW (rok publikacji): 45 pkt
Punkty MNiSW (rok 2020): 200pkt

Praca została poświęcona analizie jakości consensusu. W pracy przedstawiono definicje oraz podstawowe pojęcia dotyczące wyznaczania consensusu. Przeanalizowano różne struktury wiedzy m.in. zbiory z powtórzeniami oraz wektory binarne, dla których wyznaczono kryteria pozwalające na stwierdzenie, jakie własności muszą spełniać źródła wejściowe, aby wynik ich integracji charakteryzował się najwyższą jakością. Wszystkie własności zostały udowodnione matematycznie. Każde twierdzenie zostało zilustrowane przykładami oraz wyjaśniono jego praktyczne zastosowanie. Ponadto udowodniono osiem twierdzeń, które pozwoliły na wskazanie własności źródeł wejściowych, zapewniających wysoką jakość wyniku ich integracji. Na ich podstawie wykazano, jak zmiana zawartości źródeł wejściowych lub ich ilości, może przyczynić się do poprawy jakości ich zintegrowanej wersji. Sformułowane

kryteria są warunkiem wystarczającym maksymalizacji jakości wyniku integracji źródeł wejściowych.

Indywidualny wkład poszczególnych autorów w powstanie publikacji:

- **Adrianna Kozierkiewicz (100%)**- pomysłodawca badań, opracowanie metodologii badań, zdefiniowanie podstawowych i pomocniczych pojęć, sformułowanie kryteriów w formie twierdzeń matematycznych, przeprowadzenie dowodów matematycznych dla sformułowanych twierdzeń, opracowanie przykładów ilustrujących opracowane twierdzenia

[P2] Kozierkiewicz-Hetmańska A., 2014, *The analysis of a susceptibility to the consensus for a few representations of collective knowledge*, International Journal of Software Engineering and Knowledge Engineering Vol. 24, No. 5, s. 759775. DOI: 10.1142/S0218194014500296

IF: 0.886
Punkty MNiSW (rok publikacji): 45 pkt
Punkty MNiSW (rok 2020): 15pkt

Praca została poświęcona analizie podatności profilu na consensus. W pracy przedstawiono definicje oraz podstawowe pojęcia dotyczące wyznaczania consensusu. Następnie formalnie zdefiniowano kryterium, które pozwala zweryfikować, czy elementy profilu są podatne na consensus. Inaczej ujmując, czy na podstawie zgromadzonych z różnych źródeł informacji i wiedzy, można wyznaczyć wiarygodny i spójny wynik ich integracji. W pracy przyjęto, że informacje i wiedza przechowywane są jako wektory binarne, zbiory z powtórzeniami, podziały i pokrycia uporządkowane. Dla przyjętych reprezentacji wiedzy zweryfikowano, jakie kryteria muszą spełniać źródła wejściowe, aby być podatne na consensus. Sfomułowano zestaw twierdzeń, które udowodniono analitycznie. Wykazano, że dla wielu przypadków wystarczająca jest nieparzysta liczba źródeł wejściowych. Dla wybranych reprezentacji wiedzy zaproponowano i udowodniono analitycznie bardziej złożone własności dotyczące zawartości źródeł wejściowych lub ich ilości. Innymi słowy, jeszcze przed przeprowadzeniem procesu integracji, za pomocą opracowanych kryteriów możliwa jest ocena, czy dane źródła wejściowe umożliwią wyznaczenie wiarygodnego, spójnego wyniku ich złączenia. Sformułowane kryteria są warunkami wystarczającymi podatności źródeł wejściowych na consensus.

Indywidualny wkład poszczególnych autorów w powstanie publikacji:

- **Adrianna Kozierkiewicz (100%)**- pomysłodawca badań, opracowanie metodologii badań, zdefiniowanie podstawowych i pomocniczych pojęć, sformułowanie kryteriów w formie twierdzeń matematycznych, przeprowadzenie dowodów matematycznych dla sformułowanych twierdzeń

[P3] Kozierkiewicz A., Pietranik M., Sitarczyk M., 2020, *Assessing the influence of conflict profile properties on the quality of consensus*. W: Intelligent Information and Database Systems: 12th Asian Conference, ACIIDS 2020, Phuket, Thailand, March 23-26, 2020:

proceedings. Pt. 1 / eds. Ngoc Thanh Nguyen [i in.]. Cham: Springer, cop. 2020. s. 25-36.(Lecture Notes in Computer Science. Lecture Notes in Artificial Intelligence, ISSN 0302-9743; vol. 12033)

Punkty MNiSW (rok 2020): 20pkt
Ranking Core (rok 2021): B

W pracy przedstawiono analizę jakości consensusu dla przypadku, gdy integrowane źródła wejściowe przechowywane są jako wektory binarne. W ramach badań, poszukiwano odpowiedzi na pytanie, jakie własności musi spełniać profil wiedzy (jego wielkość oraz zawartość), aby uzyskać consensus o minimalnej, założonej jakości. Prace podzielono na dwie części. W pierwszym przypadku założono, że consensus spełnia kryterium 1-optymalności, czyli odległość od elementów profilu jest najmniejsza. W drugim testowano consensus 2-optymalny, którego kwadrat odległości od elementów profilu jest najmniejszy. Pierwsza część badań, ze względu na własności consensusu 1-optymalnego, umożliwiła przeprowadzenie dowodu matematycznego, który wykazał, że na jakość consensusu wpływa jedynie jego zawartość (proporcja zer i jedynek w wektorach binarnych). Druga część badań polegała na eksperymentalnej analizie źródeł wejściowych do integracji. Wyniki badań oraz ich statystyczna analiza za pomocą testu Wilcoxona, pozwoliły wyciągnąć wniosek, że w większości przypadków, aby uzyskać wynik integracji źródeł wejściowych charakteryzujący się wysoką jakością, wystarczy rozważyć od 3 do 5 źródeł.

Indywidualny wkład poszczególnych autorów w powstanie publikacji:

- **Adrianna Kozierkiewicz (50%)** - pomysłodawca badań, opracowanie metodologii badań, opracowanie podstawowych definicji, analityczna weryfikacja hipotezy badawczej, przeprowadzenie dowodu dla sformułowanego twierdzenia
- **Marcin Pietranik (5%)** – studia literaturowe
- **Mateusz Sitarczyk (45%)** – współopracowanie metodologii badań, eksperymentalna weryfikacja założonej hipotezy, przeprowadzenie analizy statystycznej

[P4] Hnatkowska B., Kozierkiewicz A., Pietranik M., 2020, *Semi-Automatic Definition of Attribute Semantics for the Purpose of Ontology Integration.* IEEE Access 8: 107272-107284

IF: 3.745
Punkty MNiSW (rok 2020): 100pkt

Opracowanie półautomatycznej metody generowania wartości semantyki atrybutów było głównym celem badań zaprezentowanych w pracy. Semantyka atrybutów jest z jednym z kluczowych elementów oceny przyrostu wiedzy na poziomie konceptów. W artykule przedstawiono prosty przykład ilustrujący jak różnie zdefiniowane wartości semantyk dla wybranego atrybutu wpływają na ostateczny wynik przyrostu wiedzy. W pierwszej części pracy przedstawiono opracowaną metodę wyznaczania wartości semantyki atrybutów, która wykorzystuje dostępne, zewnętrzne źródło wiedzy jakim jest WordNet. Następnie, metoda została zaimplementowana w prototypie narzędzia zwanego FOKI (Framework for Ontological Knowledge Integration), którego celem jest zarówno integracja, jak i ocena efektywności

procesu integracji ontologii. W ramach eksperymentu przygotowano zestaw ontologii testowych, które są udostępnione przez organizacje OAEI. Grupa ochotników miała zaproponować wartości semantyki atrybutów na dwa sposoby: ręcznie, w oparciu o własną wiedzę ekspercką oraz za pomocą opracowanej metody. Następnie, dla tak przygotowanych ontologii, obliczono przyrost wiedzy. Analiza wyników pozwoliła wyciągnąć wniosek, że czas generowania wartości semantyki za pomocą opracowanej metody jest istotnie krótszy niż czas generowania wartości semantyki w sposób manualny. Analiza statystyczna wyników eksperymentów za pomocą testu Wilcoxona potwierdziła skuteczność zaproponowanej metody.

Indywidualny wkład poszczególnych autorów w powstanie publikacji:

- **Bogumiła Hnatkowska (40%)** - opracowanie metodologii badań, przygotowanie przypadków testowych do eksperymentu, częściowe przeprowadzenie badań, konsultacje nad zaprojektowaną metodą, implementacja środowiska do eksperymentu
- **Adrianna Kozierkiewicz (30%)** - pomysłodawca badań, opracowanie definicji ontologii oraz definicji pomocniczych, współopracowanie metody generowania wartości semantyki atrybutów, częściowe przeprowadzenie badań, przeprowadzenie analizy statystycznej wyników eksperymentu
- **Marcin Pietranik (30%)** – studia literaturowe, współopracowanie metody generowania wartości semantyki atrybutów, częściowe przeprowadzenie badań

[P5] Kozierkiewicz A., Pietranik M., 2017, *The knowledge increase estimation framework for ontology integration on the relation level.* Computational Collective Intelligence: 9th International Conference, ICCCI 2017, Nicosia, Cyprus, September 27-29, 2017: proceedings. Pt. 1 / Ngoc Thanh Nguyen [i 5 in.] (Eds.) Cham: Springer, cop. 2017. s. 44-53

Punkty MNiSW (rok 2020, ranking CORE: C): 20 pkt

Ranking Core (rok 2021): B

W pracy przedstawiono metodę oceny przyrostu wiedzy w przypadku integracji ontologii na poziomie relacji. Prace badawcze rozpoczęto od formalnej definicji ontologii oraz przedstawienia szczegółowej charakterystyki relacji ontologii. Na uwagę zasługuje fakt, że w ramach relacji wydzielono szczególny typ relacji, jakim są hierarchie. W pracy przedstawiono najpierw metody integracji relacji i hierarchii, a następnie opracowano metody oceny przyrostu wiedzy dla obu, wyszczególnionych poziomów ontologii. W celu zilustrowania poprawności opracowanych metod zaprezentowano kilka scenariuszy integracji. Opracowane metody są częścią zestawu narzędzi zwanego FOKI (Framework for Ontological Knowledge Integration).

Indywidualny wkład poszczególnych autorów w powstanie publikacji:

- **Adrianna Kozierkiewicz (75%)** - pomysłodawca badań, opracowanie metodologii badań, współopracowanie metody oceny przyrostu wiedzy, współopracowanie definicji pomocniczych, opracowanie scenariuszy użycia, studia literaturowe

- **Marcin Pietranik (25%)** – opracowanie formalnej definicji ontologii, współopracowanie definicji pomocniczych, współopracowanie metody przyrostu wiedzy

[P6] Kozierkiewicz-Hetmańska A., Pietranik M., 2017, *The knowledge increase estimation framework for ontology integration on the concept level*, Journal of Intelligent & Fuzzy Systems. 2017, vol. 32, nr 2, s. 1161-1172.

IF: 1.851
Punkty MNiSW (rok publikacji): 25 pkt

Punkty MNiSW (rok 2020): 70pkt

W pracy przedstawiono metodę oceny przyrostu wiedzy w przypadku integracji ontologii na poziomie konceptów. Prace badawcze rozpoczęto od formalnej definicji ontologii oraz przedstawienia szczegółowej charakterystyki konceptów. Zdefiniowano podstawowe pojęcia, które zostały wykorzystane do opracowania metody oceny przyrostu wiedzy. Pierwsza część badań została poświęcona opracowaniu metody integracji oraz oceny przyrostu wiedzy. W celu zweryfikowania poprawności przyjętych założeń wykonano badania eksperymentalne. Zaproponowaną miarę oceny przyrostu wiedzy zastosowano do zadania integracji wielostopniowej. Początkowo zdefiniowano formalnie ideę integracji wielostopniowej, a następnie wykazano analitycznie, że rezultat integracji jedno- i wielostopniowej jest taki sam. Innymi słowy, sformułowano i udowodniono twierdzenie, że wynik połączenia wielu źródeł wiedzy w jedną spójną wersje jest taki sam, jak wynik integracji wielostopniowej. Badania eksperymentalne były wykonywane na benchmarkowym zbiorze ontologii udostępnianych przez organizację OAEI. W trakcie badań szacowano czas integracji i przyrost wiedzy. Zgodnie z przewidywaniami, czas integracji wielostopniowej jest krótszy niż w przypadku integracji jednostopniowej, ze względu na możliwość zrównoleglenia obliczeń. Przyrost wiedzy wyznaczany w poszczególnych przypadkach testowych pozwolił na wskazanie ontologii najbardziej istotnej w procesie integracji, czyli niosącej najwięcej wiedzy. Ponadto wykazano, że dla testowanych zestawów ontologii nie można wykluczyć kolejnych etapów integrowania, gdyż w każdym z nich przyrost wiedzy był znaczący. Opracowane metody są częścią zestawu narzędzi zwanego FOKI (Framework for Ontological Knowledge Integration).

Indywidualny wkład poszczególnych autorów w postanie publikacji:

- **Adrianna Kozierkiewicz (70%)** - pomysłodawca badań, opracowanie metodologii badań, przygotowanie przypadków testowych do eksperymentu, przeprowadzenie analizy statystycznej wyników eksperymentu, współopracowanie metody oceny przyrostu wiedzy, studia literaturowe, opracowanie wyników eksperymentu, opracowanie koncepcji integracji wielostopniowej, opracowanie twierdzenia i jego matematyczny dowód
- **Marcin Pietranik (30%)** – opracowanie formalnej definicji ontologii, opracowanie definicji pomocniczych, współopracowanie metody przyrostu wiedzy, opracowanie środowiska do eksperymentu

[P7] Kozierkiewicz A., Pietranik M., 2018, *The knowledge increase estimation framework for integration of ontology instances' relations,* Databases and Information Systems: 13th International Baltic Conference, DB&IS 2018, Trakai, Lithuania / Audrone Lupeikiene, Olegas Vasilecas, Gintautas Dzemyda (Eds.). Cham: Springer, cop. 2018. s. 172-186

Punkty MNiSW (rok 2020): 70 pkt

Ranking CORE: B

W pracy przedstawiono metodę oceny przyrostu wiedzy w przypadku integracji ontologii na poziomie relacji instancji. Prace badawcze rozpoczęto od formalnej definicji ontologii oraz przedstawienia szczegółowej charakterystyki relacji instancji. Zdefiniowano podstawowe pojęcia, które zostały wykorzystane do opracowania metody oceny przyrostu wiedzy. Pierwsza część badań została poświęcona opracowaniu metody integracji oraz oceny przyrostu wiedzy. W celu zweryfikowania poprawności przyjętych założeń wykonano badania eksperymentalne. Przygotowano 20 przypadków testowych, które prezentowały proces integracji na poziomie instancji relacji. Następnie poproszono 45 ochotników o subiektywną ocenę przyrostu wiedzy. Badania przeprowadzono w dwóch etapach: w pierwszym zweryfikowano, czy użytkownicy widzą ogólny trend zmiany ilości wiedzy w procesie integracji, dlatego ich zadaniem było ocenić, czy poziom wiedzy wzrósł, zmalał czy nie uległ zmianie. W drugim etapie uczestnicy eksperymentu mieli wskazać o ile zmienił się poziom wiedzy. Do analizy wyników wykorzystano testy statystyczne: współczynnik Kappa oraz współczynnik korelacji wewnątrzklasowej (ICC). W obu przypadkach otrzymano statystycznie istotną zbieżność wyników wskazanych przez użytkowników oraz wartości wyliczonych na podstawie opracowanej metody, a zatem zweryfikowano hipotezę, że zdefiniowana miara oceny przyrostu wiedzy jest intuicyjna oraz koresponduje z oceną wskazaną przez eksperta. Opracowane metody są częścią zestawu narzędzi zwanego FOKI (Framework for Ontological Knowledge Integration).

Indywidualny wkład poszczególnych autorów w powstanie publikacji:

- **Adrianna Kozierkiewicz (75%)** - pomysłodawca badań, opracowanie metodologii badań, przygotowanie przypadków testowych do eksperymentu, przeprowadzenie eksperymentu, przeprowadzenie analizy statystycznej wyników eksperymentu, współopracowanie metody oceny przyrostu wiedzy, studia literaturowe
- **Marcin Pietranik (25%)** – opracowanie formalnej definicji ontologii, opracowanie definicji pomocniczych, współopracowanie metody przyrostu wiedzy

[P8] Kozierkiewicz-Hetmańska A., Pietranik M., Hnatkowska B., 2017, *The Knowledge Increase Estimation Framework for Ontology Integration on the Instance Level,* Intelligent Information and Database Systems: 9th Asian Conference, ACIIDS 2017, Kanazawa, Japan, April 3-5, 2017: proceedings. Pt. 1 / Nguyen Ngoc Thanh, Tojo Satoshi, Nguyen Le Minh, Trawiński Bogdan (Eds.). [Cham]: Springer, cop. 2017. s. 3-12

Punkty MNiSW (rok 2020): 20pkt
Ranking Core (rok 2021): B

W pracy przedstawiono metodę oceny przyrostu wiedzy w przypadku integracji ontologii na poziomie instancji. Prace badawcze rozpoczęto od formalnej definicji ontologii oraz przedstawienia szczegółowej charakterystyki instancji ontologii. Zdefiniowano podstawowe pojęcia, które zostały wykorzystane do opracowania metody oceny przyrostu wiedzy. Następnie opracowano algorytm pozwalający oszacować jak w procesie integracji ontologii zmienia się ilość zawartej w niej wiedzy. Działanie algorytmu zilustrowano kilkoma prostymi przykładami. Zaproponowaną metodę zweryfikowano eksperymentalnie. Przygotowano kilkadziesiąt przypadków testowych, które prezentowały proces integracji dwóch lub trzech ontologii na poziomie instancji. Następnie, poproszono ochotników o subiektywną ocenę przyrostu wiedzy dla zaproponowanych scenariuszy. Otrzymane wyniki zostały poddane analizie statystycznej. W pracy zweryfikowano hipotezę czy zdefiniowana miara oceny przyrostu wiedzy jest intuicyjna i czy koresponduje z oceną wskazaną przez eksperta. Wykorzystano test ICC (współczynnik korelacji wewnątrzklasowej), który potwierdził, że ocena przyrostu wiedzy dla zadanych przypadków testowych wyznaczona za pomocą opracowanej metody jest statystycznie zgodna z oceną wskazaną przez eksperta. Opracowane metody są częścią zestawu narzędzi zwanego FOKI (Framework for Ontological Knowledge Integration).

Indywidualny wkład poszczególnych autorów w powstanie publikacji:

- **Adrianna Kozierkiewicz (60%)** - pomysłodawca badań, współopracowanie metody oceny przyrostu wiedzy, studia literaturowe, współopracowanie metodologii badań, częściowe przygotowanie przypadków testowych do eksperymentu, częściowe przeprowadzenie badań, przeprowadzenie analizy statystycznej wyników eksperymentu
- **Marcin Pietranik (25%)** – opracowanie formalnej definicji ontologii, opracowanie definicji pomocniczych, współopracowanie metody przyrostu wiedzy
- **Bogumiła Hnatkowska (15%)** - współopracowanie metodologii badań, częściowe przygotowanie przypadków testowych do eksperymentu, częściowe przeprowadzenie badań, konsultacje nad zaprojektowaną metodą

[P9] Kozierkiewicz A., Pietranik M., Wesołowski M., 2020, *Assessing ontology mappings on a level of concepts and instances,* IEEE Access 8: 174845-174859

IF: 3.745
Punkty MNiSW (rok 2020): 100pkt

W przypadku integracji dwóch lub więcej ontologii istnieje potrzeba wygenerowania powiązań, które przekształcają jedną ontologię w drugą. Z oczywistych względów, takie odwzorowanie musi zostać wyznaczone dla wszystkich poziomów ontologii: konceptów, instancji i relacji. Istnieją liczne narzędzia do automatycznego generowania mapowań, jednakże jakość ich działania jest różna. Poprawny oraz kompletny zbiór powiązań ma istotny wpływ na wartość przyrostu wiedzy oraz na wynik końcowy integracji. Celem niniejszej pracy było opracowanie metody, która pozwoli ocenić jakość mapowania ontologii. W pierwszym kroku przedstawiono w formalny sposób ontologie oraz definicje pomocnicze. Następnie zaprezentowano cztery metody oceny jakości mapowań wykorzystując informacje o głębokości i ciągłości

mapowanych elementów. Następnie, opracowane metody zostały zweryfikowane w odpowiednio przygotowanym środowisku do eksperymentu. W oparciu o testowe zbiory ontologii dostarczane przez organizacje OAEI oraz wybrane narzędzia do mapowania, wygenerowano odwzorowania, które zostały ocenione za pomocą opracowanych metod. Następnie wyznaczono wartość miary *F*, precyzji i kompletności między wyznaczonym odwzorowaniem, a odwzorowaniem wzorcowym. Analiza statystyczna, za pomocą współczynnika korelacji wewnątrzklasowej, pozwoliła stwierdzić, że istnieje silna korelacja między wartościami opracowanych miar, a wartością kompletności i miarą *F*. Otrzymane wyniki pozwalają jeszcze przed integracją ocenić poprawność mapowań między ontologiami bez potrzeby porównywania ich ze wzorcowym odwzorowaniem.

Indywidualny wkład poszczególnych autorów w powstanie publikacji:

- **Adrianna Kozierkiewicz (33%)** - współopracowanie definicji pomocniczych, współopracowanie metod oceny mapowań, częściowe przeprowadzenie analizy otrzymanych wyników eksperymentu, studia literaturowe
- **Marcin Pietranik (34%)** – pomysłodawca badań, opracowanie formalnej definicji ontologii, współopracowanie definicji pomocniczych, współopracowanie metod oceny mapowań, częściowe przeprowadzenie analizy otrzymanych wyników eksperymentu
- **Mateusz Wesołowski (33%)** – współopracowanie definicji pomocniczych, współopracowanie metod oceny mapowań, częściowe przeprowadzenie analizy otrzymanych wyników eksperymentu, studia literaturowe, opracowanie środowiska do eksperymentu

[P10] Kern R., Kozierkiewicz A., Pietranik M., 2020, *The data richness estimation framework for federated data warehouse integration*, Information Science 513: 397-411
IF: 5.91
Punkty MNiSW (rok 2020): 200pkt

Artykuł poświęcony jest ocenie ekspresywności informacji przechowywanych w hurtowniach danych. Prace badawcze rozpoczęto od formalnej definicji hurtowni danych oraz federacji. Następnie zaproponowano szereg miar, które były podstawą do opracowania algorytmów oceny ekspresywności informacji, w przypadku tworzenia federacji hurtowni danych. Wykazano własności opracowanych miar, a także sformułowano kilka twierdzeń, które zostały udowodnione analitycznie. Następnie przygotowano środowisko do badań eksperymentalnych. W trakcie eksperymentu dołączano kolejne hurtownie do federacji i sprawdzano dokładność odpowiedzi federacji na zestaw zapytań oraz porównywano wyniki otrzymane ze wzorcowej federacji. W badaniach zweryfikowano hipotezę czy dołączanie nowych hurtowni danych do federacji zwiększa ilość poprawnych odpowiedzi na zapytania. Do opracowania wyników badań wykorzystano model wielorakiej regresji liniowej. Analiza wykazała istotną statystycznie zależność liniową między jedną ze zdefiniowanych miar oceny ekspresywności informacji, a ilością prawidłowych odpowiedzi na zapytania. Na podstawie opracowanego modelu regresji liniowej, który wykorzystuje wartości zdefiniowanych miar, możliwe jest oszacowanie efektywności procesu integracji mierzonego ilością poprawnych odpowiedzi na wysyłane do federacji zapytania. W drugiej części badań porównano zaproponowane algorytmy z inną, znaną z literatury, metodą ACCU. Do analizy statystycznej wyników wykorzystano

współczynnik korelacji wewnątrzklasowej, który porównywał spójność i zgodność testowanych miar. Badania wykazały istotną, jednak niezbyt wysoką, statycznie zależność, co jest zgodne z założeniami eksperymentu, gdyż porównujemy dwie różne miary.

Indywidualny wkład poszczególnych autorów w postanie publikacji:

- **Rafał Kern (30%)** - współopracowanie metodologii badań, przygotowanie środowiska do eksperymentu, przeprowadzenie badań eksperymentalnych, konsultacje nad zaprojektowaną metodą
- **Adrianna Kozierkiewicz (40%)** - pomysłodawca badań, konsultacje nad zaprojektowanymi algorytmami, opracowanie twierdzeń i dowodów, współopracowanie metodologii badań, przeprowadzenie analizy statystycznej wyników eksperymentu, studia literaturowe
- **Marcin Pietranik (30%)** – opracowanie formalnych definicji, opracowanie metody oceny ekspresywności informacji

Prace badawcze zostały podzielone na 3 etapy, które zostały graficznie przedstawione na schemacie na Rys. 1:

- ocena efektywności procesu integracji dla prostych struktur wiedzy: [P1], [P2], [P3]
- ocena przyrostu wiedzy pełniąca funkcje oceny efektywności procesu integracji ontologii: [P4], [P5], [P6], [P7], [P8], [P9]
- ocena ekspresywności informacji pełniąca funkcje oceny efektywności procesu integracji cząstkowych hurtowni w celu utworzenia federacji hurtowni danych: [P10]

Dodatkowe prace związane z cyklem, ale nie wchodzące w jego skład:

- [D1] Kozierkiewicz-Hetmańska A., Nguyen, N.T., 2012, *A comparison analysis of consensus determining using one and two-level methods*, [w] Advances in knowledge-based and intelligent information and engineering systems, Manuel Graña [i in.] (Eds.) Amsterdam, IOS Press, cop. 2012. s. 159-168. **(MNSiW (rok publikacji): 10pkt, szacowany udział: 80%)**
- [D2] Hernes M., Kozierkiewicz A., Nguyen T.T., 2019, *An application a two-level determination consensus method in a multi-agent financial decisions support system.* Intelligent Information and Database Systems: 11th Asian Conference, ACIIDS 2019, Yogyakarta, Indonesia, April 8-11, 2019: proceedings. Pt. 1 / Ngoc Thanh Nguyen [i in.] (Eds.). Cham : Springer, cop.[2019. s. 453-463. **(MNiSW (rok 2020): 20pkt, ranking Core (rok 2021): B, szacowany udział: 40%)**
- [D3] Kozierkiewicz A., Sitarczyk M., 2017, *The efficiency analysis of the multi-level consensus determination method.,* Computational Collective Intelligence: 9th International Conference, ICCCI 2017, Nicosia, Cyprus, September 27-29, 2017: proceedings. Pt. 1 / Ngoc Thanh Nguyen [i 5 in.] (Eds.) Cham : Springer, cop. 2017. s.103-112, **(MNiSW (rok 2020, ranking CORE: C): 20pkt, ranking Core (rok 2021): B, szacowany udział: 50%)**
- [D4] Kozierkiewicz-Hetmańska A., Pietranik M., 2017, *Assessing the quality of a consensus determined using a multi-level approach,* 2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA 2017)

Poland / Eds. Piotr Jędrzejowicz, Tülay Yildirim, Ireneusz Czarnowski. Piscataway, NJ: IEEE, cop. 2017. s. 131-136., **MNiSW (rok 2020): 20pkt, ranking CORE: C, szacowany udział: 50%)**

- [D5] Kozierkiewicz A., Pietranik M., Nguyen V.D., 2018, *The assessing of influence of collective intelligence on the final consensus quality.*, Intelligent Information and Database Systems : 10th Asian Conference, ACIIDS 2018, Dong Hoi City, Vietnam, March 19-21, 2018 : proceedings. Pt. 1 / Ngoc Thanh Nguyen [i in.] (Eds.). Cham: Springer, cop. 2018. s. 15-24. **(MNiSW (rok 2020): 20pkt, ranking Core (rok 2021): B, szacowany udział: 40%)**
- [D6] Kozierkiewicz A., Pietranik M., Sitarczyk M., 2019, *Updating consensus in case of adding or removing new elements to the profile*, Journal of Intelligent & Fuzzy Systems: 7291-7302, **(IF: 1.851, MNiSW (rok 2020): 70pkt, szacowany udział: 33%)**

Rys. 1 Schemat publikacji powiązanych tematycznie wchodzących w skład cyklu ([P1]-[P10]) oraz najważniejszych prac, z którymi związany jest cykl ([D1]-[D6])

4.3 Omówienie osiągnięć, o których mowa w art. 219 ust. 1 pkt. 2 ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (Dz. U. z 2020 r. poz. 85 z późn. zm.).

4.3.1 Wstęp i cel przeprowadzonych badań

Blisko 400 lat temu angielski filozof Francis Bacon sformułował swoją słynną sentencję: „*Scientia potentia est*" czyli „Wiedza to potęga". We współczesnym świecie słowa XVII wiecznego myśliciela są bardzo aktualne. Z roku na rok ilość produkowanych przez ludzkość danych gwałtownie wzrasta. W raporcie "Data Never Sleeps" opublikowanym przez firmę Dumo wskazano, że urządzenia informatyczne generują 2,5 kwintyliona bajtów danych dziennie. Są to dane, które nie tylko są rozproszone, ale często również niepełne i niespójne. W wyniku odpowiedniego przetwarzania oraz interpretowania danych otrzymujemy informacje, które z kolei stają się elementem składowym wiedzy wykorzystywanej do podejmowania różnorakich decyzji zarówno gospodarczych, społecznych, politycznych czy zwykłych życiowych decyzji podejmowanych przez przeciętnego człowieka, np. wybór produktów do nabycia, wybór miejsca zamieszkania itp. Wszystkie systemy wspomagania decyzji czy systemy wieloagentowe działają w oparciu o wiedzę ekspercką pochodzącą z Internetu czy innych autonomicznych źródeł. Zatem wspomniane systemy przetwarzające wiedzę, wymagają ich formalnej reprezentacji. W niniejszej pracy przyjęto, że wiedza będzie reprezentowana przez proste struktury takie jak: wektory binarne, podziały i pokrycia uporządkowane oraz zbiory z powtórzeniami, które stały się punktem wyjścia do badań nad złożonymi struktury jakimi są ontologie i federacje hurtowni danych. W wielu przypadkach wiedza ekspercka nie wymaga reprezentacji skomplikowanych związków semantycznych i wybór prostych reprezentacji wiedzy jest wystarczający. Przykładem mogą być systemy, które reprezentują wiedzę o możliwych odczynach popromiennych w radioterapii w oparciu o normę CTCAE[1]. Jednak już system terminologii klinicznej, który zawiera zależności między kodami pojęciowymi mającymi powiązania znaczeniowe, wymaga złożonej reprezentacji za pomocą ontologii SNOMED CT[2]. W związku z tym przyjęte w niniejszych badaniach reprezentacje wiedzy pozwalają na tworzenie zarówno uproszczonych, jak i złożonych opisów świata lub jego stanów.

Jak już wspomniano w powyższym akapicie, wiedza ekspercka bardzo często pochodzi z różnych autonomicznych źródeł. Jednak ich duża ilość nie zawsze jest łatwa do przetworzenia w akceptowanym przez użytkowników czasie. Zatem integracja informacji i wiedzy może być kosztowna i czasochłonna. W szczególności rozproszenie źródeł sprawia, że połączenie ich w jedną, spójną wersję może okazać się podatne na błędy, ze względu na potrzebę przesłania ich w jedno miejsce. Wszystkie wymienione czynniki stały się motywacją do podjęcia badań, w wyniku których znaleziono odpowiedzi na poniższe pytania badawcze:

Q1: *Czy analiza źródeł wejściowych pozwala na ocenę możliwości wyznaczenia wiarygodnego i spójnego wyniku ich integracji?*

Q2: *Czy na podstawie analizy źródeł wejściowych możliwa jest ocena jakości wyniku ich integracji?*

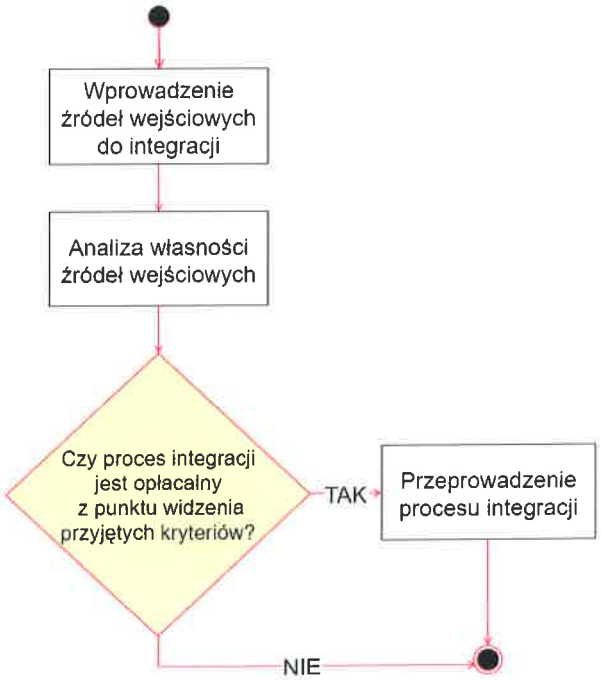
---

[1] Norma CTCAE:
ttps://ctep.cancer.gov/protocoldevelopment/electronic_applications/docs/ctcae_v5_quick_reference_5x7.pdf
[2]SNOMET CT: https://www.snomed.org/

Q3: *Czy możliwa jest ocena efektywności procesu integracji informacji i wiedzy pochodzących z różnych źródeł przed lub w trakcie procesu ich łączenia w jedną, spójną wersję?*

Celem monotematycznego cyklu publikacji było ***opracowanie metod, które dla wybranych reprezentacji wiedzy umożliwią ocenę efektywności procesu integracji jeszcze przed lub w trakcie procesu ich łączenia.***

Ogólny cel pracy został przedstawiony poglądowo na Rys. 2.



Rys. 2 Proces podejmowania decyzji o przeprowadzeniu integracji informacji i wiedzy

Powyższy cel został osiągnięty poprzez realizację mniejszych problemów badawczych, które pozwoliły na:

- wyznaczenie kryteriów, jakie muszą spełniać źródła wejściowe, aby być podatne na consensus, sformułowane kryteria są warunkami wystarczającymi podatności źródeł wejściowych na consensus,
- wyznaczenie kryteriów, które pozwalają na stwierdzenie, jakie własności muszą spełniać źródła wejściowe, aby wynik ich integracji charakteryzował się najwyższą lub zadaną jakością (według przyjętej w cyklu definicji), sformułowane kryteria są warunkami wystarczającymi maksymalizacji jakości wyniku integracji źródeł wejściowych,
- opracowanie zestawu metod do integracji i oceny efektywności integracji ontologii, które umożliwiają ocenę potencjalnego przyrostu wiedzy w procesie integracji ontologii na poziomie: konceptów, relacji, instancji, relacji,
- opracowanie metod wyznaczania wartości semantyki atrybutów,
- opracowanie metod oceny odwzorowań ontologii bez konieczności porównywania go z odwzorowaniem wzorcowym,
- opracowanie metod oceny ekspresywności informacji w federacyjnych hurtowniach danych.

Do zadania integracji informacji i wiedzy wykorzystana została teoria consensusu, która pozwala na formalną i matematyczną analizę jej efektów.

4.3.2 Przegląd istniejących rozwiązań

Jak wspomniano w poprzednim podrozdziale, z integracją informacji i wiedzy wiążą się liczne ograniczenia i problemy, w szczególności natury ekonomicznej. W związku z tym pojawia się zasadne pytanie, kiedy opłaca się integrować informacje i wiedzę pochodzące z różnych autonomicznych źródeł? Analizy, przeprowadzone przed przystąpieniem do prac nad cyklem publikacyjnym wykazały, że na powyższe pytanie brak jest jednoznacznej odpowiedzi, co stało się motywacją do podjęcia badań nad wspomnianym zagadnieniem. W dalszej części zostaną omówione najważniejsze osiągnięcia w kontekście postawionych pytań badawczych.

Q1: *Czy analiza źródeł wejściowych pozwala na ocenę możliwości wyznaczenia wiarygodnego i spójnego wyniku ich integracji?*

Przed rozpoczęciem integracji informacji i wiedzy, należy zweryfikować, czy taka operacja jest możliwa, a otrzymany wynik będzie wiarygodny i spójny. Z zadaniem integracji informacji i wiedzy wiąże się często problem konfliktu ich struktur, np. kiedy ten sam parametr (np. czas) w różnych źródłach posiada różne jednostki miary. Od ponad dwóch dekad powstaje wiele narzędzi umożliwiających efektywne mapowanie źródeł, standaryzację nazw, konwersję typów danych, eliminacje duplikatów [1], [2], które są w stanie rozwiązać istniejący problem. Druga sytuacja, w której utrudnione lub wręcz niemożliwe jest wyznaczenie jednej, spójnej wersji po integracji, wiąże się z konfliktem semantyki, np. kiedy różne źródła przechowują niespójne lub sprzeczne ze sobą informacje. Istnieją publikacje, które rozwiązują problem konfliktu wyłącznie na poziomie danych [3]. Są to m.in. strategie, które rozważają dokładność, aktualność źródeł oraz wewnętrzne zależności między nimi.

Dotychczas prowadzone prace badawcze skupiały się co najwyżej na ocenie jakości źródeł wejściowych. Przykładem takich miar są m.in. kompletność, który ocenia stopień odwzorowania faktów w systemie informatycznych, unikatowość, która odzwierciedla stopień konfliktu danych, spójność, która wyraża integralność danych względem przyjętych kryteriów, aktualność oraz dokładność, która wiąże się z poprawnością oraz precyzją danych względem świata rzeczywistego [7], [8], [9]. Mimo, że zaprezentowane w pracy [9] badania wykazały korelacje między jakością danych, a jakością podejmowanych na ich podstawie decyzji, to nie znaleziono prac, które odpowiadają na postawione pytanie badawcze.

Inne podejście zaproponowano w pracach [4], [5], [6], w którym teoria consensusu została wykorzystana do przetwarzania niespójności integrowanych źródeł. Jednakże, nie w każdej sytuacji możliwe jest otrzymanie wiarygodnego i spójnego wyniku integracji źródeł wejściowych. Jako prosty przykład rozważmy opinie użytkowników na temat sprzętu komputerowego. Cztery osoby zostały poproszone o wyrażenie swojego zdania czy wybrany model laptopa jest warty zakupu. Jeśli w wyniku ankiety otrzymamy dwie odpowiedzi „TAK" i dwie „NIE", to na podstawie zgromadzonej wiedzy nie jesteśmy w stanie podjąć decyzji o ewentualnym zakupie sprzętu komputerowego.

Taka wiedza nazywana jest niepodatną na consensus. Formalne kryterium podatności profilu na consensus zostało zdefiniowane w pracy [4] i stało się podstawą do dalszych analiz. W pracach [4], [5], [6], autor zdefiniował kryteria podatności profilu na consensus oraz wykazał

proste związki między kryterium podatności profilu na consensus, a jego spójnością (według przyjętych w pracy definicji spójności).

Prace przedstawione w ramach cyklu publikacyjnego są uzupełnieniem i rozwinięciem prac [4]. Dzięki opracowanym kryteriom możliwa jest ocena, czy na podstawie źródeł wejściowych otrzymany jeden, spójny wynik integracji.

Q2: *Czy na podstawie analizy źródeł wejściowych możliwa jest ocena jakości wyniku ich integracji?*

Jak już wspomniano powyżej, dotychczas prace badawcze skupiały się głównie na ocenie jakości danych wejściowych. Do tej grupy należą między innymi miary takie jak: kompletność, aktualność, dokładność [7], [8], [9]. W pracy [10] autorzy zdefiniowali własności, które pozwalają na zapewnienie jakości danych. Są to między innymi: duplikacja, standaryzacja, integralność, dokładność oraz interpretowalność. W pracy [11] zaproponowano model nazwany OntoQa, który zawiera szereg miar (zarówno na poziomie schematu jak i instancji) do oceny ontologii. Jednak zasadniczą wadą zaproponowanego rozwiązania jest jego prostota i brak całościowej oceny procesu integracji. Podobnie rozwiązania można znaleźć w przypadku danych przechowywanych w bazach oraz hurtowniach danych. W [12] autorzy wspomnieli o ośmiu różnych czynnikach wpływających na jakość danych, które powinny być brane pod uwagę przy szacowaniu przydatności danych do dalszej analizy oraz w procesie podejmowania decyzji. W pracy [13] autorzy opracowali 23 różne metryki jakości. Natomiast praca [14] koncentruje się na metryce, która ocenia spójność semantyczną danych. Jednakże żadna z wymienionych prac nie odpowiada na postawione pytanie badawcze.

Nieco inne podejście zaproponowano w pracy [4]. Autor opracował model integracji wiedzy w oparciu o teorię consensusu. Zdefiniował w formalny sposób jakość integracji, która stała się punktem wyjścia do dalszych analiz w ramach niniejszego cyklu publikacji. W [4] przedstawiono dowody matematyczne, które pozwoliły na sformułowanie zależności między spójnością źródeł wejściowych, a wynikiem ich integracji. Inną definicję jakości zdefiniowano w pracy [15], gdzie wynik integracji porównywany był z rzeczywistą wartością. Badania statystyczne wykazały, że zwiększenie ilości źródeł wejściowych poprawia jakość decyzji końcowej. W [16] testowano wpływ spójności źródeł wejściowych (według zdefiniowanych miar) na jakość wyniku integracji.

Niniejszy cykl publikacji wypełnia istniejącą lukę w badaniach dotyczących dziedziny problemu. Sformułowano i udowodniono matematycznie twierdzenia, które umożliwiają ocenę jakości wyniku integracji wejściowych źródeł wiedzy wyłącznie na podstawie analizy własności tychże źródeł.

Q3: *Czy możliwa jest ocena efektywności procesu integracji informacji i wiedzy pochodzących z różnych źródeł przed lub w trakcie procesu ich łączenia w jedną, spójną wersję?*

Ostatnie pytanie badawcze odnosi się do metod, które pozwalają na ocenę efektywności samego procesu integracji, która będzie rozumiana jako stopień spełnienia pewnej zdefiniowanej miary. W literaturze najczęściej można znaleźć ogólne miary takie jak: [4] kompletność (ile wiedzy zostało utracone w wyniku integracji), precyzja (estymuje, ile elementów zostało zduplikowanych oraz ile jest kompletnie nowych), optymalność (jaka jest różnica względem przyjętej metryki odległości między wynikiem integracji, a źródłami wejściowymi). Integracja ontologii oceniana jest zazwyczaj w kontekście relatywnego pokrycia oraz kompaktowości

otrzymanego rezultatu mierzona ilością wprowadzonej redundancji [17]. Podobne podejście zaprezentowano w pracy [18], w którym efektywność integracji ontologii estymowana jest na podstawie minimalności oraz kompletności integrowanych schematów. W przypadku ontologii istnieje zbiór benchmarkowych źródeł do integracji wraz z odpowiadającym im odwzorowaniem dostarczanych przez OAEI - Ontology Aligment Evaluation Initiative [19]. Pozwala on na rzetelne porównywanie między sobą różnych systemów oraz metod integracji ontologii na tych samych zbiorach wejściowych.

W przypadku baz i hurtowni danych wykorzystywane są również typowe dla systemów integracji miary takie jak: minimalność [20], dokładność oraz aktualność [21]. Praca [22] porusza kwestię integracji hurtowni danych, która powinna spełniać własność spójności, interpretowanej jako brak sprzeczności pomiędzy relacjami, zgodności rozumianej jako poprawność hierarchii między wymiarami, a także rzetelności weryfikowanej przez przynależność do wymiarów. Jednakże, wszystkie opisane miary posiadają istotną wadę- oceny można dokonać dopiero po przeprowadzeniu procesu integracji. Wiąże się to z poniesionym kosztem przeprowadzonej integracji, która nie przyniosła oczekiwanych rezultatów końcowych. Jedynie w pracy [23] znaleziono wnioski i rezultaty najbliższe tematyce, której poświęcony jest niniejszy cykl publikacji. Artykuł prezentuje algorytmy, których celem jest wybór źródeł do integracji w celu znalezienia balansu między jakością otrzymanego rezultatu, a kosztami poniesionej operacji. Jednakże problem oceny efektywności procesu integracji informacji i wiedzy pochodzących z różnych źródeł przed lub w trakcie procesu ich łączenia nie został szeroko zbadany. Brak w literaturze odpowiedzi na zadane pytanie badawcze stał się inspiracją do poniższych badań.

4.3.3 Rozwiązanie problemu

A) Proste struktury

Jak już wspomniano, proces integracji informacji i wiedzy potrafi być czasochłonny i kosztowny szczególnie, gdy przetwarzany jest duży zbiór źródeł wejściowych. Niestety, sam proces integracji może również wiązać się z naruszeniem bezpieczeństwa tychże zbiorów informacji i wiedzy, jeśli są one rozproszone i wyznaczenie jednej wersji wiąże się z ich przesyłaniem w jedno miejsce. Dlatego też ważne jest, aby jeszcze przed procesem złączenia źródeł wejściowych ocenić, czy możliwe jest wyznaczenie jednej wiarygodnej wersji i czy jej jakość będzie dla nas zadowalająca w kontekście przyjętych przez nas kryteriów.

Do zadania integracji informacji i wiedzy wykorzystano teorię consensusu, która pozwala na formalną analizę jego własności. Ogólny problem wyboru consensusu można przedstawić następująco: *dane jest uniwersum obiektów i pewien jego podzbiór z powtórzeniami. Należy znaleźć obiekt należący do uniwersum, który najlepiej reprezentuje ten podzbiór*. Formalnie, przez U oznaczono niepusty, skończony zbiór uniwersum obiektów. Każdy taki obiekt może odzwierciedlać fragment wiedzy opisujący rzeczywiste zdarzenie ze świata. Symbol $\Pi_b(U)$ reprezentuje zbiór wszystkich b-elementowych podzbiorów zbioru U z powtórzeniami, $b \in \mathbb{N}$, $\Pi(U) = \bigcup_{b \in N} \Pi_b(U)$. Każdy element X należący do zbioru $\Pi(U)$ nazywany jest profilem wiedzy (ang. knowledge profile). Problem wyboru consensusu sprowadza się do wyznaczenia takiego x z uniwersum U, który jest consensusem dla profilu X. Wyznaczenie consensusu możliwe jest wtedy, kiedy znana jest struktura uniwersum. Składają się na nią dwa elementy: mikro- i makrostruktura. Mikrostruktura określa reprezentację elementów należących do

uniwersum. W pierwszym etapie badań założono, że będą to zbiory z powtórzeniami, wektory binarne, pokrycia oraz podziały uporządkowane. W dalszej części pracy rozważano bardziej złożone struktury typu ontologie. Makrostruktura określa relację między elementami uniwersum i najczęściej definiowana jest jako funkcja:

Definicja 1. *Makrostruktura uniwersum U definiowana jest jako funkcja:*

$$\delta: U \times U \rightarrow [0,1] \tag{1}$$

która spełnia poniższe kryteria:

- $\forall_{u,v \in U}\, \delta(u,v) \geq 0$

-$\forall_{u,v \in U}\, \delta(u,v) = 0 \iff u = v$

- $\forall_{u,v \in U}\, \delta(u,v) = \delta(v,u)$.

W niniejszej pracy zakładamy, że δ nie musi spełniać nierówności trójkąta, gdyż dla wielu rzeczywistych rozwiązań jest to zbyt silny warunek. W związku z powyższym, funkcja δ nie musi być metryką, dlatego para (U,δ) będzie nazywana przestrzenią z odległością (a nie przestrzenią metryczną).

Dla zdefiniowanej przestrzeni z odległością (U,δ), możliwe jest wyznaczanie funkcji wyboru consensusu w następujący sposób:

Definicja 2. *Funkcją wyboru consensusu w przestrzeni (U,δ) nazywana jest dowolna funkcja postaci:*

$$Con: \Pi(U) \rightarrow 2^U \tag{2}$$

Przez $c(X)$ oznaczono reprezentację profilu X, a każdy element $x \in c(X)$ nazywany jest consensusem profilu X. W niniejszym monotematycznym cyklu publikacji profil rozumiany jest jako zbiór źródeł wejściowych, które integrowane są w jedną wersję nazywaną consensusem.

Istotny jest fakt, że wyznaczony consensus powinien spełniać pewne kryteria, np. niezawodności, uproszczenia, spójności, wiarygodności itp. W niniejszych pracach rozważano dwa kryteria: 1- i 2-optymalności. Pierwszy postulat pozwala wyznaczyć consensus, który jest możliwie blisko (według przyjętej funkcji odległości) od pozostałych elementów profilu. Consensus 1-optymalny jest najlepszym reprezentantem integrowanej wiedzy. Kryterium 2-optymalności pozwala wyznaczyć najbardziej „sprawiedliwy" consensus, którego kwadrat odległości od elementów profilu jest najmniejszy.

Formalnie, kryteria optymalności zostały zdefiniowane w następujący sposób:

Definicja 3. *Funkcja wyboru consensusu Con dla profilu $X \in \Pi(U)$ i przestrzeni (U,δ) spełnia postulaty:*

- *1 - optymalności wtedy i tylko wtedy gdy:* $(x \in c(X) \Rightarrow (\delta^1(x,X) = \min_{y \in U} \delta^1(y,X))$

- *2 - optymalności wtedy i tylko wtedy gdy:* $(x \in c(X) \Rightarrow (\delta^2(x,X) = \min_{y \in U} \delta^2(y,X))$

gdzie: $\delta^1(x,X) = \sum_{y \in X} \delta(x,y)$, $\delta^2(x,X) = \sum_{y \in X} (\delta(x,y))^2$.

W pierwszej części badań [P2] **dla przyjętych reprezentacji wiedzy (zbiorów z powtórzeniami, wektorów binarnych, pokryć oraz podziałów uporządkowanych) wyznaczono kryteria jakie muszą spełniać źródła wejściowe, aby być podatne na consensus. Innymi słowy, czy wynik integracji źródeł wejściowych jest wiarygodny i spójny.** Sformułowane kryteria są warunkami wystarczającymi podatności źródeł wejściowych na consensus.

Przeprowadzone analizy opierały się na formalnej definicji podatności profilu consensus:

Definicja 4. *Profil* $X \in \Pi(U)$ *jest podatny na consensus wtedy i tylko wtedy gdy:*

$$\hat{\delta}(X) \geq \hat{\delta}_{min}(X) \tag{3}$$

gdzie: $\hat{\delta}(X) = \frac{\sum_{x,y \in X} \delta^i(x,y)}{n(n+1)}, \hat{\delta}_x(X) = \frac{\sum_{y \in X} \delta^i(x,y)}{n}, \hat{\delta}_{min}(X) = \min_{x \in U} \hat{\delta}_x(X)$, *n=card(X), i={1,2}.*

W sytuacji, kiedy mikrostrukturą są zbiory z powtórzeniami (multi-zbiory), udowodniono [4], że jeśli liczba źródeł do integracji jest nieparzysta, to jest to warunek wystarczający podatności profilu na consensus. W przypadku parzystej liczebności profilu, sformułowano poniższe twierdzenia:

Twierdzenie 1. (Tw. 1 w pracy [P2]) *Dla przyjętej przestrzeni z odległością: U={0; 0,01; 0,02,...,5,99, 6,00,...,f},* $\delta(u,v) = \frac{1}{f}|u - v|$, $u, v \in U$, *profil* $X = \{a_1, a_2, \dots, a_n\} \in \Pi(U)$, *(gdzie n jest liczbą parzystą), który spełnia postulat 1-optymalności jest podatny na consensus, jeśli:*

$$x_{\frac{n}{2}} = x_{\frac{n+2}{2}} \tag{4}$$

gdzie: $x_1, \dots, x_{\frac{n}{2}}, x_{\frac{n+2}{2}}, \dots, x_n$ *są elementami profilu X uporządkowanymi w porządku niemalejącym.*

Twierdzenie 2. (Tw. 2 w pracy [P2]) *Dla przyjętej przestrzeni z odległością: U={0; 0,01; 0,02,...,5,99, 6,00,...,f},* $\delta(u,v) = \frac{1}{f}|u - v|$, $u, v \in U$, *profil* $X = \{a_1, a_2, \dots, a_n\} \in \Pi(U)$, *(gdzie n jest liczbą parzystą), który spełnia postulat 2-optymalności) jest podatny na consensus jeśli:*

1. $n = 1$ *lub*
2. $n \geq \frac{\left(\sum_{i=1}^{n} a_i\right)^2}{\sum_{i=1}^{n} {a_i}^2}$.

Popularną i wygodną reprezentacją wiedzy są wektory binarne, które pozwalają modelować większość procesów, które można zaobserwować w życiu i otaczającym nas świecie. Poniższe twierdzenie zawiera zestaw kryteriów, jakie muszą spełniać źródła wejściowe wyrażone w postaci wektorów binarnych, aby być podatne na consensus:

Twierdzenie 3. (Tw. 3 w pracy [P2]) *Dla przyjętej przestrzeni z odległością* $U = \{u_1, u_2, \dots\}$, *gdzie:* $u = (u^1, u^2, \dots, u^m)$, $u \in U$, $u^q \in \{0,1\}$, $q \in \{1, \dots, m\}$ *oraz* $\delta(u,v) = \sum_{i=1}^{m} |u^i -$

$v^i|$, $u, v \in U$, *profil* $X = \{a_1, a_2, \ldots, a_n\} \in \Pi(U)$ (*gdzie:* $a_i = (a_{i1}, a_{i2}, \ldots a_{im}), i \in \{1, \ldots n\}$, n *– liczebność profilu* X*), który spełnia postulat 1-optymalności jest podatny na consensus jeśli:*

1. *n jest liczbą nieparzystą lub*
2. *jeśli n jest liczbą parzystą oraz spełniony jest warunek:* $\underset{j=1,\ldots,m}{\forall} \sum_{i=1}^{n} a_{ij} \neq \frac{n}{2}$.

Podziały uporządkowane [24] to kolejna reprezentacja wiedzy, która pozwala na modelowanie zdarzeń i procesów wymagających podziału obiektów na klasy według przyjętych własności np. podział pacjentów chorych na gruźlicę względem zdjęć rentgenowskich płuc. Formalną definicję podziałów uporządkowanych przedstawiono w pracy [P2] (definicja nr 5, 6 i 7). Przez $\pi_k(U)$ oznaczono zbiór wszystkich k-klasowych podziałów zbioru $Z = \{z_1, z_2, \ldots, z_m\}$, a przez ξ funkcję odległości między dwoma podziałami $P, S \in \pi_k(U)$. Poniższe twierdzenie prezentuje kryteria, jakie muszą być spełnione, aby profil wyrażony w formie podziałów uporządkowanych, był podatny na consensus:

Twierdzenie 4. (Tw. 4 w pracy [P2]) *Dla przyjętej przestrzeni z odległością* $(\pi_k(U), \xi)$*, profil* $X = \{P^{(1)}, P^{(2)}, \ldots, P^{(n)}\} \in \Pi(\pi_k(U)$*, który spełnia postulat 1-optymalności jest podatny na consensus dla każdego* $j \in \{1,2,\ldots,m\}$ *jeśli:*

$$\exists \; f_{max}^{j}: \underset{i=1,\ldots,k}{\forall} f_{max}^{j} > f_i^{j} \tag{5}$$

gdzie: $f(u_j) =< f_1^j, f_2^j, \ldots, f_k^j >$, $f_i^j = \sum_{l=1}^{n} mp_{ij}^{(l)}$, $i \in \{1,2,\ldots,k\}$, $j \in \{1,2,\ldots,m\}$ *jest wektorem wystąpień elementu* u_j *w podziałach* $P^{(1)}, P^{(2)}, \ldots, P^{(n)}$*, n-liczebność profilu.*

Nie każdy obiekt może zostać zawsze przypisany tylko do jednej klasy na podstawie opisu cech. Przykładowo pacjenci z różnymi objawami mogą cierpieć na kilka schorzeń lub żadne. Zamodelowanie takich zjawisk możliwe jest dzięki pokryciom uporządkowanym, które formalnie zostały przedstawione w pracy [P2] (definicja nr 8, 9 i 10). Jeśli zbiór wszystkich pokryć uporządkowanych zbioru $Z = \{z_1, z_2, \ldots, z_m\}$ oznaczymy symbolem $\omega_k(U)$, a przez ϕ funkcję odległości między dwoma pokryciami $G, J \in \omega_k(U)$, to kryterium podatności na consensus przedstawiono w poniższym twierdzeniu:

Twierdzenie 5. (Tw. 5 w pracy [P2]) *Dla przyjętej przestrzeni z odległością* $(\omega_k(U), \phi)$ *profil* $X = \{G^{(1)}, \ldots, G^{(n)}\} \in \Pi(\omega_k(U))$*, który spełnia postulat 1-ptymalności jest podatny na consensus jeśli:*

1. *n jest liczbą nieparzystą lub*
2. *jeśli n jest liczbą parzystą oraz spełniony jest warunek:* $\underset{i=1,\ldots k}{\forall} \underset{j=1,\ldots m}{\forall} \sum_{l=1}^{n} g_{ij}^{l} \neq \frac{n}{2}$.

gdzie: n- liczebność profilu, $G^l = [g_{ij}^l]$*- macierz charakterystyczna pokrycia uporządkowanego* G^l*,* $g_{ij} = \begin{cases} 1 \; jeśli \; u_j \in G_i \\ 0 \; jesli \; u_j \notin G_i \end{cases}$.

Dowody twierdzeń 1-5 można znaleźć w pracy [P2]. W powyższych twierdzeniach (1-5) znajdują się kryteria, które poprzez analizę źródeł wejściowych umożliwiają ocenę, czy wynik ich integracji będzie wiarygodny i spójny.

Druga część badań poświęcona została **kryteriom, jakie muszą spełniać źródła wejściowe, aby wynik ich integracji miał maksymalną [P1] lub zadaną (przez użytkownika) jakość [P3].**

Jakość consensusu formalnie definiowana jest w następujący sposób:

Definicja 5. *Jakość consensusu* $x \in c(X)$ *profilu* $X \in \Pi(U)$ *wyznaczana jest za pomocą poniższej formuły:*

$$Q^i(x,X) = 1 - \frac{\delta^i(x,X)}{card(X)} \tag{6}$$

gdzie: $i \in \{1,2\}$.

W niniejszym cyklu publikacyjnym sformułowano kilka twierdzeń, które są wskazówką jak dobrać źródła wejściowe, aby wynik ich integracji cechował się maksymalną (lub założoną) jakością. W [4] udowodniono, że consensus spełniający postulat 1-optymalności jest najwyższej, możliwej jakości według definicji 5. W przypadku consensusu 2- optymalnego, pomocne są poniższe twierdzenia:

Twierdzenie 6. (Tw. 1 w pracy [P1]) *Dla przyjętej przestrzeni z odległością: U={0; 0,01; 0,02,...,5,99, 6,00,...,f},* $\delta(u,v) = \frac{1}{f}|u-v|$, $u,v \in U$ *oraz profilu* $X = \{a_1, a_2, \dots, a_n\} \in \Pi(U)$ *(gdzie n jest liczbą parzystą), consensus, który spełnia postulat 2-optymalności osiąga maksymalną jakość, jeśli spełniony jest poniższy warunek:*

$$\lim_{n\to\infty} \left| n - \frac{\left(\sum_{i=1}^n a_i\right)^2}{\sum_{i=1}^n a_i^2} \right| = 0 \tag{7}$$

Twierdzenie 7. (Tw. 7 i 8 w pracy [P1]) *Dla przyjętej przestrzeni z odległością* $U = \{u_1, u_2, \dots\}$, *gdzie:* $u = (u^1, u^2, \dots, u^m)$, $u \in U$, $u^q \in \{0,1\}$, $q \in \{1,\dots,m\}$, $\delta(u,v) = \sum_{i=1}^m |u^i - v^i|$, $u,v \in U$ *oraz profilu* $X = \{a_1, a_2, \dots, a_n\} \in \Pi(U)$ *(gdzie:* $a_i = (a_{i1}, a_{i2}, \dots a_{im})$, $i \in \{1,\dots n\}$, *n – liczebność profilu X) consensus, który spełnia postulat 2-optymalności ma jakość większą lub równą niż* $Q^2(x,X) \geq 1 - \frac{\left(\sum_{j\in W}\sum_{i=1}^n a_{ij} + \sum_{j\in E}\left(n - \sum_{i=1}^n a_{ij}\right)\right)^2}{m^2 n}$ *oraz osiąga maksymalną jakość, jeśli spełniony jest poniższy warunek:*

$$\lim_{n\to\infty} \left| n - \frac{\sum_{j\in E}\sum_{i=1}^n a_{ij} - \sum_{j\in W}\sum_{i=1}^n a_{ij}}{card(E)} \right| = 0 \tag{8}$$

gdzie: $E = \{j: \sum_{i=1}^n a_{ij} \geq \frac{n}{2}\}$, $W = \{j: \sum_{i=1}^n a_{ij} < \frac{n}{2}\}$, *n- liczebność profilu X.*

Sformułowane w Twierdzeniach 7 i 8 kryteria są warunkami wystarczającymi maksymalizacji jakości wyniku integracji źródeł wejściowych. Kolejne twierdzenia prezentują jak różne własności profilu wpływają na jakość końcową wyznaczonego consensusu:

Twierdzenie 9. (Tw. 1 w pracy [P3]) *Dla przyjętej przestrzeni z odległością* $U = \{u_1, u_2, \dots\}$, *gdzie:* $u = (u^1, u^2, \dots, u^m)$, $u \in U$, $u^q \in \{0,1\}$, $q \in \{1, \dots, m\}$, $\delta(u,v) = \sum_{i=1}^{m} |u^i - v^i|$, $u, v \in U$ *oraz profilu* $X = \{a_1, a_2, \dots, a_n\} \in \Pi(U)$ (*gdzie:* $a_i = (a_{i1}, a_{i2}, \dots a_{im})$, $i \in \{1, \dots n\}$, n – *liczebność profilu X*), *aby zapewnić wyznaczonemu consensusowi założoną jakość:* $Q^1(x,X) \geq Q_f \in \left[\frac{1}{2}, 1\right]$ *poniższy warunek musi być spełniony:*

$$p \approx \frac{1-\sqrt{2*Q_f-1}}{2} \tag{9}$$

lub

$$p \approx \frac{1+\sqrt{2*Q_f-1}}{2} \tag{10}$$

gdzie: $\frac{p}{q}$ *- stosunek ilości jedynek i zer w profilu X,* $p+q=1$.

Wniosek wynikający z twierdzenia 9 [P3] jest szczególnie ważny w przypadku integracji wektorów binarnych. Wynika z niego istotny fakt, że na jakość consensusu 1- optymalnego nie wpływa ilość źródeł wejściowych, ale jedynie zawartość profilu. W przypadku consensusu 2-optymalnego wykonano badania eksperymentalne. Zasymulowano różną wielkość profilu oraz różne rozkłady zer i jedynek w wektorach binarnych, na podstawie których był wyznaczany consensus. Otrzymane wyniki badań zostały przeanalizowane statystycznie za pomocą testu Wilcoxona. W większości przypadków, na podstawie 3 spójnych źródeł wejściowych, możliwe było wyznaczenie consensusu o satysfakcjonującym poziomie jakości. Zatem, aby zapewnić pożądaną jakość consensusu spełniającego zarówno postulat 1 i 2-optymalności, należy zwrócić uwagę na zawartość profilu (spójność), a nie jego liczebność.

Kolejne twierdzenia przedstawiają różne własności profilu takie jak: spójność oraz podatność, które wpływają na końcową jakość consensusu.

Twierdzenie 10. (Tw. 2 w pracy [P1]) *Jeśli dla każdej pary obiektów* $x, y \in U$ *spełniony jest warunek regularności:* $\delta^i(x,X) = \delta^i(y,X)$, *to każdy consensus profilu* $X \in \Pi(U)$ *ma taką samą, najwyższą jakość, gdzie:* $i \in \{1,2\}$.

Twierdzenie 11. (Tw. 3 w pracy [P1]) *Jeśli profil* $X \in \Pi(U)$ *charakteryzuje się minimalną spójnością definiowaną jako:* $cf(X) = 1 - \frac{\sum_{x,y \in X} \delta^i(x,y)}{n(n+1)}$, *to consensus dla tego profilu ma minimalną jakość:* $\mathrm{Q}^i(\mathrm{x},\mathrm{X}) = 0$, *gdzie:* $i \in \{1,2\}$, *n-liczebność profilu X.*

Twierdzenie 12. (Tw. 4 w pracy [P1]) *Załóżmy, że x jest consensusem profilu* X, a *y consensusem profilu Y, gdzie* $X, Y \in \Pi(U)$. *Jeżeli* $Q^1(x,X) \geq Q^1(y,Y)$ *to* $cf(X) \geq cf(Y)$.

Twierdzenie 13. (Tw. 5 w pracy [P1]) *Załóżmy, że x jest consensusem 1-optymalnym profilu* $X \in \Pi(U)$, *a x' jest consensusem 1-optymalnym profilu* $Y = X - \{y\} \in \Pi(U)$, *gdzie y spełnia poniższy warunek:* $\delta^1(x,y) = \max_{z \in X} \delta^1(x,z)$. *Jakość consensusu dla profilu X i Y spełnia poniższą zależność:* $\mathrm{Q}^1(\mathrm{x},\mathrm{X}) \leq \mathrm{Q}^1(\mathrm{x}',\mathrm{Y})$.

Twierdzenie 14. (Tw. 6 w pracy [P1]) *Jeśli profil* $X \in \Pi(U)$ *jest podatny na consensus, który spełnia postulat 1-optymalności, to poniższa zależność jest prawdziwa:* $\mathrm{Q}^1(\mathrm{x},\mathrm{X}) \geq \frac{2}{n+1}$, *gdzie n-liczebność profilu X.*

Twierdzenia 6-14 zostały udowodnione matematycznie w pracy [P1], [P3], które w konsekwencji pozwoliły na wskazanie:

- własności, jaką muszą spełniać źródła wejściowe (ich zawartość oraz liczba) w postaci zbiorów z powtórzeniami oraz wektorów binarnych, aby wyznaczony na ich podstawie consensus 2-optymalny cechował się maksymalną możliwą jakością. Wyniki przedstawiono w postaci matematycznego wzoru, który może służyć jako kryterium oceny efektywności procesu integracji.
- własności, jaką muszą spełniać źródła wejściowe w postaci wektorów binarnych, aby uzyskać jakość consensusu powyżej zadanej wartości.

Ponadto, na podstawie twierdzeń nr 10-14 wyciągnięto następujące wnioski, które mogą zostać wykorzystane w procesie wyboru źródeł wejściowych do integracji:

- profil, który spełnia własność regularności pozwala na wyznaczenie consensusu o najwyższej jakości.
- profil o minimalnej spójności uniemożliwia wyznaczenie consensusu o akceptowalnej jakości, dlatego ocena spójności profilu jest podstawą do oceny jakości consensusu. Zwiększenie spójności profilu zwiększa jakość ich reprezentanta. W przypadku ograniczania ilości integrowanych źródeł do uzyskania finalnej wersji o najwyższej jakości, należy wybrać najbardziej spójne źródła.
- profil, którego elementy są bardziej skupione wokół pewnej wartości pozwalają na wyznaczenie consensusu o wyższej jakości, niż profil o rozproszonych elementach.
- usunięcie elementów odstających poprawia jakość wyznaczonego consensusu. Wniosek uzyskany na podstawie przeprowadzonych badań analitycznych umożliwia redukcję ewentualnych źródeł wejściowych do integracji, których wartości znacząco różnią się od pozostałych.

Podsumowując, wszystkie przedstawione w tym podrozdziale własności zostały przedstawione w formie dowodów matematycznych. Zaproponowane własności pozwolą uniknąć kosztów integracji, których rezultat nie jest wiarygodny lub nie będzie spełniał założonej jakości. Ponadto wykazano, że zmiana lub ograniczenie źródeł wejściowych, może przyczynić się do poprawy jakości wyniku integracji.

B) Ontologie

W ostatnich latach ontologie stały się wygodną i elastyczną metodą reprezentacji wiedzy umożliwiająca jej łatwe przetwarzanie w systemach komputerowych. Złożoność ich struktury pozwala na modelowanie wielu rzeczywistych zjawisk zachodzących w świecie. Ontologie znalazły szeroki wachlarz zastosowań: od modelów prognozowania pogody po systemy ekspertowe. Formalny model ontologii składa się z kilku poziomów: konceptów, relacji między nimi, instancji oraz relacji między instancji. Poziom konceptów odpowiada za modelowanie klas obiektów z pewnego wycinka rzeczywistości. Poziom relacji jest odpowiedzialny za powiązania między konceptami. Instancje są materializacją konkretnych konceptów, a relacje instancji reprezentują powiązania między instancjami. Problem integracji ontologii został zdekomponowany na podzadania odpowiadające omówionym poziomom. Polegają one na

utworzeniu końcowej ontologii wynikowej z elementów składowych struktur poddawanych integracji. Do zadania integracji ontologii zaadaptowano algorytmy z pracy [4] oraz opracowano własne metody. Wszystkie prezentowane procedury wykorzystują teorię consensusu, która formalnie została przedstawiona w sekcji 4.3.3 A) niniejszego autoreferatu.

W pracy badawczej założono istnienie zamkniętego świata definiowanego jako para (A,V), gdzie A jest skończonym zbiorem atrybutów do opisu obiektów, a V jest zbiorem dziedzin atrybutów, $\bigcup_{a\in A} V_a$.

Definicja 6. *Ontologia oparta o świat (A,V) definiowana jest jako piątka:*

$$O = (C, T, R^C, I, R^I) \tag{11}$$

gdzie: C jest skończonym zbiorem konceptów, T opisuje hierarchię konceptów, R^C jest skończonym zbiorem relacji między konceptami, I oznacza skończony zbiór identyfikatorów instancji, R^I jest skończonym zbiorem relacji między instancjami.

Zatem każda ontologia składa się z czterech poziomów abstrakcji: konceptów, relacji między konceptami, instancji oraz relacji między instancjami. Poniżej zostanie przedstawiona krótka charakterystyka każdego elementu składowego ontologii.

Definicja 7. *Każdy koncept $c \in C$ definiowany jest jako krotka:*

$$c = (id^c, A^c, V^c, I^c) \tag{12}$$

gdzie: id^c jest identyfikatorem konceptu c, A^c jest zbiorem atrybutów konceptu c, $A^c \subseteq A$, V^c jest dziedziną wartości atrybutów: $V^c = \bigcup_{a\in A^c} V_a$), I^c jest zbiorem instancji konceptu c.

Dla uproszczenia, przez $a \in c$ oznaczono, że atrybut a należy do zbioru atrybutów A^c. Atrybuty ze zbioru A nie posiadają żadnej semantyki. Stają się interpretowalne tylko wtedy, gdy są częścią wybranego konceptu. Przykładowo, atrybut *klasa,* który należy do konceptu *szkoła* oznacza grupę uczniów uczęszczających wspólnie na zajęcia, jednak w kontekście konceptu *bilet, klasa* charakteryzuje miejsca, przedziały, wagony w środkach komunikacji publicznej zapewniające określony standard i wygody podróży.

W dalszej części pracy założono istnienie zbioru D_A, który zawiera atomowe, niepodzielne opisy atrybutów oraz operatory logicznej koniunkcji, dysjunkcji i negacji. Przez L_S^A oznaczono język formalny, który służy do opisu wartości semantyki atrybutów. Formalnie, semantyką atrybutów będzie funkcja o poniższej sygnaturze:

$$S_A : A \times C \rightarrow L_S^A \tag{13}$$

Instancje będące materializacją poszczególnych konceptów, definiowane są formalnie w następujący sposób:

Definicja 8. *Instancjami konceptów będziemy nazywać elementy zbioru I, które przyjmują postać dwójek:*

$$in = (id^{in}, v_c^{in}) \tag{14}$$

gdzie: id^{in} jest identyfikatorem instancji in, v_c^{in} jest funkcją o sygnaturze: v_c^{in}: $A^c \rightarrow V^c$.

Dla uproszczenia zapisu, przez $in \in c$ oznaczono przynależność instancji in do zbioru instancji I^c konceptu c. Dodatkowo, zbiór identyfikatorów instancji definiowany jest jako: $I = \bigcup_{c \in C}\{id^{in} | (id^{in}, v_c^{in}) \in I^c\}$, a zbiór identyfikatorów instancji konceptu c jako $Ins(c) = \{id^{in} | (id^{in}, v_c^{in}) \in I^c\}$.

Ostatnie dwa elementy ontologii to relacje między konceptami i instancjami. Przez R^C oznaczono skończony zbiór relacji między konceptami, taki, że $R^C = \{r_1^C, r_2^C, \dots, r_n^C\}, n \in N$, a każda relacja należąca do zbioru relacji $r_i^C \in R^C \;\; i \in [1, n]$ jest podzbiorem iloczynu kartezjańskiego $r_i^C \subseteq C \times C$.

Analogicznie do atrybutów, przez D_R zdefiniowano zbiór niepodzielnych cech relacji oraz operatory logiczne koniunkcji, dysjunkcji oraz negacji. Semantyką relacji konceptów będzie nazywana funkcja:

$$S_R : R^C \rightarrow L_S^R \tag{15}$$

Wśród relacji między konceptami wyróżniono szczególny typ jakim są hierarchie konceptów:

$T \subset C \times C$. Jeśli para konceptów (c_1, c_2) należy do hierarchii, oznacza to, że koncept c_1 jest bardziej ogólny niż koncept c_2 $(c_2 \leftarrow c_1)$ wtedy i tylko wtedy, gdy spełnione są poniższe warunki:

- $|A^{c_1}| \geq |A^{c_2}|$
- $\forall a_2 \in A^{c_2} \; \exists a_1 \in A^{c_1}: (a_1 \equiv a_2) \vee (a_2 \leftarrow a_1)$
- $Ins(c_1) \subseteq Ins(c_2)$.

Wspomniana relacja dziedziczenia umożliwia rozpatrywanie ontologii jako taksonomii. Jej korzeniem jest abstrakcyjna klasa *Thing*, po której dziedziczą wszystkie koncepty analizowanej ontologii.

Ostatnim elementem składowym ontologii są relacje między instancjami, które oznaczono jako: $R^I = \{r_1^I, r_2^I, \dots, r_n^I\}, n \in N$. Każda relacja ze zbioru R^C ma komplementarną relację ze zbioru R^I, dlatego $|R^C| = |R^I|$.

Ontologie są reprezentacją wiedzy, która pozwala na modelowanie różnego rodzaju związków i relacji między obiektami. Jednakże wielkość ontologii, mierzona ilością jej elementów składowych, nie zawsze przekłada się na ilość zawartych w nich informacji o rzeczywistości. Motywacją do rozpoczęcia prac była analiza, która wykazała, że w procesie integracji ontologii może dojść do kilku przypadków:

- *"przyrostu wiedzy"* - przyrost wiedzy jest dodatni, jeśli w wyniku łączenia dwóch lub więcej ontologii pojawiają nowe elementy. Połączenie dwóch kompletnie różnych ontologii ze sobą przyniesie nam 100% zysk w kontekście uzyskanej wiedzy. Dodatkowo, w przypadku integracji ontologii na poziomie relacji może dojść do "utworzenia wiedzy" np. jeżeli ontologia 1 zawiera relacje: $a \rightarrow b$, a ontologia 2: $b \rightarrow c$

to w wyniku połączenia tych dwóch ontologii dowiadujemy się na podstawie wnioskowania nowego faktu: $a \rightarrow c$, gdzie a,b,c to relacje występujące w integrowanych ontologiach. W związku z tym przyrost wiedzy może być nawet większy od 100%.

- *"utraty wiedzy"* - w przypadku łączenia ontologii, których zawartość jest sprzeczna ze sobą dochodzi do konfliktu, wynikającego z faktu, że nie jest możliwe stwierdzenie, która z ontologii wejściowych zawiera poprawną wiedzę. W konsekwencji takie informacje należy odrzucić, co powoduje, że po integracji takich ontologii część wiedzy jest tracona.
- *"braku zmiany ilości wiedzy"* - jeśli łączone są ze sobą te same lub bardzo podobne ontologie, to w efekcie ich integracji nie pojawia się żadna nowa wiedza, co jest nie opłacalne z w punktu widzenia przyrostu wiedzy.

W oparciu o zaobserwowane własności, w pracach: [P4], [P5], [P6], [P7], [P8] **przedstawiono zestaw metod, które pozwalają na ocenę potencjalnego przyrostu wiedzy w procesie integracji ontologii na poziomie:**

- **konceptów,**
- **relacji,**
- **instancji,**
- **relacji instancji.**

Opracowane algorytmy pozwalają na ustalenie kompromisu między kosztem integracji, a przyrostem wiedzy (który może być utożsamiany z pewnym zyskiem), który wyniknie z ewentualnego połączenia (integracji) poszczególnych źródeł wejściowych. W ramach przeprowadzonych badań opracowano zbiór metod, które pozwolą jeszcze przed rozpoczęciem procesu integracji (lub najpóźniej w jego trakcie), na oszacowanie ilości wiedzy jaką dzięki niemu "uzyskamy". Inaczej ujmując, opracowane algorytmy pozwalają na oszacowanie efektywności procesu integracji z punktu widzenia przyrostu wiedzy. Zaproponowane metody integracji ontologii oraz oceny efektywności tego procesu składają się na autorski zestaw narzędzi nazwany FOKI (Framework for Ontological Knowledge Integration). FOKI wykorzystują autorski model ontologii rozwijany w wielu poprzednich pracach autora niniejszego opracowania. FOKI różnią się od innych, znanych z literatury narzędzi przeznaczonych do zadania integracji, ze względu na wykorzystywane w tym procesie informacje o semantyce atrybutów.

W pracy [P4] zaproponowano półautomatyczny algorytm wyznaczania wartości semantyki atrybutów, który jest również elementem składowym FOKI. Prosty przykład zaprezentowany w pracy wykazał, że na wynik integracji oraz ocenę przyrostu wiedzy może wpływać wartość semantyki atrybutów (Tabela 1 [P4]). W początkowej fazie badań nad ontologiami założono, że wartości semantyki atrybutów wyznacza ekspert w oparciu o swoją wiedzę (podejście manualne). Jednak takie podejście obarczone jest ryzykiem błędu oraz brakiem wykorzystania opracowanych wcześniej przez siebie lub innego eksperta wartości semantyk (każdy ekspert dziedzinowy posiada inną wiedzę i doświadczenie, co przekłada się na różnorodność w wyznaczaniu wartości semantyki). Przykładowo, dla atrybutu *data_urodzin* należącego do konceptu *Osoba*, jeden z ekspertów może wyznaczyć wartość semantyki w postaci:

S_A(*data_urodzin, Osoba)* = *rok_urodzenia* ∧ *miesiąc_urodzenia* ∧ *dzień_urodzenia,* z kolei inny ekspert może zaproponować: S_A(*data_urodzin, Osoba)* = *rok_urodzenia* ∧ *miesiąc_urodzenia* ∧ *dzień_urodzenia* ∧ *wiek.* Różnice w wartościach semantyk mogą bezpośrednio przekładać się na wynik integracji dwóch ontologii.

W celu uniknięcia wspomnianej wady, opracowano autorską, półautomatyczną metodę wyznaczania wartości semantyki atrybutów w oparciu o zewnętrzną bazę wiedzy jaką jest WordNet (Algorytm 2 [P4]). Zaproponowana metoda została przebadana eksperymentalnie na benchmarkowym zestawie ontologii dostarczanym przez OAEI [19]. W badaniach wzięło udział 12 ochotników, którzy zostali poproszeni o wyznaczenie wartości semantyk dla wybranych par ontologii za pomocą opracowanego algorytmu oraz w sposób manualny. Otrzymane rezultaty zaprezentowano w Tabeli 3 [P4]. Wyniki zawierają zarówno czas generowania wartości semantyki, przyrost wiedzy na poziomie konceptów obliczony na podstawie wyznaczonych wartości semantyk oraz podobieństwo wyników integracji w przypadku ontologii, dla których wartości semantyk wyznaczono w sposób półautomatyczny (*A*) oraz ręczny (*M*).

Celem badań eksperymentalnych była odpowiedź na poniższe pytania badawcze:

1. *Czy wynik integracji ontologii, dla których wartości semantyk zostały wyznaczone za pomocą opracowanej metody oraz manualnie, są podobne?*
2. *Czy półautomatyczna metoda wyznaczania wartości semantyki jest efektywniejsza pod kątem czasowym od metody manualnej?*
3. *Czy użyteczność wartości semantyki atrybutów wyznaczonych opracowaną metodą jest większa niż wyznaczonych ręcznie przez eksperta?*

W celu odpowiedzi na pierwsze pytanie zdefiniowano miarę podobieństwa wyników integracji dla dwóch ontologii O_1, O_2:

$$sim(O_1, O_2) = \frac{\sum_{s,t \in map(O_1,O_2)} S(s,t)}{|map(O_1,O_2)|} \tag{16}$$

gdzie: $S(s,t) = max\frac{A^s \cap A^t}{A^s \cup A^t}$, $map(O_1, O_2)$- zbiór zmapowanych par konceptów <s,t>, przez *s* oznaczono koncept ontologii O_1, a przez t koncept ontologii O_2.

Następnie sprawdzono średnie podobieństwo wyników integracji, gdy wartości semantyk zostały wyznaczone manualnie lub za pomocą opracowanej metody. Ze względu na brak rozkładu normalnego próbki badawczej (kolumna *Sim(M,A)* Tab. 3 [P4]) wykorzystano test Wilcoxona, który wykazał, że podobieństwo wyników integracji jest większe niż 96% (*p-value* równe 0,0042 wartość statystyki równa 13).

Bezpośrednia analiza wyników Tabeli 3 [P4] wykazała, że czas generowania wartości semantyki metodą półautomatyczną (kolumna (*A*) Tab. 3 [P4]) jest nawet dwukrotnie krótszy niż czas generowania wartości semantyki ręcznie (kolumna (*M*) Tab. 3 [P4]). Powyższy wniosek jest odpowiedzią na drugie pytanie badawcze sformułowane w tym podrozdziale.

Odpowiedzią na ostatnie pytanie była analiza wyników zawartych w Tabeli 4 [P4], która potwierdziła większą użyteczność wartości semantyk wygenerowanych opracowaną metodą.

W przypadku metody manualnej, zgodnie z intuicją, trudno znaleźć identyczne lub podobne wartości semantyk wyznaczone przez dwóch, pracujących niezależnie ekspertów. W przypadku metody półautomatycznej, podobieństwo wartości semantyk wygenerowanych przez dwóch ekspertów jest nie mniejsze niż 68%, co potwierdzono testem Wilcoxona.

Podsumowując, semantyka atrybutów jest istotnym elementem ontologii, wpływającym zarówno na wynik integracji, jak i wartości przyrostu wiedzy uzyskanymi w wyniku łączenia ontologii. Opracowana metoda półautomatycznego wyznaczania wartości semantyki pozwala na skrócenie czasu jej generowania, przy równoczesnym zachowaniu wysokiego stopnia użyteczności.

Jednakże głównym elementem narzędzi FOKI są algorytmy integracji oraz oceny potencjalnego przyrostu wiedzy na poziomie: konceptów, instancji, relacji i relacji instancji zaprezentowane w [P5], [P6], [P7], [P8]. Poprzez $\Delta_C(O_1, O_2)$, $\Delta_{RC}(O_1, O_2)$, $\Delta_I(O_1, O_2)$, $\Delta_{RI}(O_1, O_2)$ oznaczono funkcje do oszacowania przyrostu wiedzy wynikającego z przyłączenia ontologii O_2 do ontologii O_1. Opracowane miary pozwalają znaleźć odpowiedź na pytanie o ilość nowej wiedzy w procesie przyłączenia ontologii O_2 w stosunku do tego, co już było wcześniej wiadome na podstawie ontologii O_1. Tak sformułowane zadanie będziemy nazywać subiektywnym przyrostem wiedzy.

W pierwszym kroku opracowano miarę przyrostu wiedzy na poziomie konceptów:

Definicja 9. *Przyrost wiedzy w zadaniu integracji ontologii na poziomie konceptów wyznaczany jest za pomocą funkcji o sygnaturze:* $\Delta_C(O_1, O_2)$*:* $C_1 \times C_2 \rightarrow [0,1]$ *i wyliczany według poniższego wzoru:*

$$\Delta_C(O_1, O_2) = 1 - \frac{|map(O_1,O_2)| - \sum_{<s,t>\in map(O_1,O_2)} \max_{c_i \in t} d_s(s,c_i)}{|C_1| + |C_2| - \sum_{<s,t>\in map(O_1,O_2)} |t|} \tag{17}$$

gdzie: funkcja d_s: $C_1 \times C_2 \rightarrow [0,1]$, $d_s(c_1, c_2) = dist(ctx(c_1), ctx(c_2))$,

$ctx(c\ ) = S_A(a_1, c) \wedge S_A(a_2, c)) \wedge \ldots \wedge S_A(a_p, c)$.

Szczegółowy sposób wyznaczania wartości funkcji *dist* można znaleźć w pracach innych autorów, np. [25]. Służy ona do wyznaczenia odległości między wartościami semantyki atrybutów przypisanych do poszczególnych konceptów.

Aplikowalność opracowanej metody przyrostu wiedzy na poziomie konceptów ([P6]) wykazano poprzez zastosowanie jej do zadania integracji wielostopniowej. Zadanie integracji jednostopniowej polega na znalezieniu najlepszej reprezentacji O^{1*} dla wszystkich ontologii wejściowych: $O_1^1, O_2^1, \ldots, O_n^1$. Zadanie integracji wielostopniowej polega najpierw na podziale ontologii wejściowych na mniejsze grupy, a następnie na ich integracji w ramach jednej grupy. Wyniki cząstkowych integracji stają się ontologiami wejściowymi do zadania integracji na kolejnym poziomie. Formalnie, integracja wielostopniowa na *m*-tym poziomie polega na znalezieniu najlepszej reprezentacji O^{m*} dla ontologii wejściowych: $O_1^{m-1*}, O_2^{m-1*}, \ldots, O_n^{m-1*}$, które są wynikami integracji ontologii na poziomie *m-1*, gdzie $m \geq 2$. Do zadania integracji zaadaptowano algorytm z pracy [4] zaprezentowany w publikacji z cyklu jako Algorytm 1 [P6]. Analiza matematyczna algorytmu integracji wykazała, że rezultat integracji jedno- i wielostopniowej jest taki sam. Sformułowano i udowodniono twierdzenie, które wskazuje, że

wynik połączenia wielu źródeł wiedzy w jedną, spójną wersję jest taki sam, jak wynik integracji wielostopniowej:

Twierdzenie 15. (Tw. 1 w pracy [P6]) *Wynik integracji jednostopniowej* O^{1*} *jest zawsze równy wynikowi integracji wielostopniowej* O^{m*} *w przypadku łączenia ontologii na poziomie konceptów, gdzie:* $m \geq 2$.

Ze względu na konsekwencje wynikające z twierdzenia 15, badania eksperymentalne zostały poświęcone jedynie analizie czasu potrzebnego do integracji jedno- i wielostopniowej. Badania były wykonywane na benchmarkowym zbiorze ontologii udostępnianych przez organizację OAEI [19]. W trakcie eksperymentów szacowano czas integracji i przyrost wiedzy (Tabela 1 [P6]). Zgodnie z przewidywaniami, czas integracji wielostopniowej jest nawet o 20% krótszy niż w przypadku integracji jednostopniowej, ze względu na możliwość zrównoleglenia obliczeń. Przyrost wiedzy wyznaczany w poszczególnych przypadkach testowych pozwolił na wskazanie ontologii najbardziej istotnej w procesie integracji, czyli niosącej najwięcej wiedzy (Tabela 2 [P6]). Ponadto wykazano, że dla testowanych zestawów ontologii nie można wykluczyć kolejnych etapów integrowania, gdyż w każdym z nich przyrost wiedzy był znaczący. Podsumowując, opracowane metody szacowania przyrostu wiedzy w zadaniu integracji wielostopniowej mogą służyć zarówno jako kryterium wyboru ontologii wejściowych jak i warunek stopu dla kolejnych poziomów integracji, a zatem wpływać bezpośrednio na jakość ostatecznego rezultatu integracji ontologii.

Metodę oceny przyrostu wiedzy na poziomie konceptów zweryfikowano również w pracy [P4]. Za pomocą testu Wilcoxona zweryfikowano czy wartość przyrostu wiedzy, w którym wartość semantyk została wyznaczona w sposób półautomatyczny (kolumna Δ(A) Tab. 3 [P4]) oraz w sposób manualny (kolumna Δ(M) Tab. 3 [P4]) jest równa. Analiza wyników wykazała (*p-value* równe 0,74, wartość statystyki równa 14,5), że wartości przyrostu wiedzy są statystycznie niezależne od metody generowania wartości semantyki.

Kolejnym elementem składowym ontologii są relacje między konceptami. Wśród relacji można wyróżnić ich szczególną postać jakim są hierarchie konceptów.

Definicja 10. *Przyrost wiedzy w zadaniu integracji ontologii na poziomie relacji wyznaczany jest za pomocą funkcji o sygnaturze:* $\Delta_{R^C}(O_1, O_2)$: ${R^C}_1 \times {R^C}_2 \rightarrow (-\infty, \infty)$.

Na poziomie relacji zbiór wartości jest większy niż dla konceptów, gdyż na tym poziomie możemy w wyniku integracji „utworzyć" nową wiedzę jak i „stracić". Do wyznaczenia wartości przyrostu wiedzy na poziomie relacji zaproponowano Algorytm 3 [P5], a dla hierarchii Algorytm 4 [P5]. Działania obu algorytmów zilustrowano kilkoma przykładami.

Definicja 11. *Przyrost wiedzy w zadaniu integracji ontologii na poziomie instancji wyznaczany jest za pomocą funkcji o sygnaturze:* $\Delta_I(O_1, O_2)$: $C \times I \rightarrow [-1,1]$ *i wyliczany według poniższego wzoru:*

$$\Delta_I(O_1, O_2) = \frac{1}{|C^*|} \sum_{c \in C^*} \left(\frac{1}{|I^C|} \sum_{i \in I^C} \widehat{\Delta_I}(O_1, O_2) \right) \tag{18}$$

gdzie: C^**- jest zbiorem zintegrowanych konceptów pochodzących z ontologii* O_1, O_2, $\widehat{\Delta_I}$ *wartość wyznaczona na podstawie Algorytmu 1 [P8] dla zbioru instancji należących do wybranego konceptu* c^*.

Zaproponowaną w powyższej definicji miarę oceny przyrostu wiedzy zweryfikowano eksperymentalnie. Przygotowano 27 pytań, które zawierały proste przykłady integracji dwóch lub trzech ontologii na poziomie instancji. Nastepnie poproszono 21 ochotników, którzy mieli ocenić w skali [-100%, 100%] potencjalny przyrost wiedzy, który zaobserwowali w wyniku łączenia ontologii zaprezentowanych w zadaniach. Podejście, gdzie metody czy algorytmy są weryfikowane przez ekspertów pełniących rolę potencjalnych użytkowników, jest szeroko stosowana i akceptowana w dziedzinie przedmiotu [26]. Następnie, dla przypadków testowych wyznaczono przyrost wiedzy zgodnie z definicją 11. Tak przygotowane próbki zostały poddane analizie statystycznej za pomocą testu badającego współczynnik korelacji wewnątrzklasowej (ICC). Zarówno wynik spójności oraz bezwzględnej spójności był bardzo wysoki (ok. 0,975) oraz istotny statyczynie (*p-value* mniejsze niż 0,0000001). Oznacza to, że zaproponowana metoda wyznaczania przyrostu wiedzy na poziomie instancji koreluje z intuicyjnym sposobem oceny przyrostu wiedzy przez człowieka. Dowodzi to poprawności przyjętych założeń.

Ostatnim elementem składowym ontologii, który wymaga integracji, są relacje instancji. Relacje między konceptami wskazują potencjalne powiązania miedzy instancjami, natomiast relacje instancji je konkretyzują.

Definicja 12. *Przyrost wiedzy w zadaniu integracji ontologii na poziomie relacji instancji wyznaczany jest za pomocą funkcji o sygnaturze:* $\Delta_{R^I}(O_1, O_2)$*:* ${R^C}_1 \times {R^C}_2 \rightarrow [-1, \infty)$.

Podobnie jak poprzednio, zbiór wartości nie mieści się w przedziale [0,1], gdyż na tym poziomie w wyniku integracji możemy „utworzyć" nową wiedzę jak i „stracić". Do wyznaczenia wartości przyrostu wiedzy na poziomie instancji relacji zaproponowano Algorytm 1 [P7]. Poprawność jego działania zilustrowano kilkoma przykładami. Następnie, podobnie jak w przypadku instancji, przygotowano 20 przypadków integracji dwóch ontologii na poziomie instancji relacji. W eksperymencie wzięło udział 45 ochotników pełniących rolę ekspertów. Badania zostały podzielone na dwie części. W pierwszej uczestnicy mieli za zadanie ocenić ogólną zmianę, wynikająca z połączenia ontologii (wzrost, spadek wiedzy oraz brak zmiany). W drugiej części mieli podać precyzyjne wartości z przedziału od -100% do ∞. Do analizy statystycznej pierwszych wyników wykorzystano współczynnik korelacji Kappa Cohena. Otrzymano *p- value* mniejsze niż 0,000006 oraz współczynnik Kappa równy 70,15%, co oznacza, że wartości wskazane przez ekspertów oraz wyznaczone na podstawie definicji 12 są zbliżone. Uczestnicy eksperymentu zauważają ogólny trend zmian wiedzy w przypadku integracji ontologii na poziomie instancji relacji.

Porównanie dwóch próbek- konkretnych wartości zmiany przyrostu wiedzy wyznaczonej przez eksperta oraz wyznaczonych za pomocą Algorytmu 1 [P7] - przeprowadzono za pomocą współczynnika korelacji wewnątrzklasowej (ICC). Otrzymano wyniki istotne statystycznie (*p-value* 0,0045), jednak poziom absolutnej zgodności wyniósł tylko 0,579 i spójności 0,714. Wyniki badań wykazały, że opracowane procedury szacowania przyrostu wiedzy są intuicyjne dla ludzi i pokrywają się z ich sposobem myślenia. W przypadku poziomu relacji instancji, uczestnikom eksperymentu łatwiej było zaobserwować ogólny trend zmiany (przyrost, spadek albo brak zmiany w ilości wiedzy), niż szczegółowe wartości.

Dla poprawności oraz efektywności procesu integracji kluczowe znaczenie ma również mapowanie między ontologiami (nazywane również odwzorowaniem). Problem odwzorowywania ontologii wiąże się przede wszystkim z różnicami w sposobie opisu świata rzeczywistego. Ten sam byt może posiadać dwie różne nazwy w różnych ontologiach lub po

prostu być inaczej zdefiniowany. Co więcej, rozpatrywane ontologie mogą zawierać tę samą wiedzę, ale zapisaną w różnych językach. Proces integracji ontologii polega na utworzeniu końcowej ontologii wynikowej z elementów składowych struktur poddawanych integracji. Jednak identyfikacja fragmentów ontologii, które należą do tego samego modelowanego wycinka rzeczywistości, jest zadaniem związanym z wyznaczeniem odwzorowania (mapowania) między ontologiami. Do wyznaczania mapowania między ontologiami powstało wiele narzędzi m.in. LogMap, ALIN, XMap, Kepler, Lily itp., jednak jakość odwzorowań wygenerowanych przez wspomniane narzędzia jest różna. Niestety, słaba jakość mapowań między ontologiami, pociąga za sobą problemy związane z procesem integracji, której końcowy wynik może nie być zadowalający w kontekście przyjętych kryteriów. Brak poprawnych mapowań powoduje, że ontologie po integracji są redundantne oraz niespójne. W związku z tym poprawne i kompletne odwzorowanie jest kluczowe dla efektywności procesu integracji. Od 2004 roku OAEI (Ontology Aligment Evaluation Initiative) [19] organizuje zawody, których głównym zadaniem jest wyłonienie najlepszych narzędzi do mapowania. W tym celu OAEI udostępnia zestawy ontologii wraz z ich odwzorowaniami wzorcowymi, które można określić mianem „złotego standardu". Odwzorowania wygenerowane przez różne narzędzia biorące udział w konkursie OAEI są następnie porównywane z takimi benchmarkowymi odwzorowaniami za pomocą miar precyzji, kompletności oraz miary F. Takie podejście do oceny odwzorowań, pomimo wielu zalet, jest efektywne jedynie podczas eksperymentu. W praktyce jest mało przydatne, gdyż w rzeczywistych sytuacjach, wymagających odwzorowywania ontologii, nie istnieją żadne wzorcowe mapowania. Omówione problemy stały się motywacją do podjęcia prac związanych z oceną mapowań, które później wykorzystywane są do procesu integracji ontologii.

Celem pracy [P9] było zdefiniowanie zestawu miar służących do oceny odwzorowania ontologii biorących udział w procesie integracji. W ramach prac zaproponowano cztery metody oceny jakości mapowań wykorzystując informacje o głębokości i ciągłości mapowanych elementów zarówno na poziomie konceptów jak i instancji. Opracowane metody są istotnym elementem składowym zestawu narzędzi FOKI. Definicje 13 i 14 prezentują miary, które umożliwiają ocenę odwzorowań na poziomie konceptów. W dalszej części pracy przez $\tilde{O}$ oznaczono zbiór wszystkich ontologii, a przez $\widetilde{Alin}$ zbiór wszystkich mapowań między ontologiami. Formalnie, odwzorowanie $Alin$ dla dwóch ontologii O_1, O_2 rozumiane jest jako skończony zbiór 4-elementowych krotek $< e_1, e_2, sr, cv >$, gdzie e_1, e_2 to w zależności od rozważanego elementu ontologii mogą być koncepty, instancje albo relacje, $sr \in \{=, <.>, \%\}$ to semantyczna relacja opisująca rodzaj powiązania między mapowanymi elementami, $cv \in [0,1]$ stopień pewności danego powiązania.

Definicja 13. *Miara oceny mapowania między ontologiami na poziomie konceptów w oparciu o głębokość mapowanych klas wyznaczana jest za pomocą funkcji o sygnaturze:* $\lambda_D: \tilde{O} \times \tilde{O} \times \widetilde{Alin} \rightarrow \mathbb{R}$*:*

$$\lambda_D(O_1, O_2, Alin) = \sum_{corr_{cc'} \in Alin} \gamma_D(O_1, O_2, Alin, corr_{cc'}) \tag{19}$$

gdzie: $corr_{cc'} =< c, c', =, cv >$ *oznacza mapowanie między dwoma konceptami c, c',* γ_D*-wyznaczany jest z poniższego wzoru:*

$$\gamma_D(O_1, O_2, Alin, corr_{cc'}) = \begin{cases} \frac{1}{Depth(O_1)-Depth(O_1,c)+1} + \frac{1}{Depth(O_2)-Depth(O_2,c')+1} & corr_{cc'} \sim Alin \\ 0 & corr_{cc'} \nsim Alin \end{cases} \quad (20)$$

$Depth(O)$ oznacza głębokość hierarchii ontologii, która mierzona jest liczbą relacji subsumpcji, dzieląca nadrzędną klasę *Thing* od najbardziej odległej w hierarchii dziedziczenia klasy należącej do tej ontologii. Natomiast $Depth(O, c)$ symbolizuje głębokość konceptu w hierarchii ontologii, czyli liczbę relacji subsumpcji w hierarchii ontologii, dzieląca koncept c od nadrzędnej klasy *Thing*. $corr_{cc'} \nsim Alin$ oznacza, że mapowanie konceptów *c, c'* w odwzorowaniu *Alin* są niespójne zgodnie z Definicją 6 w pracy [P10]. Przez $corr_{cc'} \sim Alin$ oznaczono spójność mapowań między konceptami *c, c'*.

Definicja 14. *Miara oceny mapowania między ontologiami na poziomie konceptów w oparciu o ciągłość mapowanych klas wyznaczana jest za pomocą funkcji o sygnaturze:* $\lambda_C: \tilde{O} \times \tilde{O} \times \widetilde{Alin} \to \mathbb{N}$*:*

$$\lambda_C(O_1, O_2, Alin) = \sum_{t \in SubTrees(O_1,\ Alin)} \gamma_C(t) + \sum_{t \in SubTrees(O_2,\ Alin)} \gamma_C(t) \quad (21)$$

gdzie: $\gamma_C(t) = |t|^2$*,* $|t|$*- liczba klas w poddrzewie t,* $SubTrees(O_1, Alin)$*,* $SubTrees(O_2, Alin)$ *są zbiorami poddrzew występujących w odpowiednio w ontologiach* O_1 *i* O_2 *dla konceptów odwzorowanych w Alin wyznaczanych na podstawie Algorytmu 2* [P10].

Sposób wyznaczania wartości miar λ_D, λ_C zaprezentowano odpowiednio za pomocą Algorytmu 1 i 2 [P10]. Kolejne dwie definicje (15 i 16) w sposób formalny prezentują miary, które można wykorzystać do oceny odwzorowań ontologii na poziomie instancji:

Definicja 15. *Miara oceny mapowania między ontologiami na poziomie instancji w oparciu o głębokość mapowanych klas wyznaczana jest za pomocą funkcji o sygnaturze:* $\sigma_D: \tilde{O} \times \tilde{O} \times \widetilde{Alin} \to \mathbb{R}$ *:*

$$\sigma_D(O_1, O_2, Alin) = \sum_{corr_{cc'} \in Alin} \psi_D(O_1, O_2, Alin, corr_{in\ in'}) \quad (22)$$

gdzie: $corr_{in\ in'} =< in, in', =, cv>$ *oznacza mapowanie między dwoma instancjami in, in',* $in \in c, in' \in c$ γ_D*- wyznaczany jest z poniższego wzoru:*

$$\psi_D(O_1, O_2, Alin, corr_{cc'}) = \begin{cases} \frac{1}{Depth(O_1)-Depth(O_1,c)+1} + \frac{1}{Depth(O_2)-Depth(O_2,c')+1} & c \equiv c' \\ 0 & c \not\equiv c' \end{cases} \quad (23)$$

gdzie: $c \equiv c'$ *oznacza, że dwa koncepty są równoważne,* $c \not\equiv c'$ *nie są równoważne.*

Definicja 16. *Miara oceny mapowania między ontologiami na poziomie instancji w oparciu o ciągłość mapowanych klas wyznaczana jest za pomocą funkcji o sygnaturze:* $\sigma_C: \tilde{O} \times \tilde{O} \times \widetilde{Alin} \to \mathbb{N}$*:*

$$\sigma_C(O_1, O_2, Alin) = \sum_{c \in C_1} \psi_C(Ins(O_1, c), Alin) + \sum_{c \in C_2} \psi_C(Ins(O_2, c), Alin) \quad (23)$$

gdzie: $\psi_C(Ins(O, c), Alin) = |s_{AL}|^2$, s_{AL}*-zbiór instancji Ins(O,c), które posiadają mapowanie w Alin.*

Sposób wyznaczania wartości miar σ_D, σ_C zaprezentowano odpowiednio za pomocą Algorytmu 3 i 4 [P10].

Opracowane algorytmy zostały przebadane eksperymentalnie. W oparciu o testowe zbiory ontologii udostępniane przez organizacje OAEI [19] oraz wybrane narzędzia do mapowania, wygenerowano odwzorowania, które zostały ocenione za pomocą miar: λ_D, λ_C. Następnie wyznaczono wartość miary *F*, precyzji i kompletności między wyznaczonym odwzorowaniem, a odwzorowaniem wzorcowym dostarczanym przez OAEI. Szczegółowe wyniki badań prezentuje Tabela 2 [P10]. Dla otrzymanych w ten sposób próbek wyznaczono współczynnik korelacji wewnątrzklasowej. Zarówno λ_D jak λ_C wykazują wysoką, istotną statystycznie zbieżność z miarą kompletności (odpowiednio: *ICC*(3,1)=0,934 dla λ_D i *ICC*(3,1)=0,837 dla λ_C) oraz miarą *F* (odpowiednio: *ICC*(3,1)=0,837 dla λ_D i *ICC*(3,1)=0,697 dla λ_C) (Tabela 3 [P10]).

Analiza statystyczna otrzymanych wyników pozwoliła stwierdzić, że istnieje korelacja między opracowanymi miarami, a miarami wyznaczonymi na podstawie porównania wyników z benchmarkowym zbiorem ontologii i ich odwzorowań. Oznacza to, że na podstawie opracowanych miar możliwa jest ocena poprawności i jakości mapowań między ontologiami bez potrzeby porównywania ich ze wzorcowym odwzorowaniem.

W zadaniu integracji lepsze (według opracowanych miar oceny odwzorowań) mapowanie między ontologiami przekłada się bezpośrednio zarówno na wynik końcowy integracji, jak i na wartość przyrostu wiedzy będącej miarą oceny efektywności procesu integracji ontologii. Oczywista jest zatem dodatnia korelacja zdefiniowanych miar oceny mapowań z efektywnością procesu integracji ontologii.

C) Hurtownie danych

Hurtownie danych określane są mianem „wysoce wyspecjalizowanych systemów baz danych optymalizowanych na potrzeby wspomagania decyzji" [27]. Hurtownie danych pełnią kluczową rolę w zarządzaniu przedsiębiorstwem i mogą być wykorzystywane w celu:

- przetwarzania analitycznego danych transakcyjnych (ang. OLAP- Online Analiticial Processing),
- wspomaganiu decyzji jako element systemów wspomagania decyzji (ang. DSS- Decision Support System),
- informowaniu kierownictwa jako element systemu EIS (ang. Executive Infromation System) oraz w inteligencji biznesowej (ang. BI- Business Intelligence).

Informacje przechowywane w przedsiębiorstwie mogą być rozproszone geograficznie lub funkcjonalnie. W celu sprawnego zarządzania przedsiębiorstwem oraz podejmowaniu strategicznych decyzji dotyczących całej korporacji istnieje potrzeba integracji takich

rozproszonych źródeł wejściowych. Można to osiągnąć poprzez utworzenie tzw. federacji hurtowni danych. Charakteryzuje się ona tzw. globalnym schematem, będącym ujednoliconym widokiem na skończony zestaw niezależnych hurtowni danych. Widok ten udostępniony zostaje użytkownikom końcowemu, jest on jednak całkowicie wirtualny, a cząstkowe hurtownie pozostają autonomiczne. Utworzenie takiego widoku wymaga jednak przeprowadzenia procesu, który można traktować jako formę integracji hurtowni cząstkowych.

Analogicznie do ontologii, wielkość pojedynczej hurtowni nie świadczy o ekspresywności przechowywanych w nich informacji. Zatem zasadne pojawia się pytanie: z ilu hurtowni danych powinna składać się federacja, aby podejmowane na jej podstawie decyzje nadal były trafne i rzetelne? W odpowiedzi na powyższe pytanie **opracowano zestaw metod [P10], które można wykorzystać do oceny opłacalności dodania nowej hurtowni danych do istniejącej federacji. Innymi słowy, zaproponowano metody oceny efektywności procesu integracji hurtowni danych.**

Hurtownia danych składa się z dwóch głównych elementów: wymiarów i miar. Schemat hurtowni danych jest zdefiniowany następująco: $H_i = \{D_0^i, D_1^i, \dots, D_{\alpha i}^i\}$, gdzie i jest identyfikatorem hurtowni w federacji, $D_0^i = \{m_1^i, \dots, m_{\beta 1}^i\}$- tabela faktów, β_i- liczba miar w D_0^i, m_z^i- z-ta miara w hurtowni H_i, $D_j^i = \{a_{j1}^i, \dots, a_{j\gamma i}^i\}$- tabele wymiarów, $j \in [1, \dots, \alpha i]$, γ_i - liczba atrybutów w wymiarze D_j^i, a_{jz}^i- z-ty atrybut i-tego wymiaru hurtowni H_i. Federacja hurtowni danych zdefiniowana jest jako piątka: $\hat{F} = (F, H, UI, qd, l)$, gdzie $F = \{D_0, D_1, \dots, D_n\}$- jest schematem federacji, D_0- schemat tabeli faktów, D_j- schematem tabel wymiarów, H- zbiorem hurtowni, UI- interfejsem użytkownika, qd- dekompozycją zapytań oraz l - procedurą integracji odpowiedzi. Między tabelą wymiarów D_j a tabelą faktów zachodzi relacja typu $1 - \infty$, co oznacza, że jeden wiersz z tabeli wymiarów może być skojarzony z więcej niż jednym wierszem tabeli faktów, ale jeden fakt ma przypisany dokładnie jeden wiersz z danego wymiaru. Stopień podobieństwa między wymiarem federacji D_j , a wymiarem hurtowni danych D_x^i oznaczany jest jako $DS(D_j, D_x^i)$. Jeśli $DS(D_j, D_x^i) > t$, oznacza, że wymiary te są dostatecznie podobne przy zadanej wartości progowej podobieństwa t, co zostało zapisane jako: $D_j \sim D_x^i$. Stopień podobieństwa między miarą federacji m, a miarą hurtowni danych m' oznaczono przez $MS(m, m')$. Dwie równoważne miary, czyli takie które spełniają warunek $MS(m, m') = 1$ zapisano jako $m^i \diamond m$. Metoda wyznaczania podobieństwa między miarami i wymiarami została zaczerpnięta i szerzej przedstawiona w pracy [28].

W ramach prowadzonych prac badawczych zdefiniowano szereg miar, które były podstawą algorytmów oceny ekspresywności informacji w przypadku tworzenia federacji hurtowni danych. Opracowane metody podzielono na dwie części: odpowiedzialne za poziom wymiarów i poziom miar. Ponadto założono, że ocenę ekspresywności można przedstawić z punktu widzenia hurtowni dołączanych do federacji (ocena subiektywna) oraz z punktu widzenia federacji (ocena obiektywna). W związku z tym, w ramach prac badawczych zaproponowano 4 różne metody szacowania ekspresywności, które formalnie zaprezentowano w poniższych definicjach:

Definicja 17. *Ocena ekspresywności informacji na poziomie wymiarów wyznaczana jest na podstawie funkcji o sygnaturze* $\delta_D^i: 2^{\hat{D}_{i-1}} \times 2^{\{a^i\}} \longrightarrow [0,1]$ *i obliczana w następujący sposób:*

$$\delta_D^i = w^+ * \delta_{D+}^i + w^- * \delta_{D-}^i \tag{24}$$

gdzie: $\widehat{D}_{i-1}$*- oznacza zbiór wymiarów w trakcie* $(i-1)$*-tej iteracji tworzenia federacji hurtowni, wagi:* $w^+, w^- \in [0,1]$ *spełniają warunek* $w^+ + w^- = 1$,

Wartość funkcji δ_{D-}^i obliczana jest za pomocą poniższego wzoru:

$$\delta_{D-}^i\left(\widehat{D}_{i-1}, \{D_1^i, \dots, D_{\alpha i}^i\}\right) = \frac{\sum_{d^i \in X_{D-}^i} (1 - \max_{d_j \in \widehat{D}_{i-1}} DS(d_j, d^i))}{|X_{D-}^i|} \tag{25}$$

gdzie: $X_{D-}^i = \{d^i \in \{D_1^i, \dots, D_{\alpha i}^i\} \exists \hat{d} \in \widehat{D}_{i-1} : d^i \sim \hat{d}\}$.

Funkcja δ_{D+}^i została zdekomponowana na dwie funkcje. W zależności od tego, czy chcemy policzyć subiektywną czy obiektywną miarę ekspresywności informacji, należy wykorzystać jeden z poniższych wzorów:

$$\textit{ocena subiektywna: } \delta_{SD+}^i\left(\widehat{D}_{i-1}, \{D_1^i, \dots, D_{\alpha i}^i\}\right) = \frac{|X_{D+}^i|}{|\{D_1^i, \dots, D_{\alpha i}^i\}|} \tag{26}$$

$$\textit{ocena obiektywna: } \delta_{OD+}^i\left(\widehat{D}_{i-1}, \{D_1^i, \dots, D_{\alpha i}^i\}\right) = 1 - \frac{|\widehat{D}_{i-1}|}{|\widehat{D}_{i-1}| + |X_{D+}^i|} \tag{27}$$

gdzie: $X_{D+}^i = \{d^i \in \{D_1^i, \dots, D_{\alpha i}^i\} \nexists \hat{d} \in \widehat{D}_{i-1} : d^i \sim d^i\}$.

W pracy [P10] sformułowano 5 postulatów jakie musi spełniać funkcja δ_{D-}^i. Na ich podstawie sformułowano i udowodniono twierdzenia, które prezentują własności funkcji δ_D^i.

Twierdzenie 16. (Tw. 1 w pracy [p10]) *Dla każdej federacji poniższy warunek jest zawsze prawdziwy:*

$$\widehat{D}_{i-1} \subseteq \{D_1^i, \dots, D_{\alpha i}^i\} \Leftrightarrow \delta_{SD+}^i\left(\widehat{D}_{i-1}, \{D_1^i, \dots, D_{\alpha i}^i\}\right) = \delta_{OD+}^i\left(\widehat{D}_{i-1}, \{D_1^i, \dots, D_{\alpha i}^i\}\right) \tag{28}$$

gdzie: $i \in \{2, \dots N\}$, *N- liczba hurtowni biorąca udział w tworzeniu federacji.*

Powyższe twierdzenie prezentuje fakt, że jeśli zbiór wymiarów *i*-tej hurtowni danych będzie zawierał wszystkie wymiary federacji $\widehat{D}_{i-1}$ w iteracji *i*-1, to wartość obiektywnej i subiektywnej miary oceny ekspresywności informacji są takie same. Inaczej ujmując, integracja przynosi te same korzyści zarówno dla *i*-tej hurtowni danych, jak i dla federacji. To twierdzenie upraszcza obliczanie subiektywnej i obiektywnej oceny przyrostu ekspresywności, gdyż wymaga obliczenia tylko jednej z wymienionych wartości. Kolejne twierdzenie wskazuje górne ograniczenie współczynnika δ_{D-}^i:

Twierdzenie 17. (Tw. 2 w pracy [p10]) *Dla każdej federacji poniższy warunek jest zawsze prawdziwy:*

$$\delta_{D-}^i\left(\widehat{D}_{i-1}, \{D_1^i, \dots, D_{\alpha i}^i\}\right) \leq (1 - t) \tag{28}$$

Funkcja oceny ekspresywności informacji na poziomie miar formalnie definiowana jest w następujący sposób:

Definicja 18. *Ocena ekspresywności informacji na poziomie miar wyznacza jest na podstawie funkcji o sygnaturze* $\delta_M^i: 2^{M_{i-1}^F} \times 2^{M^{H_i}} \longrightarrow [0,1]$ *i obliczana w następujący sposób:*

$$\textit{ocena subiektywna: } \delta_S^i M(M_{i-1}^F, M^{H_i}) = \frac{|X_{M+}^i|}{|M^{H_i}|} \tag{29}$$

$$\textit{cena obiektywna: } \delta_O^i M(M_{i-1}^F, M^{H_i}) = \frac{|M_{i-1}^F|}{|M_{i-1}^F| + |X_{M+}^i|} \tag{30}$$

gdzie: $X_{M+}^i = \{m^i \in M^{H_i} \nexists m \in M_{i-1}^F : m^i \diamond m\}$, M_{i-1}^F- *zbiór miar federacji w iteracji (i-1),* M^{H_i}- *zbiór miar i-tej hurtowni.*

Twierdzenie 18. (Tw. 3 w pracy [p10]) *Dla każdej federacji poniższy warunek jest zawsze prawdziwy:*

$$M_{i-1}^F \subset M^{H_i} \Leftrightarrow \delta_O^i M(M_{i-1}^F, M^{H_i}) = \delta_S^i M(M_{i-1}^F, M^{H_i}) \tag{31}$$

gdzie: i $\in \{2, \ldots N\}$, *N- liczba hurtowni biorąca udział w tworzeniu federacji.*

W przypadku, gdy zbiór miar *i*-tej hurtowni pokrywa się zbiorem miar federacji w (*i*-1)- tej iteracji, w celu wyznaczenia subiektywnego i obiektywnego przyrostu ekspresywności informacji na poziomie miar wystarczy wyznaczenie wartości tylko jednej z funkcji: $\delta_S^i M(M_{i-1}^F, M^{H_i})$ lub $\delta_O^i M(M_{i-1}^F, M^{H_i})$, gdyż ich wartości są sobie równe.

W pracy [P10] przedstawiono w postaci pseudokodu metody wyznaczania wartości subiektywnej i obiektywnej miary oceny ekspresywności informacji na poziomie wymiarów (Algorytm 1) i na poziomie miar (Algorytm 2).

Oprócz matematycznej analizy własności opracowanych miar, zaproponowane metody zostały przebadane eksperymentalnie (założenia i parametry zawarto w Tabeli 1 [P10]). W trakcie eksperymentu dołączano kolejne hurtownie do federacji i sprawdzano dokładność odpowiedzi federacji na zestaw zapytań. Otrzymane wyniki porównywano z odpowiedziami otrzymanymi ze wzorcowej hurtowni (składającej się ze wszystkich możliwych źródeł wejściowych). Szczegółowo procedura eksperymentu została przedstawiona na Rys. 2 [P10].

W pierwszej części badań wyznaczono odpowiedzi na zapytania za pomocą różnicy między odpowiedziami R_p udzielonymi przez wzorcową hurtownie, a odpowiedziami R_f zwróconymi przez bazową federacje:

$$distf(R_p, R_f) = \frac{\sum_{i=1}^{n} \left(\frac{\sum_{j=1}^{s} \frac{\left(R_p^{ij} - R_f^{ij}\right)}{R_p^{ij}}}{s} \right)}{n} \tag{32}$$

gdzie: R_p – odpowiedź wzorcowej hurtowni, R_f- odpowiedzi bazowej federacji, R_p^{ij} – wartość j-tego elementu i-tego rekordu odpowiedzi wzorcowej hurtowni, R_f^{ij} - wartość j-tego elementu i-tego rekordu odpowiedzi bazowej federacji, *s*- liczba kolumn w pojedynczym rekordzie z odpowiedzi, *n*-liczba rekordów.

W niniejszych badaniach założono, że w pojedynczym rekordzie poszczególne elementy (odpowiedniki różnych kolumn w schemacie) są ciągami znaków, zatem wyrażenie $\frac{(R_p^{ij}-R_f^{ij})}{R_p^{ij}}$ jest zastępowania przez wartość odległości Hamminga między ciągami znaków podzielona liczbę znaków dłuższego słowa.

Następnie po dodaniu nowej hurtowni do federacji bazowej wyznaczono odległość $distf(R_{pn}, R_{fn})$, gdzie R_{fn} - to odpowiedzi udzielone przez federacje bazową wzbogaconą o dodatkową hurtownie, a R_{pn}- to odpowiedzi udzielone przez nową hurtownie wzorcową. Różnica $distf(R_p, R_f)$- $distf(R_{pn}, R_{fn})$ pozwala na ocenę, czy dodanie nowej hurtowni do federacji zwiększa ilość poprawnych odpowiedzi. Wykonano 913 eksperymentów, dla których tylko w 223 przypadkach różnica odległości $distf(R_p, R_f)$- $distf(R_{pn}, R_{fn})$ jest większa od 0. W ramach eksperymentu podjęto się próby znalezienia zależności między zmienną zależną, czyli różnicą odległości $distf(R_p, R_f)$- $distf(R_{pn}, R_{fn})$, a opracowanymi miarami ekspresywności. Wykazano, że spełnione są wszystkie warunki wyznaczenia modelu regresji wielorakiej. Związek pomiędzy zmienną zależną $distf(R_p, R_f)$- $distf(R_{pn}, R_{fn})$, a wszystkimi zmiennymi niezależnymi przedstawia poniższe równanie: *8,15-0,303*sm-0,51*om-0,439*sd-7,47*od+0,26*, gdzie: *sd, od*- oznaczają odpowiednio subiektywną iobiektywną ocenę ekspresywności na poziomie wymiarów, *sm, om*- oznaczają odpowiednio subiektywną i obiektywną ocenę ekspresywności na poziomie miar. Opracowany model jest w miarę dobrze dopasowany, o czym świadczy mały błąd standardowy estymacji $SE_e = 0{,}261$, przeciętna wartość współczynnika determinacji wielorakiej *r*=0,3934 oraz wynik testu F analizy wariancji równej *p-value*<0,000001 (Tabela 2 [P10]). Wartości współczynników korelacji cząstkowej i semicząstkowej wskazują, że jedynie wartości miary *od* mają największy wpływ na budowany model, co potwierdzają wyniki testu *t*-Studenta (Tabela 3 [P10]). Wyniki badania (Tabela 4 [P10]) wykazały, że na podstawie wartości miary *od* można estymować efektywność odpowiedzi federacji na zapytania. Jeśli wartość obiektywnej miary ekspresywności na poziomie wymiarów rośnie, to wartość $distf(R_p, R_f)$- $distf(R_{pn}, R_{fn})$ maleje, czyli pogarsza się odpowiedź federacji na wysyłane do niej zapytania.

Drugi eksperyment był poświęcony porównaniu opracowanych miar z zewnętrzną metryką opartą na algorytmie ACCU [29]. Wykonano 530 eksperymentów. Na podstawie otrzymanych wyników przeprowadzono analizę korelacji wewnątrzklasowej, czyli stopień w jakim wartości wyznaczone przez opracowane miary ekspresywności oraz miarę ACCU są zgodne. Wyniki wykazały, że jedynie subiektywna miara ekspresywności *sd* na poziomie wymiarów koreluje z wartościami otrzymanymi za pomocą algorytmu ACCU. Otrzymane *p-value* mniejsze niż 0,000001 oznacza istotną zależność statystyczną, jednak wartości bezwzględnej zgodności równe 0,24 i 0,14 oraz spójności 0,45 świadczą o niezbyt wysokiej zbieżności obu wyników. Jest to zgodne z intuicją, gdyż miary *sd* i ACCU nie są identyczne. Miara *sd* uwzględnia tylko schematy wymiarów w hurtowniach danych, gdzie algorytm ACCU wymaga bardziej całościowej analizy.

Przeprowadzone badania potwierdziły hipotezę, że dołączenie nowych hurtowni do federacji nie zawsze zwiększa ilość poprawnych odpowiedzi na zapytania. Uwzględniając fakt, że sam proces integracji informacji jest zadaniem czasochłonnym, kosztownym oraz podatnym na potencjalne błędy podczas transmisji źródeł, selekcja najbardziej ekspresywnych hurtowni przed ich dołączeniem do federacji jest kluczowa. Analiza wyników badań wykazała istotną statystycznie zależność liniową między jedną ze zdefiniowanych miar oceny ekspresywności informacji, a ilością prawidłowych odpowiedzi na zapytania. Opracowano model regresji liniowej, który na podstawie zdefiniowanych miar, umożliwia oszacowanie efektywności procesu integracji (bez jej przeprowadzania), mierzony ilością poprawnych odpowiedzi na wysyłane do federacji zapytania.

4.3.4 Podsumowanie oraz kierunki rozwoju

W obecnych czasach, gdy ilość informacji podwaja się co 2-3 lata, ich przetwarzanie jest coraz bardziej złożonym procesem. Integracja informacji i wiedzy jest często bardzo złożona, czasochłonna i kosztowna. W związku z tym, zanim rozpoczniemy proces integracji źródeł wejściowych, chcielibyśmy wiedzieć, czy jest możliwe wyznaczenie jednej, wiarygodnej wersji. Dokładna analiza światowej literatury pokazuje, że wciąż brakuje badań na wskazany temat, dlatego też osiągnięte rezultaty naukowe charakteryzują się dużą nowością. Jest to temat wciąż aktualny, o dużym znaczeniu praktycznym.

Celem prowadzonych prac badawczych było *opracowanie metod, które dla wybranych reprezentacji wiedzy umożliwią ocenę procesu jeszcze przed lub w trakcie procesu ich łączenia.* Osiągnięte rezultaty pozwoliły na ustalenie kompromisu między kosztem przeprowadzonej integracji, a opłacalności z punktu widzenia przyjętych kryteriów. Ponadto na podstawie zaproponowanych twierdzeń sformułowano wnioski pozwalające na wybór lub modyfikacje źródeł wejściowych tak, aby otrzymany rezultat integracji był najlepszy względem przyjętego kryterium efektywności. Wszystkie rezultaty prac były prezentowane na renomowanych konferencjach i publikowane w znanych czasopismach, gdzie były poddane rzetelnym i szczegółowym recenzjom.

W ramach monotematycznego cyklu publikacyjnego postawiono 3 pytania badawcze. W pierwszej części badań weryfikowano, *czy analiza źródeł wejściowych pozwala na ocenę możliwości wyznaczenia wiarygodnego i spójnego wyniku ich integracji* Rezultaty badań pozwoliły na opracowanie warunków, jakie muszą spełniać źródła wejściowe, aby być podatne na consensus (dla zbiorów z powtórzeniami, wektorów binarnych, pokryć oraz podziałów uporządkowanych). Sformułowane kryteria są odpowiedzią na zadane pytanie badawcze, gdyż jeszcze przed procesem integracji umożliwiają ocenę wiarygodności jego wyników.

Kolejne prace badawcze pozwoliły na opracowanie kryteriów, jakie muszą spełniać źródła wejściowe, aby wynik ich integracji charakteryzował się najwyższą lub zadaną jakością. Formalna analiza własności źródeł do integracji takich jak ocena ich spójności, podatności na consensus, regularności czy liczebności pozwoliła na sformułowanie istotnych wniosków, które są odpowiedzią na drugie pytanie badawcze: *czy na podstawie analizy źródeł wejściowych możliwa jest ocena jakości wyniku ich integracji?* W wyniku przeprowadzonych badań sformułowano warunki doboru źródeł wejściowych do integracji tak, aby otrzymany rezultat charakteryzował się najwyższą lub zadaną jakością.

Trzeci etap był próbą odpowiedzi na pytanie *czy możliwa jest ocena efektywności procesu integracji informacji i wiedzy pochodzących z różnych źródeł przed lub w trakcie procesu ich łączenia w jedną, spójną wersję*? W tej części badań przeanalizowano bardziej złożone reprezentacje wiedzy jakim są ontologie i hurtownie danych. Opracowano zestaw narzędzi do integracji i oceny efektywności integracji ontologii zwanego FOKI w ramach, którego opracowano zestaw metod, które pozwalają na ocenę potencjalnego przyrostu wiedzy w procesie integracji ontologii na poziomie: konceptów, relacji, instancji, relacji instancji, opracowano metodę półautomatycznego wyznaczania wartości semantyki atrybutów, opracowano metodę oceny odwzorowań bez konieczności porównywania go z odwzorowaniem wzorcowym. Ponadto opracowano zestaw metod oceny ekspresywności informacji w federacyjnych hurtowniach danych jako kryterium efektywności procesu integracji hurtowni. Dzięki opracowanym metodom, możliwa jest ocena efektywności procesu integracji w trakcie lub przed procesem złączenia źródeł wejściowych.

Na podstawie wszystkich przeprowadzonych badań i analiz można sformułować ogólny wniosek, że na końcowy efekt integracji nie wpływa ilość źródeł wejściowych, ale inne parametry je charakteryzujące.

Wszystkie opracowane metody zostały zaimplementowane, a ich poprawność zweryfikowana eksperymentalnie. Część opracowanych wyników prac badawczych została zaimplementowana w prototypie autorskiego systemu zwanego FOKI. Jest to zestaw metod do integracji ontologii, oceny opłacalności procesu integracji, a także śledzenia ewolucji zachodzących w ontologiach.

Prezentowane, w ramach niniejszego cyklu publikacyjnego wyniki, są punktem wyjścia do kolejnych badań w tej tematyce. Mimo, że wybrane i przebadane reprezentacje wiedzy umożliwiają zamodelowanie większości zjawisk zachodzących w świecie rzeczywistym, to analiza struktur hierarchicznych takich jak np. drzewa przewidziana jest w dalszym etapie badań.

Dalsze prace skupiają się również na opracowaniu metod wyznaczania odwzorowań ontologii. Na efektywność procesu integracji ontologii wpływa nie tylko jakość źródeł wejściowych, ale również jakość odwzorowania między nimi. Opracowanie automatycznych metod, które zwiększają dokładność odwzorowań pomiędzy ontologiami jest kolejnym wyzwaniem badawczym. Pierwsze wyniki zostały zaimplementowane w systemie FOKI, a rezultaty badań zamieszczono w publikacji pt. *Fuzzy based approach to ontology relations alignment* zaakceptowanej na konferencję IEEE-FUZZ 2021

Bibliografia:

[1] M. Brodie: *Data Integration at Scale: From Relational Data Integration to Information Ecosystems.* Proc. of International Conference on Advanced Information Networking and Applications, (2010)

[2] L. Haas, *Beauty and the Beast: The Theory and Practice of Information Integration.* Proc. of International Conference on Database Theory, (2007).

[3] X. Dong F. Naumann, *Data Fusion—Resolving Data Conflicts for Integration.* Proc. of the VLDB Endowment, 2(2), (2009).

[4] N.T Nguyen, *Advanced Methods for Inconsistent Knowledge Management.* Springer London., (2008)

[5] N. T. Nguyen, *Consensus systems for conflict solving in distributed systems*, Information Sciences 147(1–4), pp. 91–122, (2002)
[6] N. T. Nguyen, Inconsistency of knowledge and collective intelligence, Cybernetics and Systems 39(6), pp. 542–562, (2008)
[7] J. Akoka, L. Berti-Equille, L., O. Boucelma, et. all, *A framework for quality evaluation in data integration systems*. In: Proceedings of the 9th International Conference on Entreprise Information Systems, Madeira, Portugal, pp. 12-16, (2007)
[8] M. Ghasemaghaei, M.,G. Calic, *Can big data improve firm decision quality? The role of data quality and data diagnosticity*. Decision Support Systems, 120, pp. 38-49, (2019)
[9] I. Taleb, M. A. Serhani. R. Dssouli, *Big data quality: A survey*. In: 2018 IEEE International Congress on Big Data (BigData Congress), pp. 166-173. IEEE., (2018)
[10] F. Boufares,, A.B. Salem, *Heterogeneous data-integration and data quality: Overview of conflicts*. In: 2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), pp. 867-874, IEEE,(2012)
[11] S. Tartir, I.B. Arpinar, M. Moore, A.P. Sheth and B Aleman-Meza, *OntoQA: Metric-based ontology quality analysis*, (2005): http://works.bepress.com/amit sheth/341/_ [ostatni dostęp: 09.02.2021]
[12] J. Sheoran, *Issues of data quality in data warehouses*, In: Proceedings of the International Conference on Advances in Computer Engineering and Applications, ICACEA, 6, pp. 6–8, (2014)
[13] S. Kumar , U. Kumar, Various data quality issues in data warehousing, Int. J. Res. 1 (10), pp. 210–213, (2014)
[14] B. Heinrich , M. Klier , A. Schiller , G. Wagner, *Assessing data quality-a probability-based metric for semantic consistency*, Decis. Support Syst. 110, pp. 95–106, (2018)
[15] V.D. Nguyen, N. T. Nguyen, *A method for improving the quality of collective knowledge*. In *Asian Conference on Intelligent Information and Database Systems*, pp. 75-84, Springer, Cham, (2015)
[16] M. Gebala , V.D. Nguyen, N.T. Nguyen , *An analysis of influence of consistency degree on quality of collective knowledge using binary vector structure*, Stud. Comput. Intell. 572, pp. 3–13, (2015)
[17] S. Raunich, E. Rahm, *Towards a benchmark for ontology merging*. In: OTM Confederated International Conferences" On the Move to Meaningful Internet Systems", pp. 124-133, Springer, Berlin, Heidelberg, (2012)
[18] M. Mahfoudh, G. Forestier, M. Hassenforder, *A benchmark for ontologies merging assessment*. In: International Conference on Knowledge Science, Engineering and Management, pp. 555-566, Springer, Cham, (2016)
[19] OAEI: http://oaei.ontologymatching.org/ [ostatni dostęp: 09.02.2021]
[20] M. Batista , C.M. Da , A.C. Salgado, *Minimality quality criterion evaluation for integrated schemas*, In: Proceedings of the 2nd International Conference on Digital Information Management, ICDIM, 1, pp. 436–441, (2007)
[21] A. Marotta , R. Ruggia, *Managing source quality changes in a data integration system*, In: Proceedings of the CAiSE (Doctoral Consortium), (2005)
[22] R. Torlone , *Two approaches to the integration of heterogeneous data warehouses*, Distrib. and Parallel Databases 23 (1) pp. 69–97, (2008)
[23] X.L. Dong , B. Saha , D. Srivastava , *Less is more: selecting sources wisely for integration,* In: Proceedings of the VLDB Endowment, 6 (2) pp. 37–48, (2012)
[24] C. Daniłowicz, N.T. Nguyen, *Metody wyboru reprezentacji podziałów i pokryć uporządkowanych*, Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław, (1992)
[25] M. Pietranik and N.T. Nguyen, *Semantic Distance Measure Between Ontology Concept's Attributes*, Proceedings of 15th International Conference, KES 2011 Lecture Notes in Artificial Intelligence, Vol. 6881, Springer, pp. 210–219, (2011)
[26] Resnik, P., *Using information content to evaluate semantic similarity in a taxonomy.* In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, vol. 1, pp. 448–453, (1995)

[27] S. Berger and M. Schrefl. *From federated databases to a federated data warehouse system.* In: Proceedings of the 41st Annual Hawaii International Conference, on System Sciences, HICSS '08, pages 1530-1605, Washington, DC, USA, IEEE Computer Society, (2008)
[28] R. Kern, *Data warehouses federation as a single data warehouse*, In: Proceedings of International Conference on Computational Collective Intelligence, Springer, Cham, pp. 356–366, (2016)

[29] X.L. Dong, L. Berti-Equille, D. Srivastava, *Integrating conflicting data: the role of source dependence*, In: Proceedings of the VLDB, 2, pp. 550–561, (2009)

5. Informacja o wykazywaniu się istotną aktywnością naukową albo artystyczną realizowaną w więcej niż jednej uczelni, instytucji naukowej lub instytucji kultury, w szczególności zagranicznej

Prowadzone przeze mnie prace badawcze powstały w wyniku aktywności naukowej realizowanej na kilku uczelniach zarówno polskich, jak i zagranicznych. Wśród aktywności można wyróżnić:

- pobyty stażowe: Leipzig University, 17.02-1.03.2020
- krótkie pobyty naukowe:
  - Uniwersytet Nguyen Tat Thanh oraz Vietnam National University 11.03-19.03.2016, 16.03-31.03.2018
  - Uniwersytet Ekonomiczny 7.01.2019-15.02.2019
- inne aktywności naukowe.

Pobyty stażowe:

1) **Uniwersytet w Lipsku,** Information System Institute at Leipzig University, Niemcy- staż naukowy 17.02-1.03.2020.

Uniwersytet zajmuje 273 miejsce na świecie w internetowym rankingu Webometrics Ranking of World Universities, rankingu oceniającym naukowe publikacje internetowe uniwersytetów. W rankingu ARWU 2020 Lipsk wspina się na poziom 201–300, natomiast w rankingu QS World University Ranking 2020 znajduje się na pozycji 483. W trakcie pobytu na Uniwersytecie w Lipsku prowadzone były praca badawcze na temat zastosowania głębokich sieci uczenia do rozpoznawania obrazów. Pobyt stażowy został zakończony publikacją:

- Franczyk B., Hernes M., Kozierkiewicz A., Kozina A., Pietranik M., Roemer I., Schieck M., 2020, *Deep learning for grape variety recognition*, Procedia Computer Science. 2020, vol. 176, s. 1211-1220, (**kat. B w rankingu CORE, 70 pkt MNISW**)

Krótkie pobyty naukowe:
W okresie **11.03-19.03.2016 oraz 16.03-31.03.2018 odbyłam, krótkie wyjazdy na poniższe uczelnie.** W ramach pobytu w Wietnamie nawiązałam współpracę z kilkoma naukowcami z obu uniwersytetów, które zaowocowały zarówno licznymi publikacjami naukowymi, a także innymi aktywnościami naukowymi.

1) **Nguyen Tat Thanh University (NTTU) w Ho Chi Minh City w Wietnamie** - wspólne badania naukowe w ramach grantu 4005/0011/17, "Smart University", Nguyen Tat Thanh University, Wietnam uwieńczone wspólną publikacją:

- Hernes M., Kozierkiewicz A., Nguyen T.T., 2019, *An application a two-level determination consensus method in a multi-agent financial decisions support system.* Intelligent Information and Database Systems: 11th Asian Conference, ACIIDS 2019, Yogyakarta, Indonesia, April 8-11, 2019: proceedings. Pt. 1 / Ngoc Thanh Nguyen [i in.] (Eds.). Cham : Springer, cop. 2019. s. 453-463. **(MNiSW (rok 2020): 20pkt, ranking Core (rok 2021): B)**

2) **Vietnam National University w Ho Chi Minh City w Wietnamie**- wspólne badania naukowe uwieńczone wspólnymi publikacjami:

- Huynh B., Nguyen L.T.T , Huynh H.M., Kozierkiewicz A., Yunf U., Kominkova Oplatkova Z., Vo B., 2021, *A novel approach for mining closed clickstream patterns*, Cybernetics and Systems, DOI: https://doi.org/10.1080/01969722.2020.1871225, **(IF: 1.681, MNSiW (2020): 40pkt)**
- Nguyen T.D.D., .Nguyen L.T.T., Kozierkiewicz A., Pham T., Vo B. 2021, An efficient approach for mining high-utility itemsets from multiple abstraction levels, W: Intelligent Information and Database Systems : 13th Asian Conference, ACIIDS 2021, Phuket, Thailand, April 7-10, 2021 : proceedings / eds. Ngoc Thanh Nguyen [i in.]. Cham : Springer, cop. 2021. s. 92-103, (Lecture Notes in Computer Science. Lecture Notes in Artificial Intelligence, ISSN 0302-9743; vol. 12672) **(MNiSW (rok 2020): 20pkt, ranking Core (rok 2021): B)**
- Nguyen H.-L., Nguyen L.T.T., Kozierkiewicz A., 2021, Mining Class Association Rules on dataset with missing data, W: Intelligent Information and Database Systems : 13th Asian Conference, ACIIDS 2021, Phuket, Thailand, April 7-10, 2021: proceedings / eds. Ngoc Thanh Nguyen [i in.]. Cham : Springer, cop. 2021. s. 104-116. (Lecture Notes in Computer Science. Lecture Notes in Artificial Intelligence, ISSN 0302-9743; vol. 12672) **(MNiSW (rok 2020): 20pkt, ranking Core (rok 2021): B)**
- Nguyen T.-N., Nguyen L.T.T, Vo B., Kozierkiewicz A., 2020, *Efficient Method for Mining Maximal Inter-transaction Patterns*, W: Computational Collective Intelligence: 12th International Conference, ICCCI 2020, Da Nang, Vietnam, November 30 - December 3, 2020 : proceedings / eds. Ngoc Thanh Nguyen [i in.]. Cham : Springer, cop. 2020. s. 316-327.(Lecture Notes in Computer Science. Lecture Notes in Artificial Intelligence, ISSN 0302-9743; vol. 12496) **(MNiSW (rok 2020, ranking CORE: C): 20pkt, ranking Core (rok 2021): B)**
- Pham M. Q., Nguyen T.T.S., Do P.M.T., Kozierkiewicz A., 2020, *Incremental SVD-based Collaborative Filtering Enhanced with Diversity for Personalized Recommendation,* W: Advances in Computational Collective Intelligence: 12th International Conference, ICCCI 2020, Da Nang, Vietnam, November 30 December, 2020 : proceedings / eds. Marcin Hernes, Krystian Wojtkiewicz, Edward Szczerbicki. Cham : Springer, cop. 2020. s. 212-223. (Communications in Computer and Information Science, ISSN 1865-0929; vol. 1287) **(MNiSW (rok 2020, ranking CORE: C): 20pkt,**

ranking Core (rok 2021): B)

- Kozierkiewicz A., Pietranik M., Nguyen L.T.T, 2020, *Updating ontology alignment on the instance level based on ontology evolution,* W: Database and Expert Systems Applications: 31st International Conference, DEXA 2020, Bratislava, Czech Republic, September 14-17, 2020: proceedings. Pt 2 / Eds. Sven Hartmann [i in.]. Cham: Springer, cop. 2020. s. 301-311, (Lecture Notes in Computer Science, ISSN 0302-9743; vol. 12392) **(MNiSW (rok 2020): 70pkt, ranking CORE: B)**
- Nguyen T.T.S., Tran T.T.T., Dao T.H.Ch., Kozierkiewicz A., Nguyen T.T.T, *A clustering-based approach to mining sequential large data for web recommendations,* praca zgłoszona do czasopisma Journal of Computer Science and Technology **(IF: 1.185, MNSiW (2020): 70pkt)**

Ponadto, w ramach nawiązanej współpracy z uczelnią Vietnam National University w Ho Chi Minh City oraz dodatkowo z naukowcem z Technical University of Madrid wspólnie zorganizowaliśmy sesję specjalną FMKI 2021: Special Session on **Fuzzy-based Methods for Knowledge Integration** zgłoszona na konferencje **FUZZ-IEEE 2021.** Prestiżowa konferencja znajduje się bardzo wysoko **w rankingu CORE- kat. A.**

3) **Uniwersytet Ekonomiczny we Wrocławiu-** współpraca została rozpoczęta w 2016 roku wspólną organizacją seminarium międzyuczelniane pt. *Inteligencja Kolektywna w Systemach Informacyjnych.* Od 2018 w ramach Centrum Inteligentnych Systemów Zarządzania zainicjowano poszukiwanie partnerów naukowych również z otoczenia gospodarczo-biznesowego oraz wspólne badania naukowe w ramach rozpoczętych projektów. Efektem nawiązanej współpracy był **krótki pobyt naukowy w dniach 7.01.2019-15.02.2019 na Uniwersytecie Ekonomicznym we Wrocławiu**, który rozpoczął prace koncepcyjne nad projektem w ramach programu "Dolnośląski bon na innowacje", Dolnośląska Agencja Rozwoju Regionalnego, U/0180/82/2019, 02-07.2019, *Opracowanie prototypu doświadczalnego modułu automatycznej oceny zdolności kredytowej klienta dla instytucji finansowej, z wykorzystaniem narzędzi sztucznej inteligencji.*

Wspólne badania naukowe uwieńczone wspólnymi publikacjami:

- Hernes M., Kozierkiewicz A., Pietranik M., Sobieska-Karpińska J., 2018, *A new distance function for consensus determination in decision support systems.* Computational collective intelligence : 10th International Conference, ICCCI 2018, Bristol, UK, September 5-7, 2018: proceedings. Pt. 1 / Ngoc Thanh Nguyen [i in.] (Eds.). Cham : Springer, cop. 2018. s.155-165. **(MNiSW (rok 2020, ranking CORE: C): 20pkt, ranking Core (rok 2021): B)**
- Hernes M., Chojnacka-Komorowska A., Kozierkiewicz A., Pietranik M., 2018, *Agents' knowledge conflicts' resolving in cognitive integrated management information system - case of budgeting module.,* Computational collective intelligence: 10th International Conference, ICCCI 2018, Bristol, UK, September 5-7, 2018: proceedings. Pt. 1 / Ngoc Thanh Nguyen [i in.] (Eds.). Cham: Springer, cop. 2018. s. 297-308. **(MNiSW (rok 2020, ranking CORE: C): 20pkt, ranking Core (rok 2021): B)**
- Hernes M., Kozierkiewicz A., Pietranik M., 2019, *The agents' selection methods for a*

consensus-based investment strategy in a multi-agent financial decisions support system. W: New Trends in Databases and Information Systems, ADBIS 2019 Short Papers, Workshops: BBIGAP, QAUCA, SemBDM, SIMPDA, M2P, MADEISD and Doctoral Consortium: Bled, Slovenia, September 8-11, 2019: proceedings / eds. Tatjana Welzer [i in.]. Cham: Springer, cop. 2019. s. 60-67. **(MNiSW (rok 2020): 70pkt, ranking CORE: B)**

- Hernes M., Kozierkiewicz A., Nguyen T.T. 2019, *An application a two-level determination consensus method in a multi-agent financial decisions support system.* Intelligent Information and Database Systems: 11th Asian Conference, ACIIDS 2019, Yogyakarta, Indonesia, April 8-11, 2019: proceedings. Pt. 1 / Ngoc Thanh Nguyen [i in.] (Eds.). Cham: Springer, cop. 2019. s. 453-463. **(20 pkt MNISW, ranking Core (rok 2021): B)**
- Hernes M., Kozierkiewicz A., Maleszka M., Rot A., Kozina A., Janus J., Wróbel E., *Deep learning for repayment prediction in leasing companies*, praca zgłoszona do czasopisma Algorithmic Finance **(MNiSW (rok 2020): 40pkt)**

Inne aktywności współpracy naukowej:

1) **Politechnika Śląska-** recenzje w czasopiśmie Studia Informatica, wspólne badania naukowe ukończone publikacją:

- Jureczek P., Kozierkiewicz-Hetmańska A., (2014), *Privacy-preserving framework for mining continuous seguences in trajectory systems*, 2014 European Network Intelligence Conference, ENIC 2014 [Dokument elektroniczny] : Wroclaw, Poland, 29-30 September 2014 : proceedings. [Piscataway, NJ]: IEEE, cop. 2014. s. 52-58.

2) **3LO we Wrocławiu** - Luty-czerwiec 2015- zajęcia projektowe z uczniami z 3 LO we Wrocławiu w ramach projektu Uniwersytet Młodych Wynalazców zakończony publikacją i prezentacją 07.09.2015 w ramach Warsztatów Młodych Informatyków.

- Kozierkiewicz-Hetmańska A., 2015, *System wieloagentowy przetwarzający opinię ekspertów*, Raport serii SPR nr 4, 8s.

3) **Państwowa Wyższa Szkoła Zawodowa w Nysie**, Październik 2010, przeprowadzenie badań eksperymentalnych

4) **Budapesti Műszaki és Gazdaságtudományi Egyetem** - współpraca 2013/14, przygotowanie wniosku do programu unijnego 7PR ICT Call 10

5) **KES International, Londyn**, Wielka Brytania,- 08.2010-01.2011 - współpraca przy organizacji konferencji naukowych

6. Informacja o osiągnięciach dydaktycznych, organizacyjnych oraz popularyzujących naukę lub sztukę

6.1 Podręczniki i inne pomoce dydaktyczne

- Opracowanie materiałów dydaktycznych dla kursu "*Projektowanie baz danych*" w ramach projektu „ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej", nr WND-POWR.03.05.00-00-Z301/17
- Opracowanie karty przedmiotu oraz materiałów do nowego kursu "*Techniki inteligencji obliczeniowych*" (wykład i ćwiczenia laboratoryjne). W ramach prac przygotowano zarówno instrukcje do laboratoriów jak i szczegółowe materiały pomocnicze do prowadzenia wykładów (prezentacje do każdego wykładu oraz testy ewaluacyjne).
- Opracowanie materiałów do laboratorium kursu „*Zaawansowane metody analizy danych*" w których zawarto szczegółowe opis zadań do wykonania przez studentów oraz wstęp teoretyczny z przykładami

6.2 Opieka nad pracami inżynierskimi i magisterskimi

Pod moim nadzorem powstały oraz zostały obronione liczne prace inżynierskie i magisterskie:
Prace inżynierskie: 29
Prace magisterskie 30
Recenzje: 29
Promotor pomocniczy w otwartych przewodach doktorskich: 2
Prace naukowe powstałe w wyniku pracy ze studentami i doktorantami: 13

1. Kozierkiewicz-Hetmańska A., Zyśk D., 2013, *A method for determination of an opening learning scenario in intelligent tutoring systems*, W: Intelligent Information and Database Systems: 5th Asian Conference, ACIIDS 2013, Kuala Lumpur, Malaysia, March 18-20, 2013 : proceedings. Pt. 2 / Ali Selamat, Ngoc Thanh Nguyen, Habibollah Haron (eds.). Berlin ; Heidelberg : Springer, cop. 2013. s. 129-138. (Lecture Notes in Computer Science. Lecture Notes in Artificial Intelligence, ISSN 0302-9743; vol. 7803) **(MNiSW (rok publikacji): 15pkt, MNiSW(rok 2020): 20pkt, ranking Core (rok 2021): B)**
2. Kozierkiewicz-Hetmańska A., Poniatowski R., 2014, *An item bank calibration method for a computer adaptive test*, W: Intelligent Information and Database Systems: 6th Asian Conference, ACIIDS 2014, Bangkok, Thailand, April 7-9, 2014: proceedings. Pt. 1 / Ngoc Thanh Nguyen [i in.] (Eds.). Cham [i in.]: Springer, cop. 2014. s. 375-383. (Lecture Notes in Computer Science. Lecture Notes in Artificial Intelligence, ISSN 0302-9743; vol. 8397) **(MNiSW (rok publikacji): 15pkt, MNiSW(rok 2020): 20pkt, ranking Core (rok 2021): B)**
3. Bernacki J., Kozierkiewicz-Hetmańska A., 2014, *Creating Collaborative Learning Groups in Intelligent Tutoring Systems, Computational Collective Intelligence*. W: Computational collective intelligence: technologies and applications : 6th International Conference, ICCCI 2014, Seoul, Korea, September 24-26, 2014: proceedings / Dosam Hwang, Jason J. Jung, Ngoc Thanh Nguyen (eds.). Cham [i in.]: Springer, cop. 2014. s.

184-193. (Lecture Notes in Computer Science. Lecture Notes in Artificial Intelligence, ISSN 0302-9743; vol. 8733) **(MNiSW (rok publikacji): 15pkt, (MNiSW (rok 2020, ranking CORE: C): 20pkt, ranking Core (rok 2021): B)**

4. Bernacki J., Kozierkiewicz-Hetmańska A., 2015, *A conception for use of user profile to prediction learning effects in Intelligent Tutoring Systems*, The Second European Network Intelligence Conference, ENIC 2015 [Dokument elektroniczny]: 21-22 September 2015, Karlskrona, Sweden: proceedings. Piscataway, NJ : IEEE, cop. 2015. s. 140-144. **(MNiSW (rok publikacji): 15pkt, MNiSW(rok 2020): 20pkt)**
5. Bernacki J., Kozierkiewicz-Hetmańska A., 2015, *The comparison of creating homogeneous and heterogeneous collaborative learning groups in intelligent tutoring systems*. W: Intelligent Information and Database Systems: 7th Asian Conference, ACIIDS 2015, Bali, Indonesia, March 23-25, 2015: proceedings. Pt. 1 / Ngoc Thanh Nguyen, Bogdan Trawiński, Raymond Kosala (eds.). Cham [i in.]: Springer, cop. 2015. s. 46-55.(Lecture Notes in Computer Science. Lecture Notes in Artificial Intelligence, ISSN 0302-9743; vol. 9011) **(MNiSW (rok publikacji): 15pkt, MNiSW(rok 2020): 20pkt, ranking Core (rok 2021): B)**
6. Kozierkiewicz-Hetmańska A., Marciniak A. , Pietranik M., 2016, *User authentication through keystroke dynamics*, Computational Collective Intelligence: 8th International Conference, ICCCI 2016, Halkidiki, Greece, September 28-30, 2016: proceedings. Pt. 1 / Ngoc Thanh Nguyen, Yannis Manolopoulos, Lazaros Iliadis, Bogdan Trawiński (Eds.). [Cham]: Springer, cop. 2016. s. 345-355, **(MNiSW (rok publikacji): 15pkt, (MNiSW (rok 2020, ranking CORE: C): 20pkt, ranking Core (rok 2021): B)**
7. Kozierkiewicz-Hetmańska A., Marciniak A., Pietranik M., 2016, *Multilevel verification of user authentication framework using keystroke dynamics*, Computational Collective Intelligence: 8th International Conference, ICCCI 2016, Halkidiki, Greece, September 28-30, 2016: proceedings. Pt. 1 / Ngoc Thanh Nguyen, Yannis Manolopoulos, Lazaros Iliadis, Bogdan Trawiński (Eds.). [Cham]: Springer, cop. 2016. s. 379-355, **(MNiSW (rok publikacji): 15pkt, (MNiSW (rok 2020, ranking CORE: C): 20pkt, ranking Core (rok 2021): B)**
8. Kozierkiewicz A., Sitarczyk M., 2017, *The efficiency analysis of the multi-level consensus determination method.,* Computational Collective Intelligence: 9th International Conference, ICCCI 2017, Nicosia, Cyprus, September 27-29, 2017: proceedings. Pt. 1 / Ngoc Thanh Nguyen [i 5 in.] (Eds.) Cham: Springer, cop. 2017. s.103-112, **(MNiSW (rok 2020, ranking CORE: C): 20pkt, ranking Core (rok 2021): B)**
9. Kozierkiewicz A., Sitarczyk M., 2018, *Heuristic algorithms for 2-optimality consensus determination.,* Intelligent Information and Database Systems: 10th Asian Conference, ACIIDS 2018, Dong Hoi City, Vietnam, March 19-21, 2018: proceedings. Pt. 1 / Ngoc Thanh Nguyen [i in.] (Eds.). Cham: Springer, cop. 2018. s. 48-58. **(MNiSW (rok 2020): 20pkt, ranking Core (rok 2021): B)**
10. Kozierkiewicz A., Pietranik M., Sitarczyk M., 2020, *Assessing the influence of conflict profile properties on the quality of consensus*. W: Intelligent Information and Database Systems: 12th Asian Conference, ACIIDS 2020, Phuket, Thailand, March 23-26, 2020: proceedings. Pt. 1 / eds. Ngoc Thanh Nguyen [i in.] Cham: Springer, cop. 2020. s. 25-36. (Lecture Notes in Computer Science. Lecture Notes in Artificial Intelligence, ISSN 0302-9743; vol. 12033). **(MNiSW (rok 2020): 20pkt, ranking Core (rok 2021): B)**

11. Kozierkiewicz A., Pietranik M., Wesołowski M., 2020, *Assessing ontology mappings on a level of concepts and instances,* IEEE Access 8: 174845-174859, **(IF: 3.745, MNSiW(2020): 100pkt)**
12. Bielaniewicz J., Kozierkiewicz A., 2020, *A hybrid method of facial expressions recognition in the pictures,* W: Computational Collective Intelligence: 12th International Conference, ICCCI 2020, Da Nang, Vietnam, November 30 - December 3, 2020 : proceedings / eds. Ngoc Thanh Nguyen [i in.]. Cham : Springer, cop. 2020. s. 581-590. (Lecture Notes in Computer Science. Lecture Notes in Artificial Intelligence, ISSN 0302-9743; vol. 12496) **(MNiSW (rok 2020, ranking CORE: C): 20pkt, ranking Core (rok 2021): B)**
13. Kozierkiewicz, A.; Pietranik, M. and Kania, K., 2021, *A Method for Estimating Potential Knowledge Increase after Updating Ontology Mapping.* In Proceedings of the 16th International Conference on Evaluation of Novel Approaches to Software Engineering - Volume 1: ENASE, ISBN 978-989-758-508-1, s. 173-180. DOI: 10.5220/0010362001730180 **(MNiSW (rok 2020): 70pkt, ranking Core: B)**

6.3 Inne aktywności dydaktyczne

- od 29 kwietnia 2020 - (nadal trwa) - członek Komisji Programowej kierunku "Sztuczna inteligencja" na Wydziale Informatyki i Zarządzania Politechniki Wrocławskiej
- recenzent 34. i 36. edycji konkursu Ogólnopolskiego Konkursu PTI na najlepsze prace magisterskie z informatyki

7. Inne informacje dotyczące kariery zawodowej

- 06.2020 - (nadal trwa) - ekspert oceniający wnioski Departament Programów Wsparcia Innowacji i Rozwoju Ministerstwo Funduszy i Polityki Regionalnej
- 05.2019 - (nadal trwa) - ekspert oceniający wnioski w Fundacji na rzecz Nauki Polskiej
- 03.2019 - (nadal trwa) - ekspert oceniający wnioski dla Instytutu Wymiaru Sprawiedliwości
- 07.09.2018 - (nadal trwa) - ekspert oceniający wnioski dla Narodowej Agencji Wymiany Akademickiej
- 05.10.2017 - (nadal trwa) - ekspert oceniający wnioski dla Narodowego Centrum Badań i Rozwoju

Achiama Kozierkiewicz