

Т.Т. Чинибаеваның 6D070400 – Есептеу техникасы және бағдарламалық қамтамасыз ету мамандығы бойынша философия докторы (PhD) дәрежесін алуға ұсынылған «Гетерогенді құрылымы бар деректерді басқаруға арналған үлгілер мен әдістер (Big Data)» тақырыбындағы диссертациялық жұмысына

АНДАТПА

Жұмыстың жалпы нәтижесі. Қазіргі кезеңдегі қоғам мен технологиялардың дамуы адам қызметінің жаңа бағыттарын ақпараттандырумен ғана емес, сонымен бірге құзыретті басқару шешімдерін әзірлеу үшін деректерді зерттеу және талдау технологияларын кеңінен енгізумен де байланысты.

Бұл мәселені бүкіл әлемде, атап айтқанда Қазақстанда дамытуға көп көңіл бөлінеді. Еліміздің цифрлық дамуының негізгі бағыттарын айқындайтын маңызды құжат 2017 жылғы 12 желтоқсанда қабылданған «Цифрлық Қазақстан» мемлекеттік бағдарламасы дәлел. Осы жобаның төлқұжатында мәліметтер көлемінің едәуір артуына байланысты үлкен деректі талдау технологиялық орталығын құрып, және ұлттық, мемлекеттік ақпараттық ресурстарының сенімді жұмыс істеуін, сақталуын, бүтіндігін, оның ішінде қолданыстағы бастамалар негізінде қамтамасыз етуде мемлекет тарапынан жәрдем болады делінген.

Өңдеу үрдісін жылдамдататын жоғары өнімділікті есептеу жүйелері мәліметтерді талдау арқылы алынатын қажетті білім жеткіліксіз. Жүйе архитектурасымен сипаттамасын құрған кезде өзара үйлесімділік мәселесі тысқара қалып қояды. Электронды ақпараттық қорды стандарттау, ақпараттық қорды біріктіруде іздеу дәлдігін және толықтығын арттыруда пайдаланатын сәйкес құралдардың келісімі әлсіз қолданылады.

Зерттеу тақырыбының өзектілігі ғылыми ұйымдардың қызметін сипаттайтын ақпараттарды бақылау мен талдау үшін қолданылатын гетерогенді құрылымы бар үлкен деректерді басқарудың үлгілері мен әдістерін ұсынумен анықталады.

Зерттеу тақырыбының ғылыми зерттелу деңгейі. Соңғы бес-он жылда үлкен деректер мәселесімен байланысты зерттеуге алыс және жақын шетел ғалымдары өз үлесін қосуда. Атап айтсақ: А.Ф. Тузовский, Л.В. Найханова, А.Н. Бездушный, В.А. Серебряков, И.С. Михайлов, Ю.А. Загоруйко, Ditmar Bayer, К.И. Шахгельдян, N. Guarino, N. Noy, M. Ehrig, A. Maedche, Yuong Im Cho және отандық авторлар: А.А. Қуандықов, Р.К. Ускенбаева, Н.К. Мукажанов, Т.Г. Балова, А.А. Шарипов.

Өртүрлі ақпарат дереккөздерінен жиналған гетерогенді құрылымы бар деректерді біріктіру мәселесінің ғылыми сарапмасының қорытындысы бойынша аталмыш пәндік аймақтағы білім жүйеленбеген, әрі үлкен деректерді басқарудағы тәсілдемесінде бірыңғай әдістің жоқтығы ғылыми зерттеудің құрылымын, мақсатын және тапсырмасын анықтады.

Зерттеу жұмысының мақсаты – берілген ғылым саласындағы ғалымдардың жарияланымдарын сипаттайтын ақпаратты іздеу, жүйелеу, сақтау, аннотациялау және талдаудың математикалық үлгілерін, алгоритмдерін және құжат корпусын қолданып бағдарламалық жасақтамасын әзірлеу.

Зерттеу жұмысында қойылған мақсатқа келесі тапсырмалар қатарын шешу арқылы қол жеткізіледі:

1. Қолданыстағы ғылыми ақпаратты басқаруға арналған әдістер мен тәсілдеріне зерттеу жұмысын жүргізу.

2. Мәтіндер коллекциясынан берілген терминдерді тақырып бойынша бөліп іздеп табу алгоритмін әзірлеу және қолдану.

3. Ғылыми білім саласының онтологиясын құрастыру алгоритмін әзірлеу.

4. Жүйе архитектурасын дамыту, қолданыстағы жүйелермен интеграциялау мүмкіндігімен тиімді жұмыс істеуін қамтамасыз ету.

5. Әзірленген алгоритмдердің нәтижелерін тестілеу және тәжірибелік бағалау.

6. Дайындалған алгоритмдердің нәтижелерін бейнелеу.

Зерттеу объектісі гетерогенді құрылымы бар ғылыми мәліметтер болып табылады.

Зерттеу пәніне құжаттарға семантикалық өзара үйлесімділікті қамтамасыз ету мақсатында гетерогенді құрылымы бар деректерді басқаратын үлгілер мен әдістер жатады.

Зерттеу әдістері. Зерттеу барысында қойылған тапсырмалар табиғи тілмен мәтінді талдау, классификация және бағдарламалық инженерия әдістерімен шешілді. Нәтижелер математикалық статистика және математикалық логика аппаратымен ұсынылды.

Диссертациялық жұмыстың **ғылыми жаңалығы** ғылыми конференция хабарландыруларынан және ғаламтордағы іздеу жүйелерін қолдана отырып интеллектуалды жүйе прототипінде ізденуші жасаған тақырып бойынша қойылған талаптарды қанағаттандыратын жұп–терминдерді мәтін жинақтамаларынан бөлу және ғылыми білімдердің жеке саласының онтологиясын құру алгоритмдері, онтология және SPARQL тілін пайдалана отырып жасалынған жүйе сұраныстардың ресми сипаттамасы болып табылады.

Диссертациялық жұмысты қорғауға келесі нәтижелер ұсынылады:

- ғылыми ұйымның қызметінің нәтижесін сипаттауда пәндік аймақ зерттеуінің нәтижесі негізінде сұраныс ақпаратын: талдау, беру, толтыру және сақтау жүйесіне онтологияны қолдана отырып әзірленген математикалық үлгі мен алгоритм, технологиялық және архитектуралық шешім;

- сұраныстарды есептеуде және жүйе кодының барлық өмірлік кезеңдерінде тиімді тексеруде қосымша мүмкіндіктер кепілдігін беретін онтология және SPARQL тілін пайдалана отырып жасалынған жүйе сұраныстардың ресми сипаттамасы;

- ғылыми білімдердің жеке саласының онтологиясын құру және тақырып бойынша қойылған талаптарды қанағаттандыратын жұп–терминдерді мәтін жинақтамаларынан бөлу алгоритмдері;

- математикалық үлгілерді қолдану арқылы қалыптастырылған бағдарламалық қамтаманың күрделілігіне аналитикалық баға беру.

Жұмыстың теориялық және тәжірибелік маңыздылығы: жүргізілген ғылыми зерттеудің ғылыми жаңалығы мен тәжірибелік маңызы жоғарғы

деңгейге ие. Зерттеу барысында алынған нәтижелер гетерогенді деректерді біріктіруде, оларды ары қарай өңдеу мақсатында қолдануға арналады.

Ғылыми жетістіктің шынайылығы зерттеу барысында түрлі әдістерді қолдана отырып алынған нәтижелерді ғылыми әдебиетте берілген нәтижелермен салыстырып, теориялық есептеулер тәжірибе нәтижелерінің сәйкестігімен расталды.

Автордың жеке үлесі. Диссертацияда баяндалған теориялық және қолданбалы зерттеулердің негізгі ғылыми нәтижелері, қорытындылары ғылыми кеңесшілердің басшылығымен автор өзбетімен алды. Бірлескен авторлық жарияланымдардың басым бөлігінде диссертациялық жұмыста қойылған тапсырмалармен байланысты үлгілер мен алгоритм, бағдарламалық жүзеге асыру мен эксперименттік зерттеуге қатысты нәтижелер ізденушіге тиесілі.

Диссертациялық жұмыстың апробациясы және жарияланымдары. Диссертацияның негізгі нәтижелері халықаралық ғылыми конференцияларда баяндалды және талқыланды: The 14th International Conference on Control, Automation and Systems, ICCAS 2014 (Оңтүстік Корея, Бусан, 2014); The 34th Chinese Control Conference and SICE Annual Conference 2015 (CCC&SICE2015) (Қытай, Хангжоу, 2015); The 15th International Conference on Control, Automation and Systems, ICCAS 2014 (Оңтүстік Корея, Гуанджоу, 2015).

Диссертациялық жұмыс Халықаралық ақпараттық технологиялар университетінің «Компьютерлік инженерия және ақпараттық қауіпсіздік» кафедрасының және Гачон университетінің «Компьютерлік инженерия» факультетінің (South Korea, Seoul) семинарларында талқыланды.

Диссертациялық жұмысты орындау барысында алынған нәтижелер 12 ғылыми еңбектерде жарияланды, олардың ішінде 5 мақала Қазақстан Республикасы Білім және ғылым министрлігінің Білім және ғылым саласындағы бақылау комитеті ұсынған журналдарда, 7 мақала халықаралық ғылыми-тәжірибелік конференцияларында жарияланған. Scopus компаниясының деректер базасына кіретін халықаралық ғылыми басылымда 1 мақала жарияланған (процентиль 36%).

Диссертациялық жұмыстың құрылымы мен көлемі. Диссертация құрылымы кіріспе, төрт бөлім, қорытынды, пайдаланылған әдебиет тізімі және қосымшадан тұрады. Жұмыстың жалпы көлемі 119 беттен тұрады, соның ішіне 40 сурет, 15 кесте, 74 атаудан тұратын әдебиет тізімі, 2 қосымша кіреді.

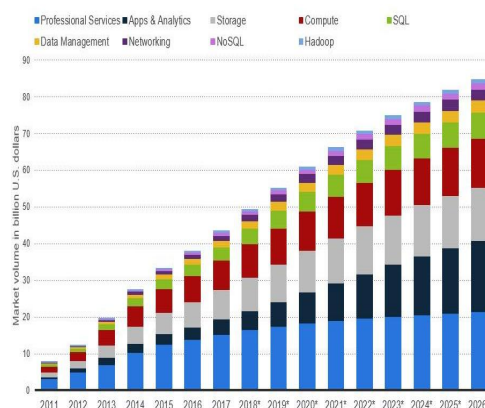
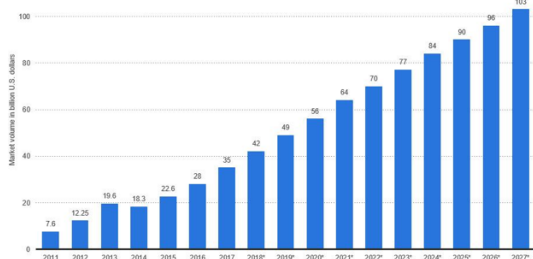
Кіріспеде пәндік сала бойынша қысқаша шолу жасалынып, саланың негізгі мәселелері келтірілді. Диссертациялық жұмыстың маңыздылығы негізделіп, мақсаты мен талаптары қалыптастырылды.

Бірінші бөлімде үлкен деректер технологиясын заманауи жай-күйі мен нарықта алатын орнын көңіл бөлінді.

Әлемдік үлкен деректер нарығында бағдарламалық жасақтама мен қызметтерді сатудан түсетін кіріс 2018 жылы 42 миллиард доллардан 2027 жылы 103 миллиард долларға дейін өсіп, GAGR жылдық өсу қарқыны 10,48% жетеді (сурет 1).

Forecast Revenue Big Data Market Worldwide 2011-2027

Big Data Market Size Revenue Forecast Worldwide From 2011 To 2027
(in billion U.S. dollars)



Сурет 1 – Үлкен деректер нарығының болжамы

Үлкен деректер технологиясы ғылыми ақпаратты басқаруда айрықша орын алады. Ұқсас мәселелерді шешу үшін пайдаланылатын және Интернетте ұсынылған ақпараттық жүйелердің талдауы жүйелердің бірнеше тобын анықтауға мүмкіндік берді, олардың көпшілігі библиографиялық және дерексіз мәліметтер базасы, атап айтқанда Web of Science, Scopus, Google Scholar, ресейлік портал eLibrary.ru. Олар бір немесе екі дәрежеде индекстеу және зерттеу жұмыстары сияқты функцияларды біріктіреді. Жүйелердің бір бөлігі, мысалы М.В. Ломоносов атындағы Мәскеу мемлекеттік университетінің «ИСТИНА» жүйесі, Астрахан мемлекеттік университетінің «Ғылыми қызмет нәтижелері» ақпараттық -аналитикалық жүйесі, Elsevier-дің PURE жүйесі ұйымның ғылыми қызметі мен өнімділігін бағалауға кеңінен мониторинг жүргізеді. Әлемдегі ғылыми ақпараттарды басқаруда қолданылатын ірі веб қызметтерінің сипаттамаларының салыстырмасы кесте 1 берілген.

Бірінші тараудың соңында ғылыми мәліметтерді өңдеу мен талдаудың қазіргі кезде белгілі жүйелерінің негізгі кемшіліктері келтірілген. Бұл кемшіліктерге мыналар жатады: мәліметтерді енгізудің күрделілігі; ақпаратты іздеудің күрделілігі мен төмен қабілеттілігі; білім саласының қатаң және ақпаратсыз модельдерін қолдану, жүйелердің икемділігінің болмауы назар аударылды.

Зерттеудің нәтижесі гетерогенді құрылымы бар ғылыми ақпаратты өңдейтін интеллектуалды бағдарламалық кешеннің прототипінде көрсетілді.

Кесте 1 – Ірі веб қызметтерінің сипаттамаларының салыстырмасы

	№	Атауы	Құзіретті орган	Артықшылығы	Кемшілігі	Деректер форматы
Ірі веб қызмет	1	Web of Science	Thomson Scientic	Жүйеде 1900 жж басталатын мақалалар бар	Сұраныс кілт сөзбен ғана орындалады	.TXT
	2	Scopus	Elsever	Жалпы пәндік аумақты қамтиды	Сұраныс кілт сөзбен ғана орындалады	.TXT
	3	Google Scholar	Google	Баспаға қабылданған, бірақ жарыққа шығып үлгермеген мақалалар есепке алынады	сапасыз және жалған ғылыми жарияланымдар кездеседі	.TXT
Шет елдік проект	1	Bibster	Карлсруэ университеті, Амстердам университеті, Dresden Bank	Жүйеден деректерді RDF форматында шығарады	Жүйеге ақпарат құрылымдалған файл арқылы жүктеледі	BibTeX
	2	JeromeDL	Гданьска (Польша) Технологиялық университет, DERI (Ирландия) сандық технологияларын зерттеу Институты	Жүйе қордағы электронды ақпаратты жіктей және мазмұндай алады	Жүйеге ақпарат құрылымдалған күйде немесе қолмен енгізіледі. күрделі сұраныстар орындалмайды, ақпарат қолмен енгізіледі	BibTeX, Marc21, Dublin Core
	3	Flink	Амстердам университеті	Кілт сөз негізінде ғалымдардың интерфейстік аумағын анықтайды	Кажет пәндік аумақтағы онтологияны қолмен құрастырады	FOAF, SWRC
	4	AIR	Вульверхэмтон (Ұлыбритания) мен Аликанте (Испания) университеті	Жүйе DC құрылымдағы ақпаратты веб парақшаларынан ақпаратты жинайды	Пән аймағын үлгілейтін күрделі онтология жоқ	Dublin Core
СДҚ		Семантикалық дерекқор	Open Sourse	Кез-келген бағдарлама жасақтаушыға қол жетімді	Логикалық байласты қамтамасыз ету	RDF(s), OWL, SPARQL
Ресейлік жүйе	1	«ИСТИНА»	Ресей, Мәскеу	Кез-келген бағдарлама жасақтаушыға қол жетімді	Логикалық байласты қамтамасыз ету	RDF(s), OWL, SPARQL
	2	Астрахан университетінің «ғылыми қызметінің нәтижелері»	Ресей, Астрахан	Кез-келген бағдарлама жасақтаушыға қол жетімді	Логикалық байласты қамтамасыз ету	RDF(s), OWL, SPARQL

Екінші бөлімде ғылыми ақпаратты автоматты басқару жүйесі кезінде қолданылған архитектура-технологиялық шешімдер көрсетілген.

Жалпы семантикалық ақпараттық жүйе 5 логикалық деңгейге бөлінген (сурет 2). Берілген архитектурада:

1) деректердің шығу деңгейі ақпарат беруші жерлердің барлығынан алынады, мысалы семантикалық ДҚ, деректер базасы, түрлі форматтағы файлдар, Web-қызмет, URI қол жетімді кез-келген сыртқы онтология, және т.б

2) деректердің қол жетімділігі деңгейінде форматтарды ажырату мен іске асыру ерекшеліктерін жасыруға мүмкіндік беретін абстрактілену іске асады.

3) базалық онтологиялық қызметтердің деңгейі келесідей құрамнан тұрады:

- Онтологияны тіркеу қызметі (жариялыным, ізденіс);

- Онтологиямен жұмыс жасайтын қызмет, мысалы элементтерді алу мен оларды өзгерту;

- Деректерді алу үшін жіберілетін сұраныс қызметі;

- Логикалық қортынды қызметі;

- Онтология мен жұмыс журналын жүргізу қызметі.

- Семантикалық ДҚ қауіпсіздігін қолдау қызметі. Семантикалық ДҚ қауіпсіздігін қолдау қызметі семантикалық ақпараттық жүйелердің жұмыс істеу үрдісінде аса маңызды рөл атқарады. Олар мынандай мүмкіндіктерге ие болуы керек:

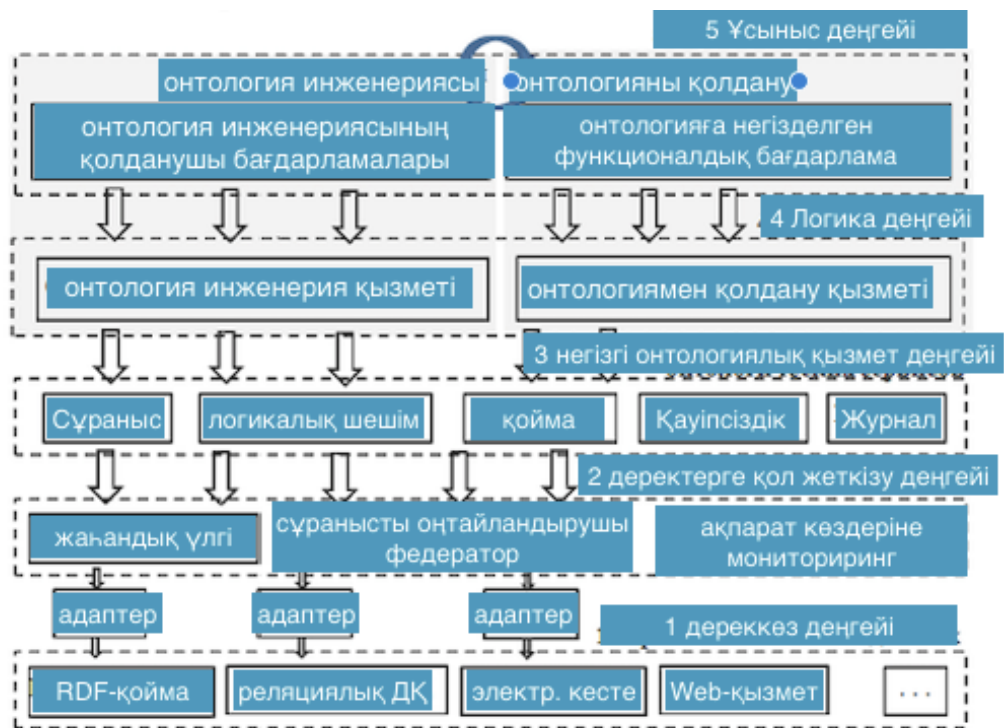
- Қоймада сақталынған деректерге қолданушылардың қол жетімділігін қамтамасыз ету;

- ДҚ сұранысын орындауды бақылау;

- Жіберілген сұраныстарды орындау кезінде алынған логикалық қорытындыларды қамтамасыз ету.

4) Логика деңгейі онтология инженерия қызметі мен базалық операцияларды орындау қызметінен, онтология мен мета деректерден тұрады (түсініктер экземпляры), олар қолданылатын нұсқалар үшін өте тиімді болып келеді (өңдеу логикасын қосады) және нақтылы объектілі үлгілермен жұмыс істейді (нақтылы деректермен). Берілген қызметтер таңдауды орындау мен семантикалық деректерді басқару үшін онтологияның өмір сүру циклын туғызады (ontology lifecycle services).

5) таныстыру деңгейі қолданушының графиктік интерфейсінен тұратын компоненттерден тұрады, мысалы web-қосымша мен desktop-қосымша, бұлар логика деңгейіндегі компоненттермен өзара әрекет етеді.



Сурет 2 – семантикалық деректер негізіндегі ақпараттық жүйенің жалпы құрамы

Автоматтандырылған жүйенің негізгі прототипі болып табылатын ғылыми-техникалық ақпараттың қорытындысы мен күрделі ұймдастырған есептеу үрдісінің жалпы формальды үлгісі.

D ғылыми білім аймағы берілді деп есептейік (мысалы, Computer science). I – білімнің осы аймағының шеңберіндегі ғылыми-техникалық ақпараттың бірліктерін сипаттау жиыны болсын (атомдық өлшем). Бұндай бірліктерге мыналар жатады: ғылыми мақалалар; патенттер; есептер; конференцияларда оқылатын есеп; есепке арналған тезистер; монографиялар; білім беруге арналған көмекші құрал мен басқа авторлық туындылар (рефераттар, аудармалар). I жиынының әрбір элементі сәйкес келетін объектінің кейбір мәтіндік сипаттамасынан тұрады.

Жүйенің негізгі мақсаты іздеу-аналитикалық сұранысының орындалуы болып табылады. Типтік сұраныстардың жиынын Q белгісімен белгілейміз. Тапсырма $I_q \subseteq I$ ғылыми-техникалық ақпарат бірлігінің сипаттамаларымен $q \in Q$ $r1: Q \rightarrow 2^I$ көрінісі арқылы беріледі.



Сурет 3 – операциялар тізбегі

Жүйенің жалпы сызбасы сурет 3 көрсетілген және олар мынандай үлгілерден тұрады:

- Білімнің осы аймағына бағытталған ғылыми-техникалық конференцияны мәтіндік сипаттаудан D ғылыми білім аймағын сипаттайтын терминдерді ерекшелеп алу;
- D ғылыми білімнің аймағындағы қарастырылатын онтологияны құрау;
- Жұмысшылардың ғылыми негізінің нәтижесі жайлы деректерді жүктеу;
- Білім аймағында құрастырылған экземпляр мен ғылыми салалардың нәтижелері жайлы жүктелген ақпараттардың арасындағы байланысты орнату;
- Алынған ақпаратқа аналитикалық сұранысты орындау;
- Жалпы сызба келесі сатылардан тұрады:
 - D ғылыми білім аймағын сипаттайтын терминдерді ерекшелеп алу (кілт сөз);
 - D ғылыми білім аймағының онтологиясын құрастыру;
 - Ақпараттардан алынатын деректерді жүктеу ғылыми саламен ерекшеленеді;
 - Құрастырылған онтология түсінігі мен қолданушылардың ғылыми қорытындыларының арасындағы байланысты орнату;
 - Сұраныстарды орындайтын үлгі алынған ақпараттың қорытындысын қамтамасыз етеді.

Келесі кезекте әр қадамның сипаты берілетін болады. Білім аймағын сипаттайтын терминдерді ерекшелеуде семантикалық, атап айтсақ лингвистикалық және статистикалық әдістерді қолдана отырып алынды. терминдерді ерекшелеу алгоритмін құрастыруда келесі анықтамалар қалыптастырылды.

Анықтама 2.13 берілген диссертацияда термин деп - оның бір немесе бірнеше тақырыптарына сәйкес келетін құжатты сипаттайтын жұп сөздерді атаймыз.

Тапсырманың формальды шешімі келесі жолмен іске асады. $r_1, \dots, r_n$ айдарына бөлінген Doc көптеген құжаттары берілсін делік. Тапсырма *Terms* терминдерінің жиынынан тұрады. Әрбір $A \in Terms$ термин реттелген екі сөзден тұрады: $A = (A_1, A_2)$ (формальды үлгінің толық сипаттамасы 3.1.1 бөлімде).

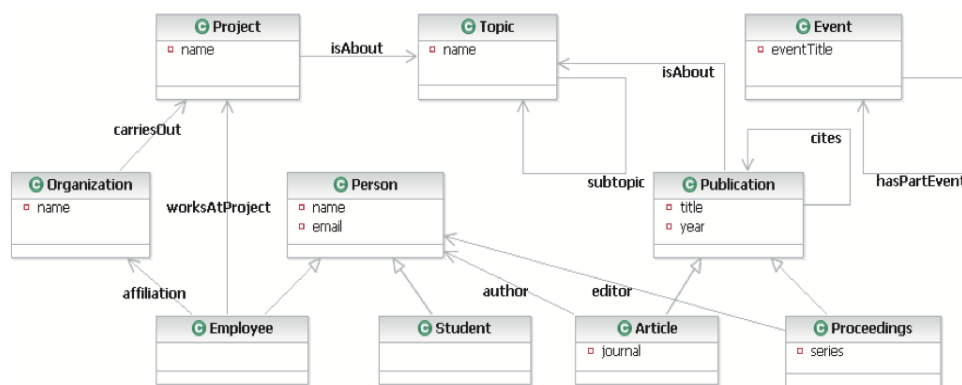
Анықтама 2.14 $O = (I, A)$ онтологияны құрауға арналған тапсырма оның формальды (автоматизацияланған) қорытындысының негізінде қалыптасуын және пән аймағының онтологиясын жасаушыны қызықтыруға тұрарлықтай Doc мәтіндер коллекциясының экспертін (немесе эксперт тобы) таңдаудан тұрады: N_C түсінігінің атаулар жиыны; N_R қатынасының атаулар жиыны; N_x экземплярларының атаулар жиыны; түсініктерді енгізу аксиомасының $I = TBox$ қорытынды жиыны (онтологияның терминологиялық бөлімі); экземплярды бекітудің $A = ABox$ қорытынды жиыны (онтологияның фактологиялық бөлімі).

Онтологияны құрастыру түсіктер жиынының ерекшеленуі мен байланыстардың атауларынан, сонымен қатар осы түсініктердің экземпляры мен олардың арасындағы байланыстан тұрады. Онтологияны толтыру түсінік пен білімді дамытады, және түсініктер экземпляры мен олардың арасындағы қатынасты іздеуді мақсат етеді. Соңғы кезде автоматпен жиі қолданылатын тілдердің ғылыми білім теориясының аймағындағы үлгі бойынша осы тапсырмалардың арасындағы әртүрлі ерекшеліктерді анықтаймыз. Онтологияның терминологиялық бөлімін құрастырудың қорытындысы тапсырманы орындау үшін арналған әдіс ретінде терминмен олардың байланысы, зерттеу аймағы ретінде осындай түсініктердің жиыны, әдістер, тапсырмалар, шешімдер, терминдер болуы мүмкін, зерттеу аймағы арнайы тапсырмалардан тұрады, термин әдістемеде қолданылады, термин зерттеу аймағындағы кілттік түсінік болып табылады. Қазіргі уақытта онтологияны құру мен толтыру келесі себептер бойынша маңызды болып табылатынын ескере кету керек.

Онтология мынандай концептерден тұрады, адам, ұйым, мақала, конференция, жоба, сонымен қатар олардың арасындағы байланыс та бар. Ғылыми саланың жалпы сипаттамалық бағытымен (берілген мақала, конференцияларға жасалатын қорытындысыз) қолданылатын құжаттарда болатын деректерді формалды түрде сипаттау үшін қажет көптеген түсініктер осы онтологияда қалыптасқан. SWRC онтологиясының кілттік түсінігі сурет 2.12 берілген.

Деректерді енгізудің келесі тәсілдері жоспарланады:

- библиографиялық сілтемелерді реттеу;
- жүктелетін мета деректерді реттеу (BibTeX, MathML, LaTeX, FinXML);
- жолдарды қолмен толтыру.



Сурет 4 – SWRC онтологиясы (фрагмент)

Библиографиялық сілтемелерді реттеу. Библиографиялық сілтемелерден ақпаратты алу ақпаратты құрамы жасалынбаған мәтіннен алу тапсырма болып табылады. Берілген библиографиялық сілтемелердің реттелу әдісін тестілеудің нәтиже көрсеткіші бойынша аса жоғарғы әсерін көрсеткен Conditional Random Fields (CRF) алгоритмі. АҚШ Браун университетінде жасалынған FreeCite бағдарламалық комплекс осы алгоритмді іске асыратын CRF++ кітапханасында қолданды.

R.Uskenbayeva, T.Chinibayeva. Model, data integration algorithms of information systems based on ontology // Journal of Theoretical and Applied Information Technology E-ISSN 1817-3195 ISSN 1992-8645 Vol.99 May 2021 No 09. pp 2125-2143	"R.Uskenbayeva, ", "T.Chinibayeva", "Model, ", "data ", "integration ", "algorithms ", "of ", "information ", "systems ", "based ", "on ", "ontology ", "Journal ", "of ", "Theoretical ", "and ", "Applied ", "Information ", "Technology ", "E-ISSN ", "1817-3195", "ISSN ", "1992-8645 ", "Vol.99 ", "Vol.99 ", "2021 ", "No 09. ", "2125-2143".
--	---

["R.Uskenbayeva, ", "T.Chinibayeva "] => "R.Uskenbayeva, T.Chinibayeva";
 "09: 2125-2143" => { :volume =>09, spage => 2125, :epage => 2143 }.

Білім аймағындағы қалыптасқан үлгі мен ғалымдардың ғылыми еңбектерінің нәтижесінен тұратын жүктелген мәтіндерден алынған ақпараттардың арасындағы байланыстың орнауы аналитикалық сұраныстарды орындау үшін аса қажет болып табылады. Бұл деңгейге дейін қолданылатын құжаттан қызметкердің ғылыми еңбегі жайлы мөлшері бар ақпарат қана ерекшеленіп алынады.

O_D (білім аймағындағы термин) онтологиясының $t \in N_x^D$ экземпляры мен O_S (мысалы, мақалалар) онтологиясының $e \in N_x^S$ экземплярының арасындағы *Sim* семантикалық жақындық деңгейін анықтау үшін берілген диссертацияда мынандай формула қолданылады:

$$Sim(e, t) = sim_{edit}(title(e), t),$$

мұндағы *title(e)*– мақаланың атауы, ал $sim_{edit}(s_1, s_2) = \frac{1}{1+editDist(s_1, s_2)} - s_1$ жолының s_2 жолына айналуы үшін қажет түзетулер санынан тең (қою, өшіру,

ауыстыру) $editDist(s_1, s_2)$ Левенштейн қашықтығының негізінде жалғастырылатын жолдардың ұқсас функциясы. Егер $Sim(e, t)$ функциясының мәні C_{sim} константтың мәнінен асатын болса, онда онтология экземплярлары мен ғылыми мақаланың арасында `swrc:isAbout` байланысы орнатылады.

Білімге байланысты онтологиялық әрекет аналитикалық сұраныстарды орындайтын алгоритмдердің қазіргі және бұрынғы апробациясын қолдануға мүмкіндік береді. Жекелей алғанда, онтологияны қолдану арқылы сұранысты қайта жазу логикалық қорытындының механизмінің көмегі арқылы автоматты түрде орындауы мүмкін.

SPARQL тілінің синтаксисін көрсету мақсатында бағдарламалық камтамасыз етуді (“Software Engineering and IS”) жасауға арналған 2020 жылғы жарияланымдарды алуға мүмкіндік беретін сұранысқа мысал келтіреміз.

```
PREFIX swrc:<http://nauka.iitu.kz/ontologies/swrc#>
PREFIX cs:<http://nauka.iitu.kz/ontologies/computer_science#>
SELECT DISTINCT ?pub
WHERE {
  ?pub a swrc:Publication.
  ?pub swrc:year 2020.
  ?pub swrc:isAbout cs:Software_Engineering and IS.
}
```

Уақыт бойынша зерттеудің жеке бағытына зерттеушілердің қызығушылық динамикасы. Бұл сұранысты былай деп жазайық: жылдар бойынша топтастырылған соңғы 10 жыл ішінде $T = \{t_1, \dots, t_n\}$ берілген бағыт бойынша ғылыми еңбектің нәтижелер санын беру. Оны SPARQL тілінде формаландырамыз.

```
SELECT DISTINCT ?res ?year
WHERE {
  ?res a swrc:Result .
  ?res swrc:year ?year .
  { ?res swrc:isAbout t_1 }
  UNION { ?res swrc:isAbout t_2 }
  ...
  UNION { ?res swrc:isAbout t_n } .
  FILTER ( ?year > 2006 && ?year < 2020 )
}
```

Осы сұраныстың нәтижесінде соңғы 5 жыл ішіндегі T бағыты бойынша барлық ғылыми еңбектің нәтижелерінің тізімі беріледі. Осыдан кейін қолданылатын тілдің көмегімен бұл деректерді жылдарға бөлу қажет болады. Бұл тапсырма жұмыстың стандартты функциясының көмегімен іске асады.

Қызықтыратын бағытқа арналған конференциялар тізімі. Бұл сұранысты былай деп жазайық: $T = \{t_1, \dots, t_n\}$ бағытымен байланысты конференциялар тізімі беріледі. Оны SPARQL тілінде формаландырамыз.

```
SELECT DISTINCT ?conf
WHERE {
  ?conf a swrc:Conference .
  { ?conf swrc:isAbout t_1 }
  UNION { ?conf swrc:isAbout t_2 }
  ...
  UNION { ?conf swrc:isAbout t_n } .
}
```

Осындай жолмен келесі бекітілген ережелерді қабылдауға мүмкіндік беретін SPARQL тілінде формаландырамыз.

1 Ереже. O_D ғылыми білім аймағы немесе оның пәндік аймақтағы мүмкін болған барлық терминдермен толықтырылған O_D онтологиясы мен олардың арасындағы байланыс берілді деп есептейік. Жеке ғалымдардың ғылыми еңбектерінің нәтижелері жайлы деректермен толтырылған O_S ғылыми саладағы онтология берілді деп есептейік. Осы онтологиялардың арасында `swrc:isAbout` түрінің мүмкін болған барлық байланыстары орнатылған, яғни ғалымның ғылыми еңбегінің әрбір нәтижесіне оның тематикасын сипаттайтын көптеген белгілер бар. Сонда SPARQL сұраныс тілі мен O_D және O_S онтологияларының ұқсастығы сұраныстарға кепілдік берілген жауап алуға мүмкіндік береді.

Берілген тематикалық бөлімдермен бірге мәтіндер жиынтығынан терминдерді іріктеп алу алгоритмінің математикалық үлгісі. $W - \varepsilon$ – бос сөзін қосқанда Doc коллекциялары арқылы беріліп, құжаттардың барлығында кездесетін барша сөздердің көбі болсын делік, ал PW – барлық реттелінген сөздердің жұптары болсын, яғни $PW = W * W$. d құжаты берілген құжаттар жиынтығындағы n -ші бағытта тұрған сөздің әрбір натуралдық n санына сәйкес келетін $d: N \rightarrow W$ бейнелейді. Сөзі жоқ жағдайдағы нөмірлер (құжаттың соңынан кейін). Осыған сәйкес p жаңа жолы көрсетілген абзацта n -ші жағдайда тұрған n сөзінің әрбір натуралдық санына сәйкес келетін $p: N \rightarrow W$ көрінісі ретінде берілген. Сөз жазылмаған орынның номері бос сөз ретінде белгіленеді. Жиынтықтағы P арқылы барлық жаңа жолдарды белгілейміз. r айдарын шығаратын көптеген құжаттар ретінде, ал нақтырақ айтсақ – $r \in 2^{Doc}$ белгілейді. Айдардың қуаттылығын ондағы құжаттардың мөлшері сияқты $|r|$ арқылы белгілейтін боламыз. Берілген айдарлардың көпшілігін R арқылы белгілейміз.

Сонымен қатар, тағы да бірнеше қосымша белгілерді анықтаймыз:

- $\tau_1: PW \rightarrow W, \tau_2: PW \rightarrow W$ – жұптар сөзі бірінші сөзге (соған сәйкес екінші) сәйкес келетін көптеген басқа сөздерге арналған жұптардың жобасы;
- $Freq: PW * Doc \rightarrow \mathbb{N} \cup \{0\}$ - $d \in Doc$ құжатында $p\omega \in PW$ жұбының енгізілу санын анықтайтын функция;
- $Freq: W * Doc \rightarrow \mathbb{N} \cup \{0\}$ - $d \in$ құжатында $w \in W$ жұбының енгізілу санын анықтайтын функция;
- $L(d) = |\{n \in \mathbb{N} | d(n) \neq \varepsilon\}|$ - d құжатының ұзындығы;
- $id(a) = a$ – ұқсас бейнелеу;
- $Av(f, A) = \frac{\sum_{a \in A} f(a)}{|A|}$ - A соңғы көпше түріндегі f функциясының орташа мәні.

Мысалы, $Av(| \cdot |, R)$ – айдардағы құжаттардың орташа саны, $Av(L, Doc)$ – құжаттың орташа ұзындығы, $Av(id, A)$ - $A = \{a_1, \dots, a_k\}$ жиынтығының орташа арифметикалық саны.

Алгоритм үшін шығарылған деректер жиыны құжаттарда кездесетін, соған сәйкес сөздердегі кестеден тұрады. Ескере кетейік, алгоритмді қолданар алдында лемматизация - құжаттарға ең алдымен лингвистикалық өңдеу жұмысын жүргізуге кеңес беріледі, сөз формаларын дұрыс (сөздік бойынша) формаға келтіріп өңдеу керек. Мысалы, қазақ, орыс тіліндегі зат есім үшін бұндай форма жекеше түрдегі атау септігі болып табылады. Шығару кестесінің әрбір тармағында жазылған төрт сан бар: айдар нөмірі, құжат нөмірі, жаңа жол (абзац)

нөмірі, сөз нөмірі. Егер А сөзі абзац бойында Б сөзінен бұрын кездесетін болса, онда кестеде де А сөзінің тармағы Б сөзінің тармағынан жоғары тұрады. Кесте алғашқы үш баған бойынша іріктелінген. Осындай жолмен, нақтылы сөз жайлы қандай құжатта, қанша рет, және қандай жерінде кездесетіні ғана белгілі.

Алгоритм төрт сатыдан тұрады. Олардың әрқайсысында кейбір ережелердің көмегімен алдыңғы қадам арқылы алынған M_i мен M_{i-1} , жиынтығы таңдалып алынады. Бірінші сатыда тандау PW (сөздердің барлық жұптары) жиынтығынан алынып, іске асады, яғни $M_0 = PW$. M_4 жиынтығы термин – жұптар болып табылады, ол барлық төрт өлшемді де қанағаттандырады.

Келесі өңдеуге арналған жұптардың алғаш таңдап алынуы терминді білдіретін сөздер мәтіннің ішінде айтарлықтай жақын орналасқандығы (дегенмен, жақын орналасуы міндетті емес) жайлы болжамға негізделеді:

$$M_1 = \{pw \in M_0 | \exists p \in P: |p^{-1}(\tau_1(pw))| \leq MAX_{DIST}\}$$

Мәтінде MAX_DIST-1 көп кездесетін басқа сөздердің арасындағы бір абзацта орналасқан екі сөз жұпты құрайды.

Базис ақпараттық болу үшін онда мәтіннің ішінде көп кездесетін сөз болуы керек. Бұл талапты қамтамасыз ету мақсатында барлық жиынтықтарда MIN_FREQ қарағанда кем кездесетін M_1 жұптарының көбінен ерекшеленетін нақты өлшемдер қолданылады:

$$M_2 = \{pw \in M_1 | \sum_{r \in R} \sum_{d \in r} Freq(pw, d) \geq MIN_FREQ\}$$

Сипаттамалық өлшем – алгоритмнің негізгі өлшемі. Оның мәні терминді анықтауда жатыр, жұп кейбір айдар үшін сай келуі керек.

Айдардағы жұптың салмағы. Әрбір жұпка сандар жиынтығы сәйкес келеді – айдардың әрқайсысындағы жұптардың салмағы. r айдарындағы pw жұбының салмағы келесі формула арқылы алынады:

$$Weight_r(pw) = \frac{\sqrt{\sum_{d \in r} \ln \left(\frac{\sqrt{Freq(pw, d)}}{\ln \left(\frac{r}{Av(L, Doc)} + 1 \right)} + 1 \right)}}{\ln \left(\frac{r}{|Av| \cdot |R|} \right)} + 1$$

Осыған сәйкес $Weight_r(pw)$ айдарындағы сөздің салмағы анықталады, нақтырақ айтсақ – жоғарыда берілген формулада $Freq(pw, d)$ белгісі $Freq(w, d)$ ауысады.

Қызметтің жалпы көрінісі. Қалыптасқан талаптарды қанағаттандыратын қызметтерді іріктеп алу үшін $Weight_r$ функциясының келесі жалпы көрінісі таңдап алынады:

$$Weight_r(\bar{x}, \bar{y}, z) = h(g(f(x_1, y_1), \dots, f(x_k, y_k)), z)$$

Функция $f(x, y)$ құжаттағы жұптардың салмағын анықтап, құжатқа қатысты ұзындық пен құжатқа енетін жұптардың санына байланысты болып келеді. Содан кейін жұптардың салмағын анықтау үшін жұптарда кездесетін құжаттар санына тура тәуелділікті қамтамасыз ететін $g(x_1, \dots, x_k)$ функциясы қолданылады. Осыдан кейін айдарға қатысты қуаттан айдардағы жұптардың

салмағының тәуелділігін көрсететін $h(x, y)$ функциясы қолданылады. Тестілеуді жеңілдету мақсатында $Weight_\tau$ функциясының арнайы түрі таңдап алынған, нақтырақ айтсақ $-h(x, y)$ функциясы $f(x, y)$ тең функциядан алынған. Бұл таңдау осы функциялардың ұқсас сипаттамасына негізделген: оларың әрқайсысы ($h(x, y)$) айдарының, не болмаса ($f(x, y)$) құжатының өлшемі мен жұптардың енетін санының салмағына тәуелділігін анықтайды. Сонымен, $Weight_\tau$ функциясының қорытынды, соңғы көрінісі мынандай болады:

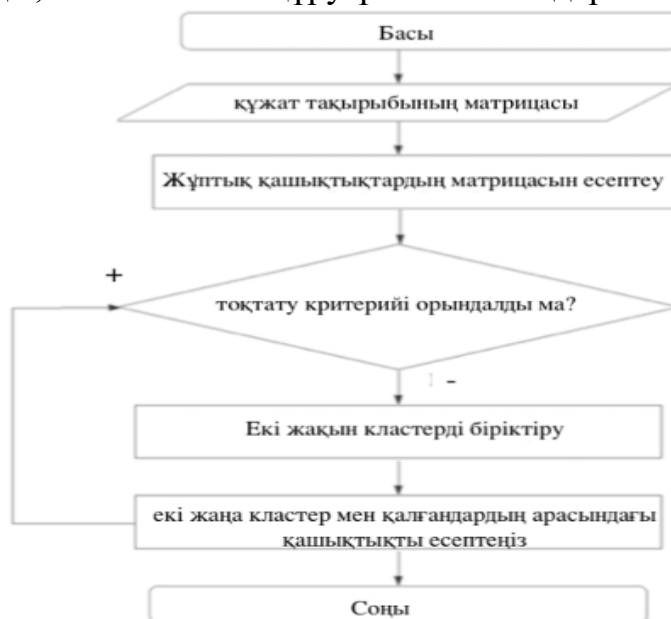
$$Weight_\tau(\bar{x}, \bar{y}, z) = f(g(f(x_1, y_1), \dots, f(x_k, y_k)), z)$$

f мен g функциялары төмендегідей анықталу жолы мен мәніне ие болуы керек:

$$D(f): x \in [0, +\infty), y \in (0, +\infty), E(f) = [0, +\infty),$$

$$D(g): x_i \in [0, +\infty), i = \overline{1, k}, E(g) = [0, +\infty)$$

Ғылыми білімнің жеке аймағындағы онтологияны құрастырудың алгоритмі ғылыми конференциялардың хабарландыруларының жинағына, тақырыптарға бөлінген автордың ғылыми білім саласының онтологиясын құруға арналған авторы әзірлеген Sonmake алгоритмі, сондай -ақ Интернеттегі іздеу жүйесінен алынған ақпарат берілген. Контактілерге арналған хабарламалар (CFP) деп аталады, онтологияны құру үшін негізгі деректер көзі ретінде қолданылды.



Сурет 5 - ғылыми білімнің жеке аймағындағы онтологияны құрастырудың алгоритмдік сызбасы

Ғылыми білім аймағындағы онтологияны құрастыру алгоритмінің келесі қадамы екі деңгейден тұратын $Terms_1$ терминдерінен алынған фильтрация болып табылады.

Фильтрацияның бірінші деңгейінде терминнің өлшеміне жатпайтын кейбір жұптар жойылады. Ол үшін ары қарай тізбектелінген төрт өлшем қолданылады. $A \in Terms$ — A_1 және A_2 екі сөзінен тұратын термин-кандидат болсын, сонда бұл өлшемдер мынандай жолмен қалыптасады:

Википедия онлайн энциклопедиясында A деп аталатын мақала бар;

$$- \frac{hits("A is a term")}{hits(A)} > C_1;$$

- $\frac{hits("A \text{ is a concept}")}{hits(A)} > C_2;$
- $\frac{hits("A_1 \text{ AND } A_2")}{\min(hits(A_1), hits(A_2))} > C_3.$

Фильтрлеудің екінші деңгейінің мақсаты D білімінің берілген аймағымен байланысты болмайтын сөздердің жұбын жою болып табылады. Ол үшін мына өлшем қолданылады:

$$\frac{hits("A \text{ AND } D")}{hits(A)} > C_4$$

Келесі сатының мақсаты баланысты терминдердің жұптарын ерекшелеу болып табылады, яғни семантикалық жақын болып келетін жұптардың N_x^D жиынындағы мүмкін болатын термин жұптарымен байланысты. Екі терминнің арасындағы семантикалық жақындық деңгейін анықтау үшін Normalized Google Distance (NGD) кеңінен таралған шарты қолданады. A мен B – терминдер, ал N – ізденіс жүйесімен индекстелетін парақшалардың жалпы саны болсын. Сонда A мен B арасындағы NGD семантикалық жақындық деңгейі мына формула арқылы анықталады:

$$NGD(A, B) = \frac{\max\{\log hits(A), \log hits(B)\} - \log hits("A \text{ AND } B")}{\log N - \min\{\log hits(A), \log hits(B)\}}$$

Алгоритмнің келесі деңгейі терминдердің иерархиясының құрылуы. Херст өңдеп шығарған лингвистикалық шаблондардың көмегімен иерархия түсінігін қалыптастырудың классикалық алгоритмі ғылыми бағыттағы иерархияны қалыптастыру үшін онша әсерлі болмай шықты

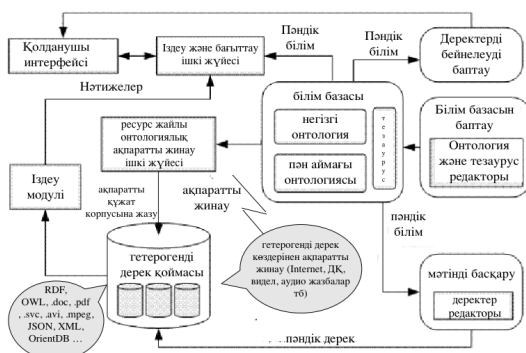
Берілген зерттеу жұмысындағы зерттеуді жүргізу барысында ғылыми деңгейдегі иерархия түсінігін қалыптастыру үшін лингвистикалық шаблондар жасалынған. Негізгі шаблондар мынандай:

$$A \text{ is } * \text{ keyword } * \text{ prep(aux)}? B$$

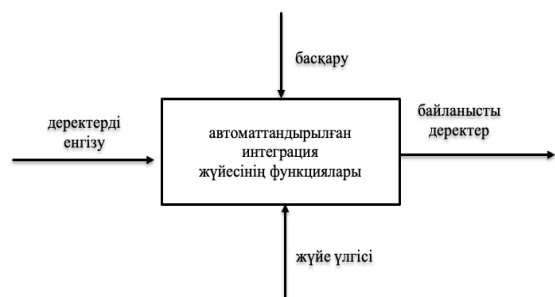
Терминдерді категорияларға жіктеу

$A \in Terms_2$ термині жататын класстардың кішкентай жиынын анықтау үшін келесі формула бойынша $C \in N_C^D$ әрбір классына жататын оның деңгейі есептелінеді:

$$score(A, C) = \frac{hits("A \text{ is a } C")}{hits(A)}$$



Сурет 6 – Жүйелік архитектура



Сурет 7 - Функционалды модель

Диссертациялық жұмыста сипатталған әдістер мен алгоритмдерді жүзеге асыру үшін гетерогенді дерек көздерден қажетті ақпаратты шығарып, олармен байланысты деректер түрінде ұсынуға арналған бағдарламалық жасақтама құрылды.

Қорытынды. Бұл диссертацияда ғылыми ақпараттарды басқару жүйесінің құрылу тәсілі мен әдістерінің сипаттамалары берілген. Әрекет етудің теоретикалық негізі онтология болып табылады. Жүйенің авторы ұсынған нұсқа сұраныстарды орындау үлгісі, онтологияны құрау мен ғалымдардың ғылыми жұмыстарының нәтижесі туралы деректермен жүйеге енгізілген байланысты орнатушы үлгі, ақпараттарды жүктеуші үлгі, білім аумағындағы формальды үлгіні құраушы үлгі сияқты үлгілерді өз құрамында сақтайды.

Бұл диссертацияны дайындау бойынша жүргізілген жұмыс барысында төмендегідей негізгі нәтижелер алынған.

- Пәндік аймақты зерттеу негізінде ғылыми ұйымның жұмысының нәтижесін сипаттайтын ақпаратты сұрату бойынша қорытынды, сақтау мен толықтыру жүйелерін құрау үшін онтологияға сүйенетін технологиялық және архитектуралық шешімдер қарастырылған, математикалық үлгілер мен алгоритмдер құрастырылған. Онтология мен SPARQL тілін қолдану арқылы оның бүкіл өмір сүру циклындағы жүйенің кодының тиімділігі үшін қосымша мүмкіндіктер мен есептеудің кепілдігін туғызатын жүйеге деген сұраныстардың формальды сипаттамасы ұсынылған.

- Ғаламтордағы іздеу жүйесіндегі ақпараттарды қолдануға, ғылыми конференциялардың жарнамаларындағы терминдерді ерекшелеуге негізделген ғылыми білім аумағындағы онтологияны құраушы алгоритм жасалынған. Оның бағдарламасының іске асырылуының күрделілігін сипаттайтын аналитикалық баға алынған.

Жұмыстың апробациясы:

1. R. Uskenbayeva, Y. Chinibayev, A. Kassymova, T. Temirbolatova, K. Mukhanov. Technology of integration of diverse databases on the example of medical records//Proceedings of the 14th International Conference on Control, Automation and Systems (ICCAS 2014) - Gyeonggi -do, Korea, 2014. P 282-285. ISSN: 2093- 7121.

2. R.Uskenbayeva, T.Temirbolatova, Young Im Cho, Z.Uskenbayeva, G.Bektemyssova, A. Kassymova. Recursive decomposition as a method for integrating heterogeneous data sources//Proceedings of the 15th International Conference on Control, Automation and Systems (ICCAS 2015). – Busan, South Korea. October 13-16, 2015 – P.2076-2079. ISSN: 2093 - 7121

3. Р.К. Ускенбаев, Т.Т.Темірболатова, А.Б. Касымова. Бұлттық есептеуде mapreduce технологиясымен үлкен деректерді өңдеу - // Вестник КазНТУ имени К.Сатпаева No5 (111). – 2015. С.50 - 53. ISSN 1680- 9211

4. Р.Ускенбаева, Г. Бектемысова, Т.Темірболатова. Интеграция больших неоднородных данных с использованием языка R и HADOOP - Вестник КазАТК - №4 2015-11-01

5. Ускенбаева Р.К., Аманжолова С.Т., Темірболатова Т.Т. Анализ и локализация инцидентов снижения работоспособности распределенных

вычислительных систем. Труды международного форума «инженерное образование и наука в XXI веке: проблемы и перспективы», посвященного 80-летию Каз НТУ им. К.И. Сатпаева

6. T. Temirbolatova, D. Beisenov Automatic asynchronous exchange of business object between heterogeneous systems - The 12th ICIT&M 2014. 2014 April 16-17, 2014, Information Systems Management Institute, Riga, Latvia

7. T. Temirbolatova, A.Khamitov, A. Keldybay, T.Sembayeva Manage different-structured Big Data - The 12th ICIT&M 2014. 2014 April 16-17, 2014, Information Systems Management Institute, Riga, Latvia

8. Temirbolatova T. Jarmukhambetov Y., Temirbolatova U. The method of extracting semantic meta descriptions from databases//2nd International scientific conference «Information Technologies in Science &Industry» International IT University, May 19, 2016 Almaty, Kazakhstan. ISBN 978-601-7407-33-9

9. T. Chinibayeva Security semantic database problems // Herald of the Kazakh-british technical university ISSN1998-6688. V INTERNATIONAL CONFERENCE "DIGITAL TECHNOLOGY IN SCIENCE AND INDUSTRY - 2019» (DTSI-2019), 10th Anniversary INFORMATION TECHNOLOGY INTERNATIONAL UNIVERSITY Vol.16, No.3 (2019), pp. 168-174

10. R.Uskenbayeva, T.Chinibayeva. Algorithm for the construction of an ontology in the field of scientific knowledge//The Bulletin of Kazakh Academy of Transport and Communications named after M. Tynyshpayev ISSN 1609-1817. Vol. 107, No.4 (2018), pp. 259-266

11. R.Uskenbayeva, T.Chinibayeva. Method of extracting meta description from databases//Herald of the Kazakh-british technical university ISSN1998-6688. Vol.15, No.4 (2018), pp. 116-123

12. R.Uskenbayeva, T.Chinibayeva. Model, data integration algorithms of information systems based on ontology // Journal of Theoretical and Applied Information Technology E-ISSN 1817-3195 ISSN 1992-8645 Vol.99 May 2021 No 09. pp 2125-2143