УДК 004.853, 004.55 https://doi.org/10.15407/pp2020.04.055

Ю.В. Рогушина, А.Я. Гладун

ЗАСТОСУВАННЯ ОНТОЛОГІЧНОГО АНАЛІЗУ ДЛЯ ОБРОБКИ МЕТАДАНИХ ПРИ ІНТЕРПРЕТАЦІЇ BIG DATA НА СЕМАНТИЧНОМУ РІВНІ

Розглядається застосування менеджменту знань для аналізу Big Data. Щоб визначати, яку саме інформацію можна отримати з Big Data, і зробити це здобуття більш ефективним, пропонується застосовувати фонові знання з онтологій предметних областей. За допомогою таких онтологій користувачі можуть формально описувати свої інформаційні потреби, задавати структуру потрібних інформаційних об'єктів та явно виділяти важливі для поточної задачі аспекти. Предметом аналізу Big Data є їх метадані, в яких відомості про семантику, як правило, представлені неструктурованим природномовним описом. Тому виникає потреба у стандартизації подання метаданих, в яких онтології визначають структуру та семантику окремих елементів.

Ключові слова: Big Data, онтологія, метадані, семантична розмітка.

Вступ

Метадані дозволяють охарактеризувати контекст, контент і структуру Big Data, а також методи керування ними. Метадані накопичуються з плином часу та документують історію Big Data. Метаданими необхідно керувати, як самими даними, оскільки вони мають бути захищені від втрати, несанкціонованого видалення, збережені або знищені, а також доступ до керування ними має бути організовано через розподіл прав доступу і виконання певних правил безпеки. Семантику Big Data відображають, як правило, неструктуровані природномовні описи, що входять до складу метаданих, але обробка такої інформації потребує значно більше зусиль порівняно з обробкою структурованої інформації. Тому ціль даної роботи – аналіз напрямків структурування метаописів Big Data з використанням існуючих стандартів.

Метадані та їх властивості

Метадані у найбільш широкому розумінні – це дані про дані. Але таке визначення надто просте й неконструктивне. Вікіпедія визначає метадані як дані з формальної системи вищого рівня, що описує задану систему даних або як структуровані дані, що характеризують певні сутності для їх ідентифікації, пошуку, оцінки та керування ними [1]. Це окремий тип інформаційних ресурсів (ІР), які потребують специфічних засобів подання, створення та обробки (ІР – це будь-яка сутність, яка спроможна передавати чи зберігати інтелектуальну інформацію або знання [2]).

Хоча спочатку метадані призначалися тільки для опису даних, проте останнім часом вони використовуються для опису найрізноманітніших інформаційних ресурсів (ІР) та об'єктів (концептуальних схем, онтологій, сервісів тощо). Вони дозволяють характеризувати життєвий цикл даних, дії та потреби різних суб'єктів обробки даних. Нині метадані дозволяють характеризувати зміст ІР, наприклад, описувати модель предметної області (ПрО) на семантичному рівні.

Розвиток інформаційних технологій став причиною істотного розширення функцій метаданих і викликав їхнє різноманіття. Зміст метаданих, їхні функції і засоби їхнього представлення визначалися тими інформаційними технологіями, що використовувалися для створення таких ІС, специфікою ПрО та тих ІР, що оброблялися цими ІС.

Розповсюдження електронних бібліотек [3], в яких зберігаються ІР різних типів, сховищ даних та знань, що впроваджують технології Semantic Web [4], викликало посилення інтересу до семантизації метаданих [5].

На сьогодні існує велика кількість визначень метаданих, що відображають

різні точки зору на цей термін та на сферу використання метаданих [6]. Метадані — це інформація, що робить дані корисними [7]. Таке визначення описує сферу застосування метаданих, але є надто загальним для практичного використання. Наприклад, для Big Data це визначає роль метаданих, але не дозволяє конкретизувати вимоги до способів їх представлення.

Метадані призначені як для комп'ютерної обробки, так і для інтерпретації людиною інформації про цифрові і нецифрові об'єкти [8]. В роботі [9] метадані визначаються як структуровані дані, що містять характеристики сутностей, які вони описують, для цілей їхньої ідентифікації, пошуку, оцінки та керування. Слід враховувати, що метадані, які використовуються для опису ресурсів Web, є, як правило, слабо структурованими, але вони відповідають погодженим моделям, що забезпечують їх операційну інтероперабельність у неоднорідному середовищі [10].

В роботі [11]. метаданими називається будь-яка дескриптивна інформація про інші джерела даних, яка сприяє організації, ідентифікації, представленню, визначенню місця розташування, забезпеченню інтероперабельності, керуванню і використанню цих даних. В роботі [12] метадані характеризують не інформаційний ресурс у цілому, а певний елемент даних, що відноситься до цього ресурсу. Такий підхід найбільш відповідає специфіці збереження Big Data у великих сховищах, тоді як ідентифікувати потрібно підмножину даних, що пертинентні конкретній задачі користувача.

Метадані можуть використовуватися для визначення семантики інформації, отже, для поліпшення її пошуку і вибірки, розуміння і використання. Наприклад, в [13] розглядається застосовуватися онтологій та тезаурусів для семантичного анотування ІР та їх елементів, що є основою для машинного навчання та здобуття знань з даних. Залежно від цілей анотування можуть застосовуватися онтології різної складності (від контрольованих словників та глосаріїв до онтологій із складними відношеннями інверсії, неперетину тощо). Dublin Core (http://www.dublincore.org/) є прикладом легкої онтології, яка широко використовується для опису характеристик електронних документів та семантизації метаданих.

Конкретний склад функцій метаданих залежить від особливостей тієї системи, що їх використовує, від характеру ІР та їх елементів, які описують ці метадані, від базових інформаційних технологій системи, від потреб її користувачів і від багатьох інших факторів.

Властивості метаданих:

1. В*ідносність* поділу ІР на дані та метадані – метадані для однієї ІС можуть розглядатися як дані в іншій, та навпаки (наприклад, онтологія, що використовується для анотування ПМ-тексту, є елементом метаданих, а та сама онтологія в репозиторії онтологій [14] є даними);

2. *Багаторівневість* опису властивостей будь-якого іншого ресурсу може здійснюватися в термінах більш абстрактної системи понять, які можуть утворювати ієрархію рівнів, яка може включати довільну кількість рівнів (наприклад, Meta Object Facility (MOF) [15] має три рівні, а Dublin Core – два);

3. *Гетерогенність* ІР та даних, що можуть описуватися метаданими: властивості, які дозволяють охарактеризувати метадані, залежать від специфіки самих даних та сфери їх використання;

4. *Відчуженість* метаданих від ІР: метадані можуть зберігатися незалежно або бути убудованими в ІР, які вони характеризують;

5. Ступінь *залежності від контента* визначається змістом самих метаданих (наприклад, дата створення і тип файлу не залежать від контенту, тоді як анотація тексту визначається контентом);

6. Ступінь *залежності від ПрО* визначається цілями створення метаописів, які можуть бути спеціалізованими або універсальними;

7. Ступінь *структурованості;*

8. Рівень *гранулярности опису ресурсів* визначає, які саме елементи ІР описуються метаданими;

9. Ступінь *динамічності* визначається тим, за яких умов та як часто можуть змінюватися метадані;

10. Ступінь *формалізованості* визначається тим, які засоби використовуються для представлення метаданих. Для представлення метаданих (ПМ, ПМ з обмеженим словником, формальні мови – наприклад, OWL [16]).

Існує багато інших властивостей метаданих, які можуть враховуватися в різних дослідженнях (наприклад, засоби представлення, способи збереження та наявність явного подання), але вони не є принциповими для опису Big Data і тому не розглядаються у даній роботі.

Недоліки систем метаданих [17] – це низька оперативність відновлення інформації; неузгоджене введення змін у метадані, що призводить до суперечливості та дублювання; недостатня автоматизація системи ведення метаданих на основі керування контентом; орієнтованість на роботу з одним типом об'єктів (ІР та їх елементів, які описують метадані); відсутність єдиної моделі метаданих для всіх типів об'єктів; відсутність спільного розуміння одиниці опису метаданих – екземпляра метаданих, який описується сукупністю параметрів, що не перетинається з іншими сукупностями, що описуються іншими метаданими; неповнота набору об'єктів метаданих, які зазвичай не містять відомості про засоби обробки та збереження даних.

Неструктуровані дані

Неструктуровані дані (НСД) – це інформація, яка не має попередньо визначеної моделі даних або не організована заздалегідь [18]. Якщо певні елементи метаданих не мають формалізованої структури, то для здобуття з них потрібної інформації необхідно застосовувати методи, що орієнтовані на аналіз НСД. Саме НСД потенційно мають найбільшу цінність як джерела нових знань, і чим більше таких даних доступні для аналізу, тим точніше результати. Більш детально властивості НСД та засоби їх обробки проаналізовано в [19].

Природномовна інформація – набори слів природної мови (ПМ) довільної довжини, поєднані за слабо формалізованими лінгвістичними правилами та представлені в електронній формі, може аналізуватися як НСД. Це обумовлюється тим, що хоча така текстова інформація містить деякі структурні елементи, але у більшості ІР такі структурні елементи не представлені явно, і тому їх здобуття потребує великого часу та зусиль.

Для аналізу НСД можна застосовувати семантичну розмітку. Найбільш корисним засобом семантичної розмітки є зв'язування елементів ІР з елементами онтології (наприклад, фрагмент ПМ-тексту пов'язується з класом або екземпляром класу онтології, а інший елемент – із значенням його властивості). Але з точки зору легкості впровадження безпосереднє застосування онтологій для семантизації ІР є недоцільним – більшість користувачів не володіють онтологічним аналізом, не знають мови подання онтологій тощо. Тому більш корисно використовувати простіші засоби семантизації, наприклад, семантичну Wiki-розмітку. Така семантична вікіфікація може виконуватися як експертами ПрО, так і технічними співробітниками.

Значний недолік цього підходу – семантична Wiki-розмітка ІР, що побудована для однієї ПрО, не може використовуватися для іншої ПрО. Тому доцільно застосовувати онтології вищого рівня, для створення яких можуть застосовуватися онлайнові енциклопедії, що побудовані на основі технологій семантичних Wiki (наприклад, портальна версія Великої української енциклопедії e-ВУЕ [20]). Семантична розмітка дозволяє також аналізувати семантичну подібність між поняттями обраної та використовувати її надалі для аналізу НСД [21].

Метадані для Big Data

Властивості метаданих, їх склад і функції істотно залежать від технологій реалізації систем, в яких вони використовуються, особливостей описуваних ними ресурсів, а також від області застосування і конкретних програм.

Певний набір даних розглядається як Big Data, якщо він володіє однією або декількома характеристиками, так званими характеристиками «5V»: *об'єм; швидкість; різноманіття; достовірність; цінність* [22]. Метадані, які характеризують Big Data, можуть містити інформацію про

джерело даних; про автора і дату створення документа; кількість записів у наборі даних; опис цих даних тощо. В обробці Big Data аналіз метаданих має ключове значення, тому що метадані містять інформацію не тільки про походження даних [23, 24], але й про їх зміст.

Метадані для Big Data [25] – це структурована або напівструктурована інформація, яка дозволяє створювати, керувати і використовувати Big Data у різний час і у різних сферах діяльності, а також робити відбір таких наборів Big Data, що релевантні задачі, яку необхідно вирішити [26]. Для опису метаданих використовуються різні природні та штучні мови. Природні мови є найбільш багатими і виразними в порівнянні з іншими засобами подання метаданих. Вони призначені не для комп'ютерної обробки, а для людей, і не забезпечують однозначності і строгості інтерпретації метаданих, і тому такі описи аналізуються як НСД.

Штучні мови, які використовуються для опису метаданих, – це мови опису даних СУБД, концептуального моделювання, опису онтологій, бізнес-процесів; мови подання онтологій OWL, RDF; мови розмітки тощо.

Стандартизація метаданих

Стандартизація метаданих – основа інтероперабельності та повторного використання як самих метаданих, так і тих IP, що характеризують ці метадані. Тому міжнародні організації зі стандартизації приділяють велику увагу розробці форматів метаданих, які призначені для формального опису різних типів IP та інформаційних об'єктів (IO). Такі стандарти включають в себе набір властивостей, що дозволяють характеризувати конкретний IO. Такі стандарти можуть бути залучені (з різною ефективністю) для опису Big Data. Нині в Україні три міжнародні стандарти, що стосуються метаданих, (ISO 15489-1:2016 [27], ISO 15836-1:2017 [28], ISO 15836-2:2019 [29]) прийнято як національні стандарти методом підтвердження [30, 31].

Стандарт *ISO 15489-1:2016 Information and documentation – Records management — Part 1: Concepts and principles (Інформація і документація. Керування документами. Частина 1: Поняття і принципи)* визначає основні поняття і принципи керування документами і інформацією. Цей стандарт може бути застосований для відображення основних властивостей Big Data: 1) автентичності; 2) достовірності; 3) цілісності; 4) придатності їх до обробки). В стандарті описано інформаційні поля, що входять в структуру метаданих. Для Big Data ці поля дозволяють відобразити наступну інформацію: опис контенту Big Data – це структура даних (форма, формат, зв'язки між блоками Big Data); середовище створення; взаємозв'язок з іншими блоками Big Data (шардинг, реплікація) і метаданими; ідентифікатори та іншу інформацію, що потрібна для видобутку і подання даних; дії і події, що пов'язані з цими Big Data (дата, час дій, зміна метаданих тощо). Big Data, які не супроводжуються такими метаданими, не можуть використовуватися повноцінно.

Стандарт *ISO 15836-1:2017 Information and documentation — The Dublin Core metadata element set — Part 1: Core elements* (Інформація та документація. Набір елементів метаданих «Дублінське ядро». Частина 1: Основні елементи) описує 15 елементів Dublin Core, які використовують для опису ресурсів. В цьому стандарті під ресурсом розуміють будь-який об'єкт, який можна ідентифікувати (наприклад, у сфері комп'ютерних наук ресурсами виступають окремі документи, тексти, аудіо- та відео-файли, Web-сторінки, бази даних тощо). Big Data та їх метадані теж відповідають такому визначенню і можуть розглядатися як ресурси. 15-елементне «ядро», зазначене в цьому стандарті, є частиною більшого набору словників метаданих та технічних специфікацій, що підтримуються Дублінською ініціативою метаданих (Dublin Core Metadata Initiative, DCMI) [32]. Основні елементи можуть використовуватися в поєднанні з термінами метаданих з інших сумісних словників у контексті профілів застосунків, як зазначено в абстрактній моделі DCMI [DCAM]. В табл. 1 приведена специфікація 15 елементів метаданих Dublin Core.

Таблиця 1. Специфікація 15 елементів метаданих Dublin Core

Назва елемента	**Мітка елемента**	**Визначення**	**Коментар**
title	**Заголовок**	**Назва ресурсу**	
creator	Автор	Сутність, відповідальна за створення контенту ресурсу	Людина, організація або сервіс; зазвичай збігається з ім'ям людини, назвою організації або сервісу
subject	Тема	Тема контенту ресурсу	Як правило, подається ключовими словами, фразами або кодами класифікації. Рекомендується вибирати значення з певного словника. Просторова або часова приналежність ресурсу повинна описуватися елементом coverage
description	Опис	Опис контенту ресурсу	Опис контенту ресурсу може включати зміст, анотацію, графічну презентацію або короткий текстовий опис ресурсу
publisher	Видавець	Сутність, що робить ресурс доступним	Людина, організація або сервіс; зазвичай збігається з ім'ям людини, назвою організації або сервісу
contributor	Учасник	Сутність, що бере участь у створенні контенту ресурсу	Людина, організація або сервіс; зазвичай збігається з ім'ям людини, назвою організації або сервісу
Date	Дата	Дата події в життєвому циклі ресурсу	Може використовуватися для подання інформації про час з будь-яким рівнем точності
type	Тип	Вид або категорія контенту ресурсу	Рекомендується вибирати значення з певного словника, такого як DCMI Type Vocabulary. Фізичне або цифрове подання ресурсу визначається елементом format
format	Формат	Фізичне або цифрове подання ресурсу, вимір	Вимірювання може бути, наприклад, розміром або тривалістю
identifier	Ідентифікатор	Конкретне посилання на ресурс в цьому контексті	Рекомендується визначати ресурс за допомогою рядка або числа, що задовольняє формальній системі ідентифікації
source	Джерело	Посилання на ресурс, на основі якого складено цей ресурс	Цей ресурс може складатися з "Джерела" частково або повністю. Рекомендується визначати "Джерело" за допомогою рядка або числа, що задовольняє формальній системі ідентифікації
coverage	Охоплення	Простір або границі, з якими пов'язано вміст ресурсу	Як правило, географічне положення (назва місця або координати), часовий період (назва періоду, дата, набір дат) або підвідомча область (така як адміністративна область)
language	Мова	Національна мова вмісту	Рекомендується вибирати значення з певного словника, такого як RFC 4646
relation	Зв’язування	Посилання на зв'язаний ресурс	Рекомендується визначати "зв'язування" за допомогою рядка або числа, що задовольняє формальній системі ідентифікації
rights	Правова інформація	Правова інформація, пов'язана з ресурсом	Зазвичай "Правова інформація" містить правові угоди щодо ресурсу, включаючи інформацію про права на інтелектуальну власність

Міжнародний стандарт *ISO 15836-2:2019 Information and documentation – The Dublin Core metadata element set – Part 2: DCMI Properties and classes* (Інформація та документація. Набір елементів метаданих «Дублінське ядро». Частина 2: DCMI властивості і класи) є розширенням і доповненням першої частини цього стандарту ISO 15836-1. Розширення полягає у тому, що він надає програмістам загальну універсальну мову для створення та аналізу метаданих. Така універсальна мова забезпечує розширений опис елементів метаданих, використовуючи їх оновлені властивості та класи. Стандарт ISO 15836-2 збільшує початковий набір з 15 основних властивостей до 40 властивостей і 20 класів для підвищення точності і виразності описів у стандарті Dublin Core. Основна увага цього стандарту зосереджена на опису загальних властивостях елементів метаданих, що необхідні для базової інтероперабельності між різними мовами програмування та предметними областями їх застосування.

Такий набір властивостей і класів подається як словник RDF і може використовуватися для зв'язаних даних (Linked Data). Кожна властивість і клас ідентифікується глобальним ідентифікатором для використання в даних RDF. Розробники метаданих, що не належать до RDF, можуть використовувати словник у XML, JSON, UML та реляційних БД, не застосовуючи глобальний ідентифікатор і специфічні для RDF аспекти визначень термінів.

Значення URI можуть бути використані для створення посилань зі значень елементів на відповідні ресурси Web. URI – це уніфіковані локатори ресурсів (URL-адреси) або постійні ідентифікатори, такі як уніфіковані імена ресурсів (URN). Стандарт Dublin Core визначає лише посилання другого типу. У стандарті подані імена властивостей, які можуть бути префіксами для використання як ідентифікатори або цитуватися як повні URI, використовуючи простір імен PURL за замовчуванням.

Таким чином, важливим досягненням базового набору елементів Dublin Core є те, що його розширена семантика дає можливість опису будь-яких Web-ресурсів. Однак існують і негативні наслідки цієї позитивної характеристики.

1. Розширення семантики припускає різні інтерпретації (найбільш складними в інтерпретації є пари "relation – source", "creator – contributor", "type – format".

2. Для опису конкретних категорій ресурсів глобальний рівень є недостатнім: він не відображає важливі характеристики ресурсу. Це стосується основних ПМ-об'єктів опису в репозиторіях – статей, матеріалів конференцій, книг, дисертацій.

Тому можуть вводитися більш детальні елементи опису ресурсів з використанням: розширеного набору термів Dublin Core, які нам надає стандарт ISO 15836 Part 2: "DCMI Properties and classes "(ISO 15836-2: 2019); інших форматів метаданих, таких як MODS (Metadata Object Description Schema) на базі спрощеного набору елементів формату MARC, ETD-MS для опису дисертацій, Data Cite Metadata Schema та інших; власних наборів метаданих, які формуються на основі розширеного формату з додаванням специфічних елементів.

Для забезпечення уніфікації значень і потрібного рівня деталізації метаданих, отримуваних по OAI-PMH у форматі базового DC, репозиторії-агрегатори застосовують набір рекомендацій щодо обов'язкового використання деяких полів; уніфікації використання полів (наприклад, для статей рекомендується записувати назву журналу в поле dc: source); уніфікації формулювань значень полів, важливих для пошуку та щодо заповнення полів з можливостями структурування.

Тенденції розвитку структур метаданих йдуть у напрямку більшого різноманіття і диференціації елементів. Це пов'язано з підвищенням ролі репозиторіїв в структурі відкритої науки, з розміщенням наукових публікацій, підготовлених за підтримки фондів, у репозиторії як альтернативі публікацій в журналах відкритого доступу.

З огляду на ці тенденції, ми можемо виділяти у своїх внутрішніх структурах метаданих окремі елементи, щоб згодом передавати їх в деталізованих обмінних форматах.

Метадані та типові інформаційні об'єкти

Як показав аналіз сучасних систем метаданих, вони дозволяють описувати не тільки ІР у цілому, але й типові для певної ПрО інформаційні об'єкти, які описуються у цих ІР та є їх елементами. Типові інформаційні об'єкти (ТІО) характеризуються набором семантичних властивостей, які можуть бути описані в метаданих кожного екземпляра. ТІО можуть описувати як ІО (документи, елементи БД, мультимедійну інформацію), так і об'єкт реального світу (персоналій, організації, географічні об'єкти тощо). Доцільність створення ТІО визначається специфікою ПрО конкретної ІС: якщо в системі обробляється певна кількість елементів із подібним набором властивостей та характеристик, тоді доцільно виділити для них окремий ТІО.

Відповідно до концепції ТІО [33], які дозволяють класифікувати інформацію про різноманітні ІО зі складною структурою на семантичному рівні, значення деяких елементів метаданих Dublin Core можуть бути віднесені до певних ТІО (табл. 2), що надалі визначає правила їх аналізу та обробки. Крім того, деякі з них можуть розглядатися як ТІО – поняття ПрО, що відповідають класам та екземплярам класів онтології ПрО, тоді як інші є ПМ-описами. Визначити ТІО елементів дозволяє аналіз коментарів, що надаються у стандарті.

Таблиця 2. ТІО елементів метаданих Dublin Core

Назва	**ТІО**
title	Поняття ПрО
creator	Персоналія, Організація, Сервіс
subject	Поняття ПрО
description	ПМ-опис, НСД
publisher	Персоналія, Організація, Сервіс
contributor	Персоналія, Організація, Сервіс
Date	Структуровані дані, Дата
type	Поняття з онтології "Ресурси"
format	ТІО (поняття з онтології "Типи даних")
identifier	Посилання
source	Посилання
coverage	Поняття з онтології "Географічні об'єкти"
language	Поняття з онтології "Мови"
relation	Посилання
rights	ПМ-текст, НСД

Структура та відношення між ТІО можуть відображатися різними засобами подання знань. Наприклад, в онтологіях ТІО відповідають класи, а їх характеристикам – властивості екземплярів класів. В семантичних Wiki-ресурсах для подання ТІО використовуються шаблони, що містять категорії та набір семантичних властивостей ТІО (рис. 1).

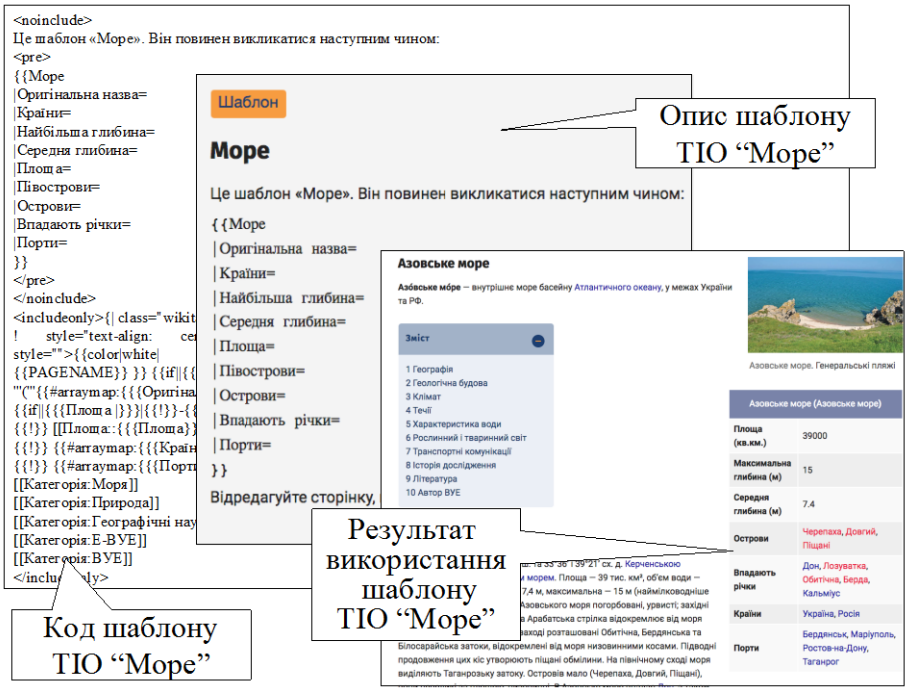

Рис. 1. Використання шаблонів в Semantic MediaWiki для подання ТІО

Використання Data Mining для аналізу метаданих Big Data

На сьогодні створено багато методів, що забезпечують здобуття знань з різних типів ІР – структурованих, частково структурованих та неструктурованих [34]. Аналіз таких методів показує, що внесення структурних елементів у дані значно зменшує простір рішень та зменшує час обробки.

Досить часто основою для створення ТІО є застосування різних напрямків Data Mining для здобуття знань з метаданих цих ТІО для більш ефективної роботи ІС. Особливо це актуально для Big Data, тому що саме аналіз метаданих такої інформації є основою для створення наборів Big Data, що можуть використовуватися як дані для машинного навчання (категоризації та кластеризації). В такому випадку властивості ТІО є параметрами вибірки даних, значення яких аналізуються методами Data Mining [35] , і тому коректне створення ТІО є визначальним фактором обробки Big Data в цілому.

Data Mining – це процес, спрямований на виявлення нових значущих кореляцій, шаблонів і тенденцій у результаті аналізу великого обсягу збережених даних з використанням методик розпізнавання зразків та застосування статистичних і математичних методів. Особливо ефективними методи Data Mining стали із розвитком та накопиченням Big Data. Можна казати, що Data Mining – це процес автоматизованого здобуття з наявних інформаційних ресурсів нових знань, які неявним чином присутніми в оброблюваній інформації.

Результати *Data Mining* у значній мірі залежать від тих даних, які вони обробляють: від їх повноти, актуальності, релевантності поставленій задачі та якості, та від знань, на основі яких обираються ці дані. Тому в тому випадку, якщо побудова набору даних базується на аналізі їх метаданих, саме склад та якість метаданих значним чином визначають якість тих знань, що можна здобути з ІР.

Інструменти Data Mining дозволяють знаходити нові закономірності у даних самостійно й також самостійно будувати гіпотези про взаємозв'язки між їх елементами. Оскільки саме формулювання гіпотези щодо залежностей є найскладнішим завданням, то перевага Data Mining у порівнянні з іншими методами аналізу є очевидною. Але для їх ефективного використання ці результати мають бути пов’язані з відповідним поняттєвим апаратом, який формалізується засобами подання знань, наприклад, за допомогою онтологій [36]. У багатьох випадках такий зв’язок встановлюється через семантичні метадані – ті елементи метаданих, що пов’язані з певним поданням знань, наприклад, з елементами онтології відповідної ПрО. Знання, що здобуваються таким чином з даних, дозволяють у свою чергу вдосконалити онтологію ПрО, яка надалі використовуватиметься для створення метаданих. Таким чином, створення метаданих та їх використання для вдосконалення онтологій є циклічним процесом, який підтримує більш ефективне збереження та використання даних.

Найпоширеніші сфери використання Data Mining пов’язані із вирішенням задач класифікації, кластеризації та прогнозування. Слід відмітити, що Data Mining характеризує не стільки конкретну інформаційну технологію, скільки процес пошуку закономірностей (кореляцій, тенденцій, взаємозв'язків) за допомогою математичних і статистичних алгоритмів, наприклад, регресійного й кореляційного аналізу тощо.

Найбільш розповсюджена задача, що вирішується за допомогою Data Mining, – це задача *класифікації*: вирішення задачі класифікації дозволяє виявити ознаки, що характеризують групи об'єктів досліджуваного набору даних – класи, за якими новий об'єкт можна віднести до того чи іншого класу. Ця задача безпосередньо пов’язана з онтологічним аналізом і дозволяє віднести екземпляри до відповідних класів. Для вирішення задачі класифікації можуть використовуватися методи: найближчого сусіда (Nearest Neighbor); k-найближчого сусіда (k-Nearest Neighbor); Байєсівські мережі (Bayesian Networks); індукція дерев рішень; нейронні мережі (neural networks).

Задачу *кластеризації* можна розглядати як логічне продовження ідеї класифікації і полягає в розподілі множини об'єктів на групи (кластери), при цьому в кожному кластері зібрані об'єкти, які схожі за параметрами. Варто зауважити, що на відміну від класифікації, кількість кластерів і їхніх характеристик визначають у процесі побудови кластерів, виходячи зі ступеня близькості поєднуваних об'єктів по сукупності параметрів. В онтологічному аналізі ця задача виникає на попередньому етапі та дозволяє побудувати набір базових класів онтології й встановити між ними ієрархічні відношення.

Задача *асоціації* – задача пошуку асоціативних правил (визначення взаємозв'язків), що полягає у визначенні наборів об'єктів, які часто зустрічаються серед множини подібних наборів. Відмінність асоціації від двох попередніх задач Data Mining: пошук закономірностей здійснюється не на основі властивостей аналізованого об'єкта, а між декількома подіями, що відбуваються одночасно.

Інші розповсюджені задачі Data Mining – задачі прогнозування, асоціації, визначення відхилень тощо – також можуть застосовуватися для вдосконалення онтологій шляхом обробки даних відповідних ПрО, доступних через Web.

Якщо дані, що обробляються в Data Mining, є ресурсами Web, то це вносить багато додаткових вимог до методів аналізу. Тому у Data Mining виокремлюють такий напрямок, як Web Mining. Системи Web Mining дозволяють знаходити закономірності в інформаційних ресурсах Web, застосовуючи технологію Data Mining для аналізу неструктурованої, неоднорідної, розподіленої і значної за обсягом інформації, яка знаходиться на Web-вузлах. У Web Mining можна виділити такі напрямки, як Web Content Mining і Web Usage Mining, Opinion Mining. В Web Mining можна виділити наступні етапи:

- *вхідний етап* (input stage) – отримання "сирих" даних із джерел (логи серверів, тексти електронних документів);
- *етап попередньої обробки* (preprocessing stage) – дані представляються у формі, необхідній для успішної побудови тієї чи іншої моделі;
- *етап моделювання* (pattern discovery stage);
- *етап аналізу моделі* (pattern analysis stage) – інтерпретація отриманих результатів.

Конкретні процедури кожного етапу залежать від поставленого завдання. У зв'язку із цим виділяють різні категорії Web Mining [37]: аналіз використання Web-ресурсів (Web Usage Mining); отримання Web-структур (Web Structure Mining); здобуття Web-контенту (Web Content Mining.

Значна частина даних – це ПМ-тексти. Саме в таких даних зазвичай міститься найбільш корисна інформація. Тому аналіз таких даних в Data Mining також виокремлюють в спеціальний підрозділ – Text Mining [38]. Технологія Text Mining містить процеси добування знань і високоякісної інформації з ПМ-масивів. Це звичайно відбувається за допомогою виявлення шаблонів і тенденцій за допомогою статистичних та лінгвістичних методів.

Значно підвищити ефективність Data Mining в усіх його напрямках дозволяє застосування фонових знань ПрО. Це дозволяє не шукати заново вже відомі користувачам закономірності та семантично збагатити зв'язки між параметрами (властивостями об'єктів, що аналізуються) за рахунок наявних знань щодо відношень між ними.

Одним з актуальних напрямків застосування фонових знань в Data Mining є аналіз Big Data та їх метаданих. Це обумовлено надзвичайно великими обсягами самих даних та їх динамічністю, що призводить до динамічності тих метаданих, що їх описують. Тому важливими вимогами до методів їх аналізу є швидкодія та наявність евристик, що дозволяють значно скоротити час аналізу. Наприклад, знання щодо відношення "клас-підклас" між параметрами метаданих дозволяє вдосконалити навчальну вибірку.

Це обумовлює необхідність отримання таких фонових знань, яке складається з наступних підзадач:

1) пошук ІР, що пертинентні задачі користувача;

2) здобуття з цих ІР необхідних фонових знань;

3) використання отриманих знань для аналізу даних.

У випадку аналізу Big Data ці задачі конкретизуються наступним чином:

1.1. Вибір сховища Big Data, в якому здійснюється пошук;

1.2. Пошук або створення онтології ПрО, що містить фонові знання щодо задачі користувача;

1.3. Аналіз метаданих Big Data з метою вибору набору даних, що пертинентні задачі користувача, з використанням фонових знань обраної онтології ПрО;

1.4. Генерація потрібного набору даних (підмножини Big Data за визначеними умовами) з використанням знань онтології;

2) Здобуття з онтології ПрО тих термінів та відношень між ними, які потрібні для більш ефективного аналізу великого обсягу інформації (наприклад, для зменшення кількості параметрів даних або для зменшення кількості записів за більш точними умовами відповідності задачі);

3) Використання отриманих знань для аналізу отриманого набору даних та для інтерпретації отриманого результату.

Таким чином, онтології дозволяють як аналізувати семантично метадані, що описують Big Data (наприклад, заміняти терміни в описі задачі на синоніми або на семантично подібні поняття, звужувати або розширювати запит), так і аналізувати самі дані (наприклад, використовуючи обмеження на можливі значення параметрів або виводячи з одних даних інші).

Семантичні Wiki-ресурси як джерело фонових знань для аналізу метаданих Big Data

Дослідження методів отримання фонових знань, які характеризують ПрО Big Data, є актуальним напрямком наукових досліджень, що спрямовані на обробку таких даних. Це обумовлено тим, що, як правило, для наборів Big Data не пропонуються пертинентні онтології тими особами або організаціями, що створюють та зберігають такі набори даних. У більшості випадків використання онтологічного аналізу для Big Data обмежується вибором онтології для визначення структури та змісту метаданих, яка не є специфічною для певної ПрО. Але використання знань ПрО може значно підвищити ефективність обробки.

Висока часова складність, на яку впливає великий розмір простору ознак у Big Data, викликає проблеми в використанні традиційних методів штучного інтелекту до такої інформації. Доцільно для їх оптимізації застосовувати наявні знання щодо ПрО, до якої відносяться як самі Big Data, так і задача, для вирішення якої здійснюється аналіз цих Big Data. Це дозволяє не здобувати ці знання повторно та використовувати їх для логічного виведення та встановлення відношень між елементами метаданих Big Data. Ефективність такого підходу визначається пертинентністю вибору бази знань та засобами подання самих знань. На сьогодні найбільш поширеним рішенням для подання розподілених знань з точки зору сумісного та повторного використання є онтології. Але побудова та пошук онтологій, що є пертинентними конкретній задачі, є складною проблемою. Значно простіше генерувати онтологічні структури за семантизованими Wiki-ресурсами. Такі онтології мають обмежену виразну здатність, але вони можуть створюватися автоматизовано за тим набором Wiki-сторінок, які обирає користувач. Крім того, такий підхід дозволяє відфільтровувати тільки ту інформацію, яка потрібна для вирішення задачі, що значно обмежує обсяг побудованої онтології та зменшує час на її використання.

Пошук пертинентної онтології неможливо повністю автоматизувати, хоча співставлення метаданих Big Data з метаописами онтологій в репозиторії дозволяє виконати попередній відбір. Проблема ускладнюється тим, що значна частина спеціалістів, що працюють з Big Data та їх метаданими, не мають достатнього досвіду у роботі з онтологіями. Тому доцільно застосовувати як джерело фонових знань такі ІР, що задовольняють наступним умовам:

1) досить прості для розуміння їх змісту та обсягу;

2) досяжні через Web;

3) зберігаються у відкритих форматах;

4) дозволяють автоматизовано генерувати онтології з фіксованим набором понять.

Таким вимогам відповідають семантично розмічені Wiki-ресурси. Виразні можливості Semantic MediaWiki [39] – семантичного розширення MediaWiki [40] – дозволяє явно фіксувати зміст відношень між Wiki-сторінками, які відповідають класам онтології.

Для того, щоб використовувати такий Wiki-ресурс як джерело фонових знань в аналізі Big Data, доцільно застосовувати Wiki-онтологію цього IP, яка є формалізованою моделлю знань ресурсу та дозволяє фіксувати характеристики його елементів, їх зв'язків, властивостей та відношень у формі, придатній для автоматичного оброблення, логічного виведення та аналізу. Wiki-онтологія – це окремий випадок онтології ПрО [41], виразні можливості якої обмежені відповідно до виразності Wiki та її семантичного розширення та не припускають застосування характеристик для об'єктних властивостей та властивостей даних. Використання цієї моделі для семантичної розмітки (як назви категорій та семантичних властивостей) забезпечує побудову уніфікованого набору ієрархічно пов'язаних категорій, шаблонів типових інформаційних об'єктів, їх семантичних властивостей та запитів, що їх використовують.

Важливою особливістю семантизованих Wiki-ресурсів є можливість генерації Wiki-онтології не для всієї сукупності сторінок, а тільки для певної підмножини, обраної користувачем явно переліком сторінок або за допомогою семантичного запиту (рис. 2). Параметрами такого запиту є категорії та умови щодо значень семантичних властивостей сторінок.

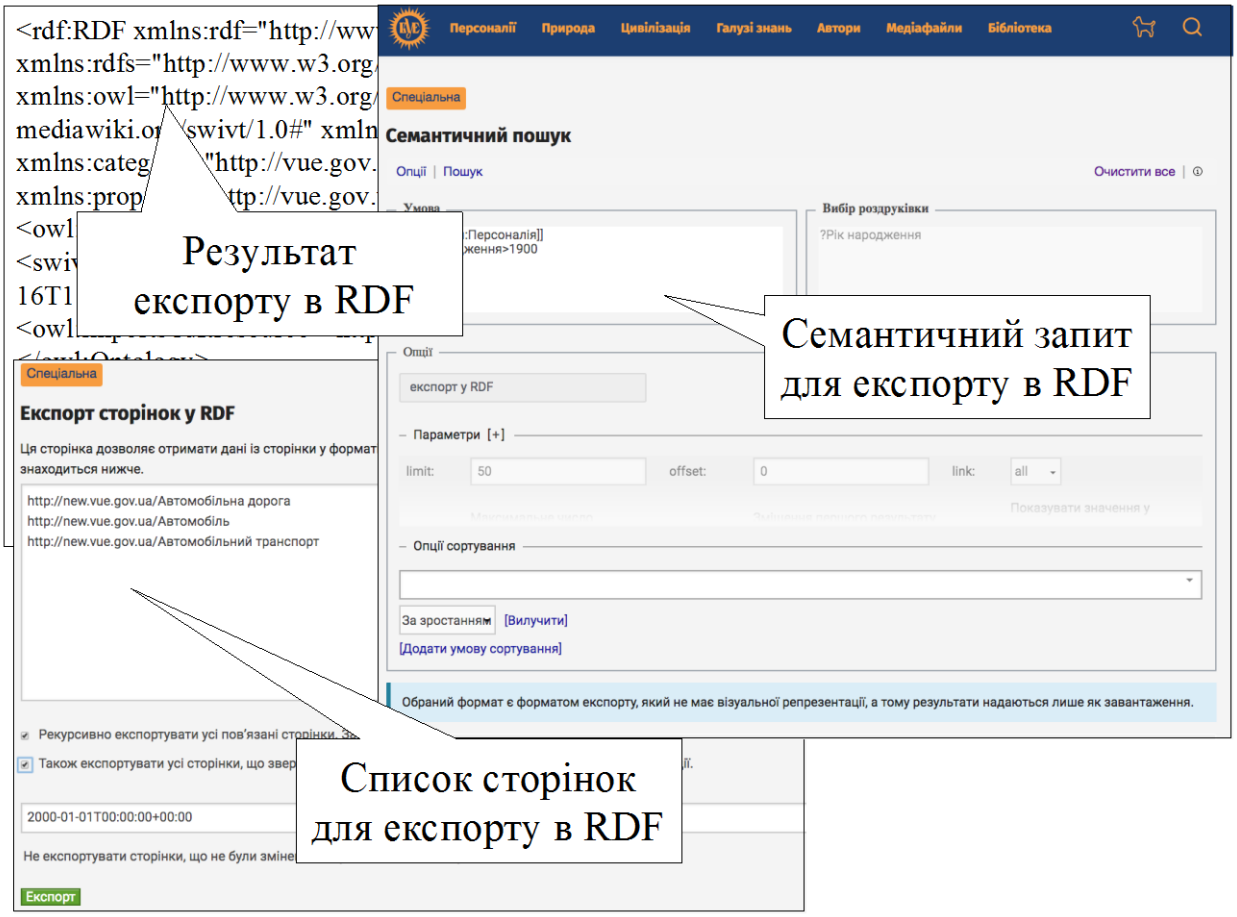

Рис. 2. Засоби Semantic MediaWiki для експорту інформації в RDF-форматі

Висновки

Для можливості інтеграції даних із внутрішніх та зовнішніх джерел та покращення керування Big Data, їх оцінювання та інтерпретації для виконання прикладних задач штучного інтелекту ми використали семантичні технології та онтології. Метадані є основними джерелами інформації про Big Data на протязі всього їх життєвого циклу. Для того, щоб правильно відбирати набори даних з Big Data, необхідно навчитись автоматично видобувати знання з їх метаданих за допомогою семантичних технологій. Доцільно застосовувати для цього такі джерела фонових знань як щодо цих метаданих, так і щодо ПрО, для якої потрібно аналізувати дані, як онтології та тезауруси.

Для семантичного аналізу метаданих ми використовуємо природномовні анотації, які входять до складу метаданих. Семантична обробка інформації метаданих дозволяє отримати від них неявні знання про самі дані. Аналіз текстів метаданих безпосередньо пов'язана із семантикою та певними логічними правилами, тому без метаданих та методів їх аналізу було б практично неможливо обійтися. Запропоновані нами методи аналізу природномовних анотацій є найбільш адекватним засобом співставлення семантики метаданих Big Data з тими задачами, для рішення яких вони можуть застосовуватися. На сьогоднішній день відсутні загальноприйняті, універсальні стандарти про метадані, а найбільш часто використовується універсальний стандарт опису метаданих Dublin Core.

Ми запропонували використовувати технології Wiki та їх семантичне розширення як джерело фонових знань щодо ПрО задачі користувача. Ці знання можуть також бути використані при оцінюванні семантичної близькості термінів домену для структурування елементів метаданих Big Data.

Новизна досліджень, які запропоновані у цій роботі, полягає у новому підході до інтеграції та структуруванні даних в інтелектуальних системах, який базується на семантичному аналізі та інтерпретації структурованих, частково структурованих та неструктурованих метаданих, які описують Big Data, та формуванні на їх основі пертинентного задачі користувача набору даних із застосуванням онтології предметної області.

Література

1. Метадані. https://uk.wikipedia.org/wiki/Метадані
2. Dublin Core Metadata Initiative. DCMI TYPE Vocabulary. http://dublincore.org/documents/demitype-vocabulary
3. Резніченко В А., Захарова О В., Захарова Е.Г. Електронні бібліотеки: інформаційні ресурси та сервіси. *Проблеми програмування*. 2005. № 4. С. 60–72.
4. Berners-Lee T., Hendler J., Lassila O. The semantic web. *Scientific american*. 2001. 284(5). P. 34–43.
5. Dunsire G., Willer M. Standard library metadata models and structures for the Semantic Web. *Library hi tech news*. 2011.
6. Когаловский М. Р. Метаданные, их свойства, функции, классификация и средства представления. Труды 14-й Всероссийской научной конференции «*Электронные библиотеки: перспективные методы и технологии, электронные коллекции*» – RCDL-2012. 2012. http://ceur-ws.org/Vol-934/paper3.pdf
7. Grotschel M., Lugger J. Scientific Information System and Metadata. Konrad-Zuse-Zentrum fur Informationstechnik. Berlin. http://www.zib.de/groetschel/pubnew/paper/groetschelluegger1999.pdf
8. Halshofer B., Klas W. A Survey of Techniques for Achieving Metadata Interoperability. ACM Computing Surveys. 2010. Vol. 42. N 2. Article 7.
9. Taylor C. An Introduction to Metadata. The University of Queensland, Australia. http://www.libraty.uq.edu.au/papers/ctmeta4.html
10. Lagose C. Metadata for the Web. Cornell University. CS 431 - March 2. 2005.
11. Feng L., Brussee R., Blanken H., Veenstra M. Languages for Metadata. In: Multimedia Retrieval. Data-Centric Systems and

Applications, Springer, 23–51. http://www.springerlink.com/ content/m276p88003533q86/.
12. Jeusfeld M.A. Metadata. In: Encyclopedia of Database Systems, Springer. 2009. P. 1723–1724. http ://www. springerlink.com/content/ h241167167r35055/.
13. Corcho O. Ontology based document annotation: trends and open research problems. *Intern. Journal of Metadata, Semantics and Ontologies*. 2006. Vol. 1. Is. 1. http://www.dia.fi.upm.es/~ocorcho/documents/IJMSO2006_Corcho.pdf .
14. Гладун А.Я., Рогушина Ю.В. Репозитории онтологий как средство повторного использования знаний для распознавания информационных объектов. *Онтология проектирования*. 2013. № 1 (7). С. 35–50.
15. Overbeek J. F. Meta Object Facility (MOF): investigation of the state of the art. 2006. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.96.4092&rep=rep1&type=pdf.
16. OWL Web Ontology Language. Overview. W3C Recommendation: W3C, 2009. – http://www.w3.org/TR/owl-features/.
17. Кобелев А. Е., Вязилов Е. Д. Сучасні підходи по створенню метаданих. *Сучасні проблеми дистанційного зондування Землі з космосу*. 2010. 7(4). С. 194–203. http://d33.infospace.ru/d33_conf/sb2010t4/194-203.pdf.
18. Unstructured_data. – https://en.wikipedia.org/wiki/Unstructured_data.
19. Рогушина Ю. В. Засоби та методи аналізу неструктурованих даних. *Проблеми програмування*. 2019. № 1. С. 57–77. http://pp.isofts.kiev.ua/ojs1/article/view/348/346.
20. Андон П.І., Рогушина Ю.В., Резніченко В.А., Киридон А.М., Арістова А.В., Тищенко А.О. Досвід використання семантичних технологій для створення інтелектуальних ВЕБ-енциклопедій (на прикладі розробки порталу Е-ВУЕ). *Проблеми програмування*. 2020. № 2–3. С. 246–258.
21. Rogushina J. Use of Semantic Similarity Estimates for Unstructured Data Analysis CEUR Vol-2577, Selected Papers of the XIX International Scientific and Practical Conference "*Information Technologies and Security*" (ITS 2019). Kyiv. 2019. P. 246–258. http://ceur-ws.org/Vol-2577/paper20.pdf.
22. Demchenko Y., De Laat C., Membrey P. Defining architecture components of the Big Data Ecosystem. In 2014 International Conference on Collaboration Technologies and Systems (CTS). 2014. P. 104–112.
23. Smith K., Seligman L., Rosenthal A., Kurcz C., Greer M., Macheret C., Eckstein A. "Big Metadata" The Need for Principled Metadata Management in Big Data Ecosystems. Proceedings of Workshop on Data analytics in the Cloud. 2014. P. 1–4).
24. Dey A., Chinchwadkar G., Fekete A., Ramachandran K. Metadata-as-a-service. 31st IEEE International Conference on Data Engineering Workshops. 2015. P. 6–9.
25. Chen M., Mao S., Liu Y. Big data: A survey. Mobile networks and applications. 2014. 19(2). P. 171–209.
26. Rogushina J., Gladun A., Pryima S. Use of Ontologies for Metadata Records Analysis in Big Data. Selected Papers of the XVIII International Scientific and Practical Conference "*Information Technologies and Security*" (ITS 2018). CEUR Vol-2318. http://ceur-ws.org/Vol-2318/paper5.pdf.
27. ISO 15489-1:2016 Information and documentation – Records management – Part 1: Concepts and principles.
28. ISO 15836-1:2017 Information and documentation – The Dublin Core metadata element set – Part 1: Core elements.
29. ISO 15836-2:2019 Information and documentation – The Dublin Core metadata element set – Part 2: DCMI Properties and classes.
30. ДСТУ ISO 15489-1:2018 Інформація та документація. Керування записами. Частина 1. *Поняття та принципи* (ISO 15489-1:2016, IDT).
31. ДСТУ ISO 15836-1:2018 Інформація та документація. Набір елементів метаданих Дублінського ядра. Частина 1. *Основні елементи* (ISO 15836-1:2017, IDT).
32. Weibel S.L., Koch T. The Dublin core metadata initiative. *D-lib magazine*. 2000. 6(12). P. 1082–9873.
33. Рогушина Ю.В. Використання тезаурусів для пошуку складних інформаційних об’єктів у Web на основі онтологій. *Проблеми програмування*. 2019. № 4. С. 11–27.
34. Гладун А.Я., Рогушина Ю.В. Семантичні технології: принципи та практики. – К.:ТОВ "ВД "АДЕФ-Україна". 2016. 308 с. http://eprints.isofts.kiev.ua/669/.

Гладун А.Я., Рогушина Ю.В. Data Mining: пошук знань в даних. К.:ТОВ "ВД "АДЕФ-Україна". 2016. 452 с.

36. Nigro H.O. ed. Data Mining with Ontologies: Implementations, Findings, and Frameworks: Implementations, Findings, and Frameworks. IGI Global. 2007. 289 p.
37. Kosala R., Blockeel H. Web mining research: A survey. *ACM Sigkdd Explorations Newsletter*. 2000. 2(1). P. 1–15. https://arxiv.org/pdf/cs/0011033.pdf
38. Berry M. W., Castellanos M. Survey of text mining. Survey of Text Mining:Clustering, Classification, and Retrieval. *Computing Reviews*. 2007. 45(9). P.548.
39. Krötzsch M., Vrandečić D., Völkel M. Semantic MediaWiki. International Semantic Web Conference. 2006. P. 935–942. https://link.springer.com/content/pdf/10.1007/11926078_68.pdf.
40. MediaWiki. URL: https://www.mediawiki.org/wiki/MediaWiki.
41. Rogushina J. Analysis of Automated Matching of the Semantic Wiki Resources with Elements of Domain Ontologies. *International Journal of Mathematical Sciences and Computing (IJMSC)*. 2017. Vol. 3. N 3. P. 50–58. URL: http://www.mecs-press.org/ijmsc/ijmsc-v3-n3/IJMSC-V3-N3-5.pdf.

References

1. Metadata. – https://uk.wikipedia.org/wiki/Метадані
2. Dublin Core Metadata Initiative. DCMI TYPE Vocabulary.– http://dublincore.org/documents/demitype-vocabulary. (in Ukrainian)
3. Reznichenko V.A., Zakharova O.V., Zakharova E.G. Electronic libraries: information resources and services. *Problems in programming*. 2005. № 4. P.60–72. (in Ukrainian)
4. Berners-Lee T., Hendler J., Lassila O. The semantic web. *Scientific american.* 2001. 284(5). P. 34–43.
5. Dunsire G., Willer M. Standard library metadata models and structures for the Semantic Web. Library hi tech news. 2011.
6. Kogalovsky M.R. Metadata, their properties, functions, classification and presentation means. Proc. of the 14th All-Russian Scientific Conference "*Digital Libraries: Promising Methods and Technologies, Electronic Collections*" – RCDL-2012, 2012. http:ceur-ws.org/Vol-934/paper3.pdf. (in Russian)
7. Grotschel M., Lugger J. Scientific Informa¬tion System and Metadata. Konrad-Zuse-Zentrum fur Informationstechnik. Berlin. http://www.zib.de/groetschel/pubnew/paper/groetschelluegger1999.pdf
8. Halshofer B., Klas W. A Survey of Techni¬ques for Achieving Metadata Interoperability. ACM Computing Surveys. 2010. Vol. 42. No. 2. Article 7.
9. Taylor C. An Introduction to Metadata. The University of Queensland, Australia. http://www.libraty.uq.edu.au/papers/ctmeta4.html
10. Lagose C. Metadata for the Web. Cornell University. CS 431 - March 2. 2005.
11. Feng L., Brussee R., Blanken H., Veenstra M. Languages for Metadata. In: Multimedia Retrieval. Data-Centric Systems and Applications, Springer, 23–51. http://www.springerlink.com/content/m276p88003533q86/.
12. Jeusfeld M.A. Metadata. In: Encyclopedia of Database Systems, Springer. 2009. P. 1723–1724. http ://www. springerlink.com/content/h241167167r35055/.
13. Corcho O. Ontology based document annotation: trends and open research problems. Intern. Journal of Metadata, Semantics and Ontologies. 2006. Vol. 1. Is. 1. http://www.dia.fi.upm.es/~ocorcho/documents/IJMSO2006_Corcho.pdf .
14. Gladun A., Rogushina J. Repositories of ontologies as a means of knowledge reuse for recognition of information objects. *Ontology of design.* 2013. N 1 (7). P. 35–50. (in Russian)
15. Overbeek J. F. Meta Object Facility (MOF): investigation of the state of the art. 2006. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.96.4092&rep=rep1&type=pdf.
16. OWL Web Ontology Language. Overview. W3C Recommendation: W3C, 2009. – http://www.w3.org/TR/owl-features/.
17. Kobelev A.E., Vyazilov E.D. Modern approaches to metadata creating. Modern problems of remote sensing of the Earth from space. 2010. 7 (4). P. 194–203. http://d33.infospace.ru/d33_conf/sb2010t4/194-203.pdf. (in Ukrainian)
18. Unstructured_data. – https://en.wikipedia.org/wiki/Unstructured_data.

19. ROGUSHINA J. (2019) Means and methods of unstructured data analysis. // Problems in programming, N 1, P. 57–77. http://pp.isofts.kiev.ua/ojs1/article/view/348/346. (in Ukrainian)
20. Andon P., Rogushina J., Grishanova I., Reznichenko V., Kyrydon A., Aristova A., Tyschenko A. (2020) Experience of the semantic technologies use for intelligent Web encyclopedia creation (on example of the Great Ukrainian Encyclopedia portal). Problems in programming, N 2-3. P. 246–258. (in Ukrainian)
21. Rogushina J. Use of Semantic Similarity Estimates for Unstructured Data Analysis CEUR Vol-2577, Selected Papers of the XIX International Scientific and Practical Conference "Information Technologies and Security" (ITS 2019). Kyiv. 2019. P. 246–258. http://ceur-ws.org/Vol-2577/paper20.pdf.
22. Demchenko Y., De Laat C., Membrey P. Defining architecture components of the Big Data Ecosystem. In 2014 International Conference on Collaboration Technologies and Systems (CTS). 2014. P. 104–112.
23. Smith K., Seligman L., Rosenthal A., Kurcz C., Greer M., Macheret C., Eckstein A. "Big Metadata" The Need for Principled Metadata Management in Big Data Ecosystems. Proceedings of Workshop on Data analytics in the Cloud. 2014. P. 1–4).
24. Dey A., Chinchwadkar G., Fekete A., Ramachandran K. Metadata-as-a-service. 31st IEEE International Conference on Data Engineering Workshops. 2015. P. 6–9.
25. Chen M., Mao S., Liu Y. Big data: A survey. Mobile networks and applications. 2014. 19(2). P. 171–209.
26. Rogushina J., Gladun A., Pryima S. Use of Ontologies for Metadata Records Analysis in Big Data. Selected Papers of the XVIII International Scientific and Practical Conference "Information Technologies and Security" (ITS 2018). CEUR Vol-2318. http://ceur-ws.org/Vol-2318/paper5.pdf.
27. ISO 15489-1:2016 Information and documentation – Records management – Part 1: Concepts and principles.
28. ISO 15836-1:2017 Information and documentation – The Dublin Core metadata element set – Part 1: Core elements.
29. ISO 15836-2:2019 Information and documentation – The Dublin Core metadata element set – Part 2: DCMI Properties and classes.
30. DSTU ISO 15489-1: 2018 Information and documentation. Records management. Part 1. Concepts and principles (ISO 15489-1: 2016, IDT). (in Ukrainian)
31. DSTU ISO 15836-1: 2018 Information and documentation. Dublin Core Metadata Element Set. Part 1. Basic elements (ISO 15836-1: 2017, IDT). (in Ukrainian)
32. Weibel S.L., Koch T. The Dublin core metadata initiative. *D-lib magazine*. 2000. 6(12). P. 1082–9873.
33. Rogushina J. The use of thesauri to search for complex Web information objects based on ontologies. *Problems of programming*. 2019. № 4, P. 11–27. (in Ukrainian)
34. Gladun A., Rogushina J. Semantic technologies: principles and practices. 2016. Kyiv. ADEF-Ukraine. 308 p. (in Ukrainian)
35. Gladun A., Rogushina J. Data Mining: search for knowledge in data. 2016. Kyiv. ADEF-Ukraine. 452 p. (in Ukrainian)
36. Nigro H.O. ed. Data Mining with Ontologies: Implementations, Findings, and Frameworks: Implementations, Findings, and Frameworks. IGI Global. 2007. 289 p.
37. Kosala R., Blockeel H. Web mining research: A survey. ACM Sigkdd Explorations Newsletter. 2000. 2(1). P. 1–15. https://arxiv.org/pdf/cs/0011033.pdf
38. Berry M. W., Castellanos M. Survey of text mining. Survey of Text Mining:Clustering, Classification, and Retrieval. Computing Reviews. 2007. 45(9). P. 548.
39. Krötzsch M., Vrandečić D., Völkel M. Semantic MediaWiki. International Semantic Web Conference. 2006. P. 935–942. https://link.springer.com/content/pdf/10.1007/11926078_68.pdf.
40. MediaWiki. URL: https://www.mediawiki.org/wiki/MediaWiki.
41. Rogushina J. Analysis of Automated Matching of the Semantic Wiki Resources with Elements of Domain Ontologies. International Journal of Mathematical Sciences and Computing (IJMSC). 2017. Vol. 3. N 3. P. 50–58. URL: http://www.mecs-press.org/ijmsc/ijmsc-v3-n3/IJMSC-V3-N3-5.pdf.

Одержано 23.10.2020

Про авторів:

Рогушина Юлія Віталіївна,
Кандидат фізико-математичних наук,
старший науковий співробітник.
Кількість наукових публікацій в
українських виданнях – 130.
Кількість наукових публікацій в
зарубіжних виданнях – 28.
http://orcid.org/0000-0001-7958-2557,

Гладун Анатолій Ясонович,
кандидат технічних наук, доцент,
старший науковий співробітник відділу комплексних досліджень інформаційних технологій.
Кількість наукових публікацій в
українських виданнях – 67.
Кількість наукових публікацій в
зарубіжних виданнях – 53.
https://orcid.org/0000-0002-4133-8169.

Місце роботи авторів:

Інститут програмних систем
НАН України, 03181, Київ-187,
проспект Академіка Глушкова, 40.

E-mail: ladamandraka2010@gmail.com.

Міжнародний науково-навчальний центр інформаційних технологій та систем НАН та МОН України,
03680, Київ, Україна,
проспект Академіка Глушкова, 40.

Тел.: +38(044) 526-2549.
E-mail: glanat@yahoo.com