УДК 681.3 http://doi.org/10.15407/pp2021.04.026

Ю.В. Рогушина, І.Ю. Гришанова

ДОСЛІДЖЕННЯ ПРИНЦИПІВ, МОДЕЛЕЙ ТА МЕТОДІВ ПАРАДИГМИ МЕНЕДЖМЕНТУ НАУКОВИХ ДАНИХ FAIR ДЛЯ АНАЛІЗУ МЕТАДАНИХ BIG DATA

Розглянуто базові принципи, моделі та методи парадигми менеджменту наукових даних FAIR (Findable, Accessible, Interoperable, Reusable як окремого випадку великих даних (Big Data), яка орієнтована на повторне використання результатів наукових досліджень. Проаналізовано, як властивості даних FAIR сприяють уніфікації й об'єднанню наукової інфраструктури у парадигмі відкритої науки. Запропоновано методи та програмні засоби, за допомогою яких властивості даних FAIR можуть відтворюватися у семантично розмічених Wiki-ресурсах, що побудовані на основі Semantic MediaWiki.
Ключові слова: метадані, Big Data, семантичні Wiki-ресурси.

Вступ

Цифрові технології усе ширше проникають у різні галузі, загострюючи проблему керування великими даними, вимагаючи оптимізації методів і підходів до обробки даних, а також ефективних способів збору даних. Одним з окремих випадків Big Data [1] є наукові дані великого обсягу. Незважаючи на те, що термін «великі дані» частіше пов'язують із соціальними мережами та фінансовою індустрією, спочатку великі дані генерувалися в рамках широкомасштабних наукових проєктів – зокрема, проєкту Великого адронного колайдера, що вимагало створення принципово нових засобів і методів обробки екстремально великих обсягів відомостей, які генерувалися у ході експериментів.

Тому, здійснюючи дослідження щодо застосування сучасних методів керування знаннями для інтелектуальної обробки Big Data та їх метаданих [2], доцільно проаналізувати існуючі підходи до подання та обробки великих наукових даних.

У поєднанні з величезною кількістю даних, які сьогодні необхідно обробляти в наукових і медичних дослідженнях, важлива системна вимога полягає в тому, щоб дані не губилися. Поширеним підходом для запобігання втраті даних є збереження їх у каталозі даних. Каталоги даних допомагають організовувати, структурувати та відстежувати метадані та згенеровані дані, щоб інформацію можна було зберігати та обмінюватися в межах організації. Використання каталогів даних може навіть привести до того, що вчені отримають більше цитат, оскільки вони створюють можливості для розробки або повторного використання попередніх досліджень.

Наприклад, каталог даних значно полегшує пошук відповідних даних. Щоб дозволити всій науковій спільноті отримати якнайбільшу користь від даних досліджень, повторне використання даних слід покращити надійним способом, захищаючи як виробника даних, так і зовнішніх повторних користувачів.

Підвищуючи якість і порівнянність даних досліджень, колеги-вчені повинні мати можливість повторно використовувати певний набір даних. Встановлення довіри між виробниками даних і зовнішніми повторними користувачами даних є проблемою, яка вимагає більш серйозних змін у поведінці вчених, ніж просто збільшення додаткових метаданих у наборі даних.

Щоб сприяти необхідним змінам, сьогодні кілька фінансових агенцій вимагають, аби одержувачі гранту надали план управління даними або план управління даними, з описом того, яким чином дані будуть доступні для колег-дослідників. Аналіз публікацій показує, що сьогодні не існує єдиної форми представлення великих наукових даних, доступної для комерціалізації, що ускладнює одержання вигоди від інвестицій у дослідницькі інфраструктури.

Саме в науці довелося вперше розбиратися зі збереженням і передачею великих масивів даних, з питаннями дотримання прав їхніх власників, створення безпечного інформаційного і правового середовища для користувачів наукового устаткування, обліку соціальних наслідків упровадження нових технологій тощо. В інших областях при роботі з Big Data акцент ставиться на ефективності використання конкретних методів і їхньої максимальної універсальності, а не на забезпеченні відкритості і доступності наявних даних. Детальний огляд таких досліджень наведено в [3].

Тому значний інтерес викликають FAIR – принципи керування даними без втручання користувача, що можна розглядати як один із перших кроків до формування цифрової інфраструктури для трансферу наукових результатів у форму, зрозумілу інвесторам, чиновникам, суспільству і придатну для контролю за обсягами наукових даних.

FAIR дані (FAIR_data) – це дані, які відповідають принципам знаходжуваності, доступності, інтероперабельності та повторного використання [4].

У березні 2016 консорціум науковців і організацій визначив базові принципи «FAIR Guiding Principles for scientific data management and stewardship", де був введений відповідний акронім FAIR (Findable, Accessible, Interoperable, Reusable) для зручності ведення дискусії.

Властивості FAIR даних

Дані FAIR мають наступні властивості:

1. *Findable*. Щоб використовувати дані, їх необхідно спочатку знайти там, де вони зберігаються. Метадані та дані повинні бути легко доступними як для людей, так і для комп'ютерів. Можливість машинної обробки метаданих є важливим для автоматичного виявлення наборів даних і служб, тому це важливий компонент процесу FAIRification.

F1. (Мета)даним призначається глобально унікальний і постійний ідентифікатор.

F2. Дані описуються докладними метаданими (визначені в R1 нижче).

F3. Метадані чітко і явно містять ідентифікатор даних, які вони описують.

F4. (Мета)дані реєструються або індексуються в пошуковому ресурсі.

Тож, для вирішення задачі пошуку даних, такі дані і додаткові матеріали мають мати достатньо повні метадані, мета-опис і унікальний постійний ідентифікатор.

2. *Accessible*. Коли користувач знаходить необхідні дані, він/вона повинен знати, як до них отримати доступ (можливо, включаючи аутентифікацію та авторизацію).

A1. (Мета)дані можна отримати за їхнім ідентифікатором за допомогою стандартизованого протоколу зв'язку.

A1.1 Протокол є відкритим, безкоштовним і універсальним.

A1.2 Протокол припускає процедуру аутентифікації та авторизації, якщо це необхідно.

A2. Метадані доступні, навіть якщо дані не доступні.

Задача доступності формулюється так, що метадані (з метаописами) і самі дані мають бути зрозумілі для людини та придатні для програмної обробки. Дані повинні зберігатися в надійному репозиторії.

3. *Interoperable*Зазвичай дані потрібно інтегрувати з іншими даними. Крім того, дані мають взаємодіяти із застосунками або робочими процесами для аналізу, зберігання та обробки.

I1. (Мета)дані використовують офіційну, доступну, спільну та широко застосовану мову для представлення знань.

I2. (Мета)дані використовують словники, які відповідають принципам FAIR.

I3. (Мета)дані включають кваліфіковані посилання на інші (мета)дані. Таким чином, задача інтероперабельності (сумісності) може бути вирішена за умови, коли для метаданих використовується формальна, доступна та широко вживана мова подання знань.

4. *Reusable*. Кінцевою метою FAIR є оптимізація повторного використання даних. Щоб досягти цього, метадані та дані повинні бути добре описані, щоб їх можна було відтворювати та/або комбінувати в різних налаштуваннях.

R1. Мета(дані) повинні бути детально описані набором точних і відповідних атрибутів.

R1.1. (Мета)дані видаються з чіткою та доступною ліцензією на використання даних.

R1.2. (Мета)дані пов'язані з їх походженням.

R1.3. (Мета)дані відповідають стандартам спільноти, що стосуються домену.

Отже, вимога повторного використання говорить про те, що дані і колекції мають однозначні ліцензії, які описують їх використання та чітку інформацію про джерело даних та їхнє походження.

Наприклад, агентство DARPA, що традиційно працює над проблемами інтеграції фундаментальних наукових праць із прикладними рішеннями, реалізує проєкт Automating Scientific Knowledge Extraction, що спрямований на автоматизацію процесів здобуття наукових знань з визначенням місцезнаходження нових інформаційних ресурсів, а також їхнього аналізу з метою отримання нових знань і генерації нових моделей.

У рамковій програмі ЄС із розвитку наукових досліджень і технологій «Горизонт — 2020» продемонстрована необхідність у створенні нових методів і підходів до обробки даних, таких як персоналізація та деперсоналізація даних, миттєвий збір даних тощо, рішення яких неможливе без ефективної організації керування потоками даних.

Для подолання різнорідності БД, для уніфікації й об'єднання наукової інфраструктури 2016 року було створено портал EOSC (European Open Science Cloud) [5] – віртуальне середовище із вільним доступом для збереження, керування, аналізу і передачі даних із усіх сфер знань в усі країни ЄС. Наукова цифрова інфраструктура ЄС містить множину регламентованих, відкритих, але спеціалізованих БД і репозиторіїв: BioMA, Global Marine Information System (GMIS), Central Core DNA Sequence Information System (CCSIS) і ін. Подібні ресурси постійно актуалізуються, мають чіткі регламенти представлення даних наукових досліджень, надають інструменти і механізми для керування контентом, однак тематика даних обмежена, а правила представлення метаданих не погоджені (різнорідні). Спроби створення універсальних сховищ даних, незалежних від тематики досліджень, призводять до розбалансування системи збереження, тому що репозиторії не мають обмежень щодо формату представлення даних і дескрипторів метаданих. Внаслідок цього інформаційна система ускладнюється, втрачає гнучкість і не забезпечує ефективного пошуку даних і їхнього повторного використання.

На рішення подібних проблем націлена ініціатива Go FAIR, що містить базові принципи поліпшення можливостей пошуку, забезпечення доступу до даних, їхньої сумісності і, що особливо важливо, повторного використання [6].

Згідно FAIR, функції пошуку, здобуття і представлення даних реалізують не користувачі, а інформаційна система. При цьому мова йде не тільки про власне дані і метадані, а й про алгоритми та інструменти керування ними. Крім того, до розробки підходів щодо керування науковими даними залучаються всі зацікавлені сторони: науково-дослідні організації й окремі вчені; оператори баз даних і видання, що публікують наукові статті і результати експериментів; організації, що фінансують ці наукові дослідження; виробники програмного забезпечення й інструментів обробки даних; компанії, що надають послуги з аналізу й інтерпретації даних. Важливо, що в коло зацікавлених сторін також включаються самі обчислювальні системи (алгоритми обробки даних) як самостійний об'єкт — залежно від їхнього рейтингу приймається рішення про включення обчислювального методу до конфігурації [7].

Для підтримки пошукових функцій (серед даних і метаданих) інформаційному блоку надається унікальний постійний глобальний ідентифікатор, а самі дані описуються розширеною множиною метаданих, які однозначно і явно включають ідентифікатор описуваних даних. Дані (та метадані) реєструються чи індексуються в доступному для пошуку ресурсі.

Для оптимізації доступу до даних потрібно керуватися наступними засадами: дані (метадані) можуть бути отримані за їхнім ідентифікатором за стандартизованими протоколами зв'язку; протокол до-

ступу до даних – відкритий і передбачає використання уніфікованого протоколу доступу — за необхідності для доступу до даних використовується процедура аутентифікації й авторизації, а метадані можуть бути доступні навіть за відсутності доступу до самих даних.

Має бути забезпечена сумісність даних не тільки з іншими даними, а й із застосунками та інструментами для їх аналізу, збереження й обробки: дані (метадані) використовують формальну, доступну і поширену мову опису даних; дані (метадані) використовують словники, що реалізовані відповідно до керівних принципів FAIR; дані (метадані) містять у собі повні посилання на інші дані (метадані).

Кінцева мета FAIR — оптимізація повторного використання даних та їх об'єднання в різних задачах: дані (метадані) докладно описують із застосуванням набору однозначних і релевантних атрибутів; дані (метадані) супроводжуються чіткою і доступною ліцензією на їхнє використання; дані (метадані) мають детальну історію їхнього походження; дані (метадані) подаються у відповідності зі стандартами тематичного наукового співтовариства.

Представлені елементи даних і метаданих взаємопов'язані, але водночас незалежні й відокремлені. Кожен з них визначає сукупність метрик (характеристик) – вимог, які передаються ресурсам, інструментам, словникам обробки даних для забезпечення їх повторного використання третіми сторонами, у тому числі коли вони не мають прямого відношення до науки. Водночас існує можливість керування рівнем входження в озера даних FAIR тих чи інших користувачів за рахунок градації у процесі визначення характеристик наданих ресурсів. Варіюючи і комбінуючи метрики опису об'єктів, можна досягти високого ступеня адаптивності представлення даних і метаданих в інформаційній системі. .

Керівні принципи FAIR не потребують будь - якої стандартизації чи конкретної технології підтримки. Принципи виступають як керівництво для створення даних для озер даних з урахуванням функціональності їх пошуку, доступності, сумісності і повторного використання.

Наведені принципи стосуються трьох типів сутностей: дані (або будь-який цифровий об'єкт), метадані (інформація про цей цифровий об'єкт) та інфраструктура. Наприклад, принцип F4 визначає, що і метадані, і дані реєструються або індексуються в ресурсі з можливістю пошуку (компонент інфраструктури).

У цілому FAIR подібний до open data, але існує ключова відмінність. Відкриті дані доступні кожному без будь-яких ліцензійних обмежень, угод, авторських прав чи патентів, тоді як FAIR допускає можливість доступу до даних (метаданих) у певний час і за певних умов. Інакше кажучи, FAIR-дані можуть бути як відкритими, так і частками, якщо вони доступні лише визначеній групі користувачів. Такий підхід є більш гнучким і дозволяє характеризувати дані на кожному етапі їхнього життєвого циклу.

Наприклад, у процесі фізичного експерименту дані доступні тільки групі експериментаторів, потім — науковому співтовариству з метою їхньої інтерпретації, а після обробки переходять у загальний доступ (open data) як результат експерименту. На практиці наукові дані неодноразово переходять через такі стадії «відкритості». У переважній більшості випадків персональні і комерційні дані не можуть бути загальнодоступними, це суперечить ідеям open data, але допустимо в FAIR.

Зараз багато європейських дослідницьких інфраструктур (DTU Library, International Neuroinformatics Coordinating Facility, TU Dublin, Biobanking and Biomolecular Resources Research Infrastructure of Czech Republic, Radboud University тощо) використовують концепцію FAIR для надання доступу до своїх наукових даних. Створено і розвиваються методичні рекомендації та інструкції з представлення даних відповідно до FAIR.

У рамках програми «Горизонт-2020» ініційовано проєкт PaNOSC, що поєднує шість великих європейських дослідницьких інфраструктур (ESRF, European XFEL, CERIC-ERIC, ELI Delivery Consortium, ESS, ILL) для розвитку Європейської хмари відкритої науки (European Open Science Cloud) — універсального міждисциплінарного репозиторію наукових даних із відкритим

доступом для дослідників у всіх галузях. У рамках PaNOSC дослідникам з таких галузей, як хімія, біологія, матеріалознавство тощо надаються сервіси й інструменти для збереження, пошуку й аналізу даних, отриманих на нейтронній і фотонній дослідницьких інфраструктурах.

За рахунок використання постійних унікальних ідентифікаторів реалізується можливість передачі метаданих між сервісами. Це дозволяє збільшити на порядок можливість повторного використання результатів наукового дослідження в масштабах прямо не зв'язаних тематичних галузей наукових досліджень. У перспективі мова йтиме про забезпечення для усього світового наукового співтовариства, незалежно від тематики досліджень, доступу через EOSC до експериментальних даних від європейських дослідницьких інфраструктур.

На поточному етапі досліджень мова йде не стільки про об'єкт цифрової наукової інфраструктури (база даних, озеро даних), скільки про послугу керування великими даними: реалізується механізм керування множиною даних, доступним різним типам користувачів — науковим співтовариствам, державним структурам тощо.

Парадигма Відкритої Науки

Парадигма Відкритої Науки є спробою світової наукової спільноти розв'язати проблему наукової невідтворюваності (scientific irreproducibility) [8, 9]. «Наукова невідтворюваність – неспроможність повторити чужі експерименти та дійти того ж висновку – [10]. Для цього запропоновано базові принципи, на яких повинні ґрунтуватися наукові дослідження:

- *Відкритий доступ*. Тобто результати досліджень, наукові публікації, які поширюються онлайн і без затрат або інших перешкод, повинні мати вільний доступ.
- *Відкрита наука*. Дослідники діляться своїми методами, програмним кодом та даними досліджень через централізовані спеціалізовані репозиторії.
- *Відкриті дані*. Дані повинні бути вільно доступні кожному для використання, повторного аналізу і публікації на свій розсуд, без обмежень з боку авторського права, патентів або інших механізмів контролю.

Виходячи з цих принципів, дослідники мають не тільки публікувати свої дані в Web, а й надавати до них доступ у такому вигляді (і форматі), щоб забезпечити їх сумісність із поширеними стандартами, а також можливість їх повторного використання. Проблема ускладнюється тим, що йдеться про дані великого обсягу, які швидко змінюються та слабо структуровані, тобто їх подання та збереження базується на технологіях Big Data.

Впровадження FAIR

Із початку 2018 року спільнота GO FAIR працює над впровадженням Керівних принципів FAIR. Результатом цих спільних зусиль є структура з трьох пунктів, яка формулює основні кроки до кінцевої мети – глобального Інтернету даних і послуг FAIR, де дані є знаходжуваними, доступними, інтероперабельними та повторно використовуваними (FAIR) для машин.

На сайті https://www.go-fair.org/fair-principles/ надано докладне роз'яснення принципів і практичне керівництво щодо того, як розробляти та використовувати FAIR дані, де їх шукати [11].

Структура FAIRification дає практичні вказівки «як це зробити» для зацікавлених сторін, які прагнуть бути FAIR.

Крім того, дотримуючись цієї структури, зацікавлені сторони можуть бути впевнені, що їхні зусилля щодо FAIRification будуть оптимально скоординовані із зусиллями інших зацікавлених сторін у спільноті GO FAIR. Структура з трьох пунктів максимізує повторне використання існуючих ресурсів, максимізує взаємодію та прискорює зближення стандартів і технологій, що підтримують дані та послуги FAIR.

Як правило, процес FAIRification починається, коли спільнота практиків розглядає свої вимоги до метаданих, що стосуються домену ПрО, та інші міркування політики, і формулює ці міркування як компоненти метаданих, що використовуються машиною. Для складання цих міркувань можна керуватися розділом Метаданих для машин (M4M) Workshops.

Схеми метаданих для повторного використання, створені в M4M, складають частину більшого профілю впровадження FAIR (FIP).

Профіль впровадження FAIR, у свою чергу, керує вибором і конфігурацією інфраструктури FAIR. Наприклад, використання точок даних FAIR (FDP) або FAIR Digital Objects (FDO), які сприяють створенню глобального Інтернету даних і послуг FAIR.

Розроблений підхід допомагає широкому колу зацікавлених сторін побачити, що для них означає «справедливий процес» на практиці, і ввійти в новий ландшафт FAIR. Це не тільки зберігає пріоритет практичних елементів FAIRification, а й дозволяє розподілити підхід до координації громади, який необхідний для швидкого масштабування та конвергенції.

З квітня 2020 року функціонують робочі групи, які розробляють методи, інструменти та документацію навколо платформи процесу FAIRification:

• Робоча група Metadata 4 Machines

• Робоча група FAIR Implementation Profile

• Робоча група FAIR Data Point

У зв'язку з викликами, пов'язаними з пандемією COVID-19, 3-кроковий FAIRification Framework активно розробляється в кількох проєктах. Безпосередньою метою цих трьох робочих груп є створення посібника, який об'єднає методи та ресурси для проведення семінарів M4M, для створення профілів впровадження FAIR та для встановлення точок даних FAIR.

Проаналізувавши розробки, що пов'язані зі створенням та використанням FAIR для наукових Big Data, можна відмітити доцільність застосування онтологій ПрО, що відповідають окремим галузям наук або є основою для інтеграції інформації з різних галузей та з різних країн. Такі онтології можуть бути використані як джерело знань для метаданих таких Big Data. Але це викликає потребу в автоматизованій побудові відповідних онтології – за пертинентними інформаційними ресурсами різного ступеня структурованості та з використанням уже існуючих онтологічних структур.

Створення джерела даних FAIR на базі Semantic MediaWiki.

Семантизовані Wiki-ресурси, такі як Semantic MediaWiki, які дозволяють створювати семантичні дані та базуються на використанні стандартів Semantic Web, надають потужне рішення для спільного редагування даних та їхніх метаописів, створення різних довільних наборів властивостей у шаблонах цих метаописів, з одночасним поданням їх як в машинно-оброблюваній формі, так і формі, придатній для розуміння людиною, що в результаті дає можливість оперувати цими даними, автоматизовано керувати, проводити аналіз, публікувати.

Однак Semantic MediaWiki не містить адекватних і ефективних вбудованих функцій імпорту та експорту між інтероперабельними форматами Semantic Web (таких, як RDF або OWL) і внутрішнім Wiki-форматом. Для вирішення цієї задачі розробляються проєкти, як наприклад, RDFIO (pharmb.io/project/rdfio.) – набір інструментів для імпорту RDF-даних в галузі біомедичних досліджень в Semantic MediaWiki з метаданими, які необхідні для експорту цих даних у формат RDF або OWL [12] .

Семантизація програмних засобів, що використовують онтології для керування метаданими наукових даних, є перспективним напрямком, який дозволить незалежно від галузі наукових досліджень, прийнятих у цій галузі стандартів, типів даних, розробити інформаційний ресурс (IP), який спільно створюється і спільно використовується відповідно до усіх принципів FAIR.

Використання для цього семантичного розширення Wiki-технології [13, 14] Semantic MediaWiki [15], що історично себе показало досить потужним інструментом, має широке використання, відкритий код та відкриті принципи розробки і підтримки, постійно розвивається, – є найбільш виправданим вибором.

Вбудовані можливості Semantic MediaWiki з завантаження файлів різного формату і додавання до них метаданих з різним набором атрибутів, які можливо змінювати, доповнювати, та які мають можливість обробки програмно, водночас зрозумілі звичайній людині. Простий і доступний інтерфейс, проста мова розмітки, інтуїтивний інтерфейс, можливість спільної роботи, відкрите розміщення в Web, роблять MediaWiki найкращим рішенням.

Розглянемо відповідність IP, що створений у середовищі Semantic MediaWiki, основним вимогам FAIR.

Findable

Першим кроком для використання даних є їх пошук.

Semantic MediaWiki надає можливості семантичного пошуку – на основі семантичних властивостей та категорій окремих Wiki-сторінок, які можуть розглядатися як гнучкий набір метаданих. Забезпечується можливість машинної обробки таких метаданих. .

F1. Таким (Мета)даним в Semantic MediaWiki надається глобальний унікальний і постійний ідентифікатор – кожній семантичній властивості або категорії в IP відповідає окрема Wiki-сторінка з унікальним ідентифікатором, яка описує характеристики та сферу застосування відповідного фрагменту метаданих. Крім того, забезпечується можливість перегляду того, де саме (на яких сторінках) використовується цей фрагмент метаданих та яких значень він набуває.

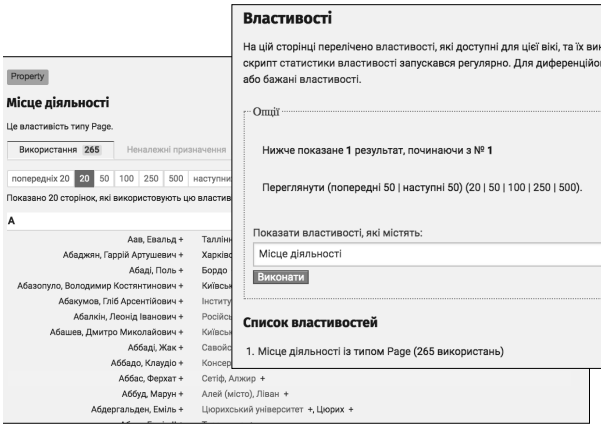

Рис.1. Пошук метаданих в середовищі Semantic MediaWiki

Певні набори метаданих, які характеризують типові для IP інформаційні об'єкти, можна об'єднувати за допомогою стандартного механізму «Форми» та «Шаблони» MediaWiki.

F2. Механізми MediaWiki мають можливість додавати докладні метадані до даних, які завантажуються до стандартного сховища MediaWiki, або вказувати гіперпосилання на зовнішнє сховище. Для цього також використовуються семантичні властивості та категорії.

F3. Метадані чітко і експліцитно містять ідентифікатор даних, які вони описують. Ідентифікатор даних в Semantic MediaWiki подається окремим значенням властивості, що додається до набору властивостей метаданих (рис.2).

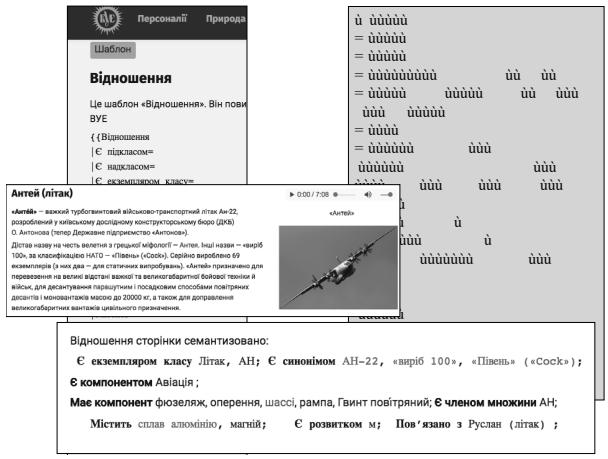

Рис.2. Приклад присвоювання значень семантичних властивостей на основі шаблонів у середовищі Semantic MediaWiki.

F4. Метадані реєструються або індексуються в пошуковому ресурсі. Оскільки Semantic MediaWiki є Web-застосуванням, вона містить файл налаштувань, де є можливість вказати тип відкритості – відкрита чи закрита система. Цей тип вказується лише раз у момент першого запуску й записується в файл налаштування LocalSettings.php. Для відповідності цьому принципу, треба встановити у цьому файлі певні права:

 # Enable/disable reading by anonymous users

$wgGroupPermissions['']['read'] = true;*

 # Enable/disable anonymous editing

$wgGroupPermissions['']['edit'] = true;*

 # Allow new user registrations

$wgGroupPermissions['']['createaccount'] = true;*

Для ефективного знаходження даних такі дані і додаткові матеріали до них повинні мати достатньо повні метадані, а також метаопис і унікальний постійний ідентифікатор. Дані, розміщені на сторінках Semantic MediaWiki з вказаними нала-

штуваннями, подаються до Web відкрито і добре індексуються глобальними пошуковими системами, такими як Google та Bing.

Accessible

Коли користувач знайде необхідні дані, він повинен знати, як до них можна отримати доступ. Можливо, включаючи аутентифікацію та авторизацію. Таку інформацію в Semantic MediaWiki можна доповнювати окремими атрибутами на сторінках із метаописами.

A1. (Мета)дані можна отримати за їхнім ідентифікатором за допомогою стандартизованого протоколу зв'язку. Зазвичай Web-системи розміщуються на серверах і мають стандартний http або захищений https протокол доступу.

A1.1 Протоколи http та https є відкритими, безкоштовними і універсальними. Вони є базовими протоколами Інтернет і Web.

A1.2 Протокол https за необхідності припускає процедуру аутентифікації та авторизації.

A2. Метадані доступні, навіть якщо дані більше не доступні. Оскільки метадані розміщуються окремо на сторінках MediaWiki (фізично – в БД MediaWiki), цей принцип виконується.

Отже, метадані (з метаописами) і самі дані в Semantic MediaWiki (з відповідними налаштуваннями серверів та інфраструктури, заданими наборами метаданих із властивостями унікальних ідентифікаторів) зрозумілі як для людини, так і для програмної обробки й зберігаються в надійному репозиторії, тобто відповідають зазначеним вище вимогам.

Interoperable

Дані, що представлені за допомогою Semantic MediaWiki, можуть бути інтегровані з іншими даними. Крім того, дані можуть взаємодіяти з іншими застосуваннями або робочими процесами для аналізу, зберігання та обробки.

I1. (Мета)дані в Semantic MediaWiki використовує базові стандарти Semantic Web, експорт даних в XML, RDF, а за можливості встановлення інших додаткових плагінів, то і в OWL або інші спеціалізовані стандарти. PDF, CSV, LaTeX, тощо. Експорт результатів семантичного пошуку в формат RDF входить до функціоналу Semantic MediaWiki.

I2. MediaWiki є відкритою системою, тобто за умови відкритого типу встановлення (як вказано в п.F4 і A1), система публікації даних і метаданих, базована на Semantic MediaWiki, може використовувати, інтегрувати, імпортувати (стандартна функція імпорту або додатковий спеціальний плагін) будь-які словники, онтології, які подані стандартною мовою Semantic Web, що відповідає принципам FAIR.

I3. (Мета)дані в Semantic MediaWiki включають посилання на інші метадані. Стандартний функціонал MediaWiki дозволяє додавати необмежену кількість посилань на інші джерела, а можливість додавати різні додаткові властивості дозволяє розширювати набори властивостей до даних в залежності від задачі.

Отож, задача інтероперабельності (сумісності) в Semantic MediaWiki вирішується простою публікацією додаткових властивостей, які для людини показуються простими описами, а для програмного запиту надаються в одному із загальновживаних форматів XML, JSON, RDF.

Reusable

Повторне використання даних у Semantic MediaWiki забезпечується механізмами семантичних властивостей та категорій Semantic MediaWiki, які підтримують гнучке внесення змін до структури метаданих ресурсу та автоматизацію деяких елементів цього процесу. Такі властивості однозначно описані на окремих Wiki-сторінках, тому їх можна відтворювати та комбінувати в різних налаштуваннях.

R1. Можливість Semantic MediaWiki додавати необмежену кількість атрибутів та категорій кожній сторінці забезпечує вимогу FAIR щодо того, що метадані повинні бути детально описані множиною точних і відповідних атрибутів.

R1.1. Механізми MediaWiki розроблялась для використання у відкритому середовищі з різними правами власності, тому вона містить певні механізми додавання інформування про різні типи ліцензій до різних об'єктів.

R1.2. Походження даних та метаданих досягається в Semantic MediaWiki шля-

хом додавання додаткового метаопису про джерело та умови походження даних.

R1.3. (Мета)дані в Semantic MediaWiki відповідають стандартам спільноти, які стосуються домену ПрО, що може бути формалізована за допомогою онтології цієї ПрО, яка використовується як основа для семантичної розмітки Wiki-сторінок і визначає набори даних, їхні імена та можливі значення.

Висновки

Наведений вище аналіз парадигми менеджменту наукових даних FAIR та порівняння основних вимог FAIR до подання даних та метаданих з виразними властивостями середовища Semantic MediaWiki свідчить про те, що інформаційні ресурси, які створюються в цьому середовищі, відповідають сучасним вимогам до відкритих даних великого обсягу. Це уможливлює використання таких Wiki-ресурсів як основи для побудови та семантичного аналізу метаданих у релевантних предметних областях. Одним із можливих напрямків застосування цього підходу є використання структури метаданих Wiki-ресурсу для аналізу метаданих Big Data.

Даний підхід апробовано в процесі створення бази знань портальної версії Великої української енциклопедії (vue.gov.ua) [16], яка є джерелом інтероперабельних знань, що придатні для повторного використання в інших інтелектуальних застосуваннях.

Література

1. Hurwitz, J., Nugent, A., Halper, F., Kaufman, M., 2013,. Big Data. New York.
2. 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.
3. Балякин А., Малышев А. Управление большими данными в исследовательских инфраструктурах // Открытые системы. СУБД, 2020, № 03. – https://www.osp.ru/os/2020/03/13055606.
4. FAIR_data. https://en.wikipedia.org/wiki/FAIR_data.
5. Gomez-Diaz, T., Recio, T. (2021). Open comments on the Task Force SIRS report: Scholarly Infrastructures for Research Software (EOSC Executive Board, EOSCArchitecture).
6. The FAIR Guiding Principles for scientific data management and stewardship. Available from: https://www.nature.com/articles/sdata201618.
7. The FAIR data principles. Available from: : https://www.force11.org/group/fairgroup/fairprinciples (дата обращения: 29.08.2020).
8. The Irreproducibility Crisis of Modern Science – CUSES, Consequences and the Road to Reform, National Association of Scholars, 2018, Available from: https://www.nas.org/reports/the-irreproducibility-crisis-of-modern-science.
9. Challenges in irreproducible research, Nature, 18-10-2018, Available from: https://www.nature.com/collections/prbfkwmwvz/.
10. Baker, 1,500 scientists lift the lid on reproducibility. Nature, 533(7604): 452-454. (2016) doi:10.1038/533452a, Available from: https://www.nature.com/articles/533452a.
11. Three-point FAIRification Framework. Available from: https://www.go-fair.org/how-to-go-fair/.
12. Lampa, S., Willighagen, E., Kohonen, P., King, A., Vrandečić, D., Grafström, R., Spjuth, O. 2017. RDFIO: extending Semantic MediaWiki for interoperable biomedical data management. Journal of biomedical semantics, 8(1), 2017, P.1-13.
13. Manual:What is MediaWiki?. Available from: https://www.mediawiki.org/wiki/Manual:What_is_MediaWiki%3F.
14. MediaWiki. Available from: https://www.mediawiki.org/wiki/MediaWiki.
15. *Krötzsch M., Vrandečić D., Völkel M.* Semantic mediawiki. International semantic web conference, 2006, pp. 935-942. Available from: https://link.springer.com/content/pdf/10.1007/11926078_68.pdf.
16. Rogushina J.V., Grishanova I.J. Ontological methods and tools for semantic extension of the media WIKI. Проблеми програмування, № 2-3, 2020. С.-61-73. Available from: http://pp.isofts.kiev.ua/ojs1/article/download/398/437

References

1. Hurwitz, J., Nugent, A., Halper, F., Kaufman, M. (2013). Big Data. New York.

2. 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. Available from: http://ceur-ws.org/Vol-2318/paper5.pdf [Accessed 18/11/2021]
3. Baliakin A., Malyshev A. (2020) Management of Big Data in research infrastructures. Open systems, 03. Available from: https://www.osp.ru/os/2020/03/13055606 [Accessed 18/11/2021]
4. FAIR_data. Available from: https://en.wikipedia.org/wiki/FAIR_data [Accessed 18/11/2021]
5. Gomez-Diaz, T., Recio, T. (2021). Open comments on the Task Force SIRS report: Scholarly Infrastructures for Research Software (EOSC Executive Board, EOSCArchitecture). arXiv preprint arXiv:2108.06127.
6. The FAIR Guiding Principles for scientific data management and stewardship. Available from: https://www.nature.com/articles/sdata201618 [Accessed 18/11/2021]
7. The FAIR data principles. Available from: https://www.force11.org/group/fairgroup/fairprinciples [Accessed 18/11/2021]
8. The Irreproducibility Crisis of Modern Science – CUSES, Consequences and the Road to Reform, National Association of Scholars, (2018), Available from: https://www.nas.org/reports/the-irreproducibility-crisis-of-modern-science [Accessed 18/11/2021]
9. Challenges in irreproducible research, Nature, 18-10-2018, Available from: https://www.nature.com/collections/prbfkwmwvz/ [Accessed 18/11/2021]
10. Baker, 1,500 scientists lift the lid on reproducibility. Nature, 533(7604): 452-454. (2016) doi:10.1038/533452a, Available from: https://www.nature.com/articles/533452a.
11. Three-point FAIRification Framework https://www.go-fair.org/how-to-go-fair/ [Accessed 18/11/2021]
12. Lampa, S., Willighagen, E., Kohonen, P., King, A., Vrandečić, D., Grafström, R., & Spjuth, O. (2017). RDFIO: extending Semantic MediaWiki for interoperable biomedical data management. *Journal of biomedical semantics,* 8(1), 1-13 [Accessed 18/11/2021]
13. Manual:What is MediaWiki? Available from: https://www.mediawiki.org/wiki/Manual:What_is_MediaWiki%3F [Accessed 18/11/2021]
14. MediaWiki. Available from: https://www.mediawiki.org/wiki/MediaWiki.
15. Krötzsch M., Vrandečić D., Völkel M. (2006) Semantic mediawiki. *International semantic web conference,* pp. 935-942. Available from: https://link.springer.com/content/pdf/10.1007/11926078_68.pdf [Accessed 18/11/2021]
16. Rogushina J.V., Grishanova I.J. (2020) Ontological methods and tools for semantic extension of the media WIKI. *Problems in Programming,* 2-3, P.-61-73. Available from: http://pp.isofts.kiev.ua/ojs1/article/download/398/437 [Accessed 18/11/2021].

Отримано: 20.11.2021

Об авторах:

Рогушина Юлія Віталіївна,
Канд.фіз.-мат.наук, старший науковий співробітник Інституту програмних систем НАН України, публікації в українських виданнях – 170, публікації в іноземних журналах – 35, індекс Хірша (Scopus) – 5, ORCID http://orcid.org/0000-0001-7958-2557.
e-mail: ladamndraka2010@gmail.com

Гришанова Ірина Юріївна,
науковий співробітник Інституту програмних систем НАН України, публікації в українських виданнях – 19, публікації в іноземних журналах – 3, індекс Хірша (Scopus) – 1, ORCID http://orcid.org/0000-0003-4999-6294.

Місце роботи авторів:

Інститут програмних систем
НАН України, 03181, Київ-187,
проспект Академіка Глушкова, 40,
e-mail: ladamandraka2010@gmail.com,
i26031966@gmail.com
066 550 1999.