Проміжний звіт

Інтеграція анотацій в Semantic Web. Модель інформаційного середовища

Новицький О.В.

2009

1. Зв’язані дані (Linked Data)

Однією з моделей семантичного вебу можна вважати модель зв’язаних даних.

Основні принципи Linked Data висвітлено в [**1**].

Цінність зв'язаних даних полягає в тому, що корисність даних та їх інформативність збільшується чим більше вони пов'язані з іншими даними. Основні принци побудови що визначені в роботі є:

1. Слід використовувати URIs в якості імен для сутностей
2. Слід використовувати HTTP URIs щоб люди могли побачити ці імена.
3. Коли хтось шукає в URI, слід представляти корисну інформацію.
4. Містити посилання на інші URI з тим щоб вони могли дізнатися більше сутності.

Дана модель вбачається проміжною для побудови семантичного Інтернету. Водночас Semantic Web, можливо представити, як бачення або ціль де семантично багата анотація даних використовується машинними агентами для пошуку інформації. Ми перебуваємо на шляху до цієї мети або цілі, але при цьому інтерпретуючи, що Semantic Web - є більше процесом чим станом.

Саме поняття Semantic Web є багатогранним і чітко не визначеним. Це поняття слід розуміти як здатність машин обробляти та розуміти дані, які розміщені в інформаційних ресурсах. Частим є запитання як відноситься Semantic Web та Linked Data. Фактично LD це перенесення технології гіперпосилань веб-документів для зв’язування RDF трійок.

Деякі автори в свої роботах [**2**], [**3**], асоціюють це поняття з Semantic Web однак цей підхід повністю не відображає всіх аспектів Semantic Web, наприклад таких, як динамічність Semantic Web. Окрім того що інформація постійно змінюється в середовищі Semantic Web функціонують і агенти, які цю інформацію оброблюють і можуть вносити до неї певні зміни в рамках концепції LD. Тому ми схильні до думки що LD є прикладом частинної реалізації Semantic Web. І дає відповідь як шукати документи в Semantic Web.

Застосування обох принципів призводить до створення спільних даних в Веб. Ці спільні дані часто називають Веб-Даними або Semantic Web.

Доступ до Веб-даних можна отримати з використанням LD браузерів, так само, як традиційні доступ до веб-документів, за допомогою HTML-браузерів.

Однак замість того, щоб переміщатися між посилання HTML-сторінок, LD браузери дозволяють користувачам переміщатися між різними джерелами даних, виконавши RDF посилання.

Це дозволяє почати з одного джерела даних, а потім пройти через потенційно нескінченну кількість джерел Web даних, пов'язаних RDF посилання.

Наприклад, при пошуку даних про людину з одного джерела, а користувачу може бути цікава інформації у рідне місто особи.

При переході по RDF посиланню, користувач може перейти до інформації про місто, яка містяться в іншому наборі даних.

Точно так само, як в традиційних веб-документах може бути про скановано всі гіпертекстові посилання, можуть бути просканувані переходи RDF-посиланнь Веб Даних. В даному випадку робота з такими даними має ряд переваг, пошукові системи можуть надавати складні запити можливості, аналогічні тим, які передбачені звичайними реляційними базами даних. Оскільки результати запиту до структурованих даних є зновуж структуровані дані, а не лише посилання на HTML-сторінки, вони можуть бути негайно оброблені, дозволяючи, таким чином, новий клас програм, заснованих на Веб-Даних.

Основні принципи LD тісно пов'язані з архітектурою Інтернету. Одними з основних понять в архітектурі є *ресурс* та *представлення* детальний зміст цих понять наведено в [**4**].

2. Ресурс

Щоб опублікувати дані в Інтернеті, ми спочатку повинні *ідентифікувати елементи,* що *представляють інтерес* в нашому домені. Вони є сутності чиї властивості і відносини ми хочемо описати в даних. В термінології Веб Архітектури, всі елементи, що представляють інтерес, називаються *ресурсами.*

В [**5**] з розрізняти два види ресурсів: *інформаційні ресурси та не-інформаційних ресурси* (які також називається "інші ресурси"). Ця різниця є дуже важливим у цьому контексті LD. Всі ресурси, які ми знаходимо на традиційні веб-документа, як, наприклад, документи, зображень та інших мультимедійні файлі, є інформаційними ресурсами.

Поняття «інформаційний ресурс» введено в [**4**] тому, що було відмічено корисність його використання для технологій мережі Інтернет.

Насправді в Technical Architecture Group (TAG) не дає чіткої відповіді на питання різниці між інформаційними та не інформаційними ресурсами. Якщо взяти за основу підхід який викладений в [**6**], тобто, якщо на GET запит при розіменованні повертається результат з кодом 303 то це не інформаційний ресурс. Дотримуючись такого правила постає питання який ресурс відносити до інформаційних або не до інформаційних.

Але багато сутностей, дані про які ми хочемо спільно використовувати не є даними в прямому розумінні цього слова, наприклад: особи, фізичні об'єкти, місця, наукові концепції, і так далі.

Як правило, всі " об'єкти реального світу", які існують поза Інтернету є не інформаційними ресурсами .

3. Ресурсні Ідентифікатори

Ресурси ідентифікуються за допомогою *Uniform Resource Identifiers (Уніфіковані Ідентифікатори).*

У контексті LD, ми обмежемося використанням тільки HTTP URIs і не допускати інших URI схем, таких, як URNs і DOIs.

HTTP URIs хороші по двом причинам: вони забезпечують простий спосіб створення глобально унікальні імена без централізованого управління, а також URIs працюють не тільки як назва, але і як засіб доступу до інформації про ресурс через Інтернет.

4. Представлення

Інформаційні ресурси, можуть мати *представлення*. Представлення це потік байтів в певному форматі, наприклад, HTML, RDF / XML або JPEG. Наприклад, рахунок-фактура є інформаційним ресурсом. Він може бути представлений як HTML сторінка, а для друку PDF документом, або як RDF документом. Одни інформаційний ресурс може мати різні представлення, наприклад, у різних форматах, або на різних природних мовах.

5. Розіменування HTTP *URI*

Розіменування URI це процес пошуку URI в Інтернеті, щоб отримати інформацію про посилання на ресурс.

*Інформаційні ресурси: Коли ідентифікаційний URI інформаційного ресурсу є розіменованим, сервер власник URI, як правило, породжує нове уявлення, новий екземпляр інформаційного ресурсу в нинішньому стані відправляє його назад клієнту, використовуючи HTTP код відповіді 200 OK .

*Не інформаційні ресурси не можуть бути розіменовані напряму. Тому веб-архітектура використовується прийом, щоб URIs могли ідентифікувати неінформаційні ресурси, які будуть розіменовані: Замість того, щоб представити ресурс, сервер відправляє клієнту URI інформаційного ресурсу, який описує, не-інформаційних ресурс з використанням HTTP коду відповіді 303. Це називається 303 переадресацію. В якості другого кроку, клієнт розіменовує цей новий URI і отримує представлення ресурсу з описом, не-інформаційного ресурсу.

6. RDFa

В великому різноманітті технологій які пов’язані з Semantic Web важливо встановити співвідношення в якому перебувають LD до цих технологій.

Розглянемо технологію RDFa та її місце по відношенню до LD. RDFa призначений для створення семантичної розмітки контенту. Семантична розмітка або анотування являє собою явний опис семантики контенту ресурсу за допомогою понять семантичний моделі (онтології або словника). Таке явне опис семантики виконується зазначенням чіткого відповідності між певною частиною контенту ресурсу та його семантикою, описаною в семантичний моделі. Анотування при цьому базується на RDF. Сьогоднішні Web-ресурси розробляються здебільшого для використання людьми. Незважаючи на поступове поява в мережі даних, призначених для машинного сприйняття, ці дані в основному поширюються окремим файлом у певному форматі.

Притому відповідність машинної версії людському представленню досить обмежена. Як наслідок, Web-браузери можуть забезпечити користувачів лише мінімальної підтримкою в аналізі та обробці мережевих даних. Адже браузери тільки представляють інформацію. Технологія RDFa [**7**], [**8**], дозволяє супроводити графічні дані машиночитаємими підказками з допомогою набору XHTML-атрибутів. RDFa - це спосіб вираження RDF-даних в XHTML, в рамках якого дані, призначені для людини, що використовуються повторно.

Анотація це визначення семантики формальним способом. На даний момент в основу анотації покладають модель даних RDF.

Збагачення даних семантичними анотаціями процес при якому використовують спільно доступні словники термінів. Для отримання доступу до таких словників необхідні відповідні методи та технології.

Зв’язані дані дають бачення використання Інтернету для підключення відповідних даних, які раніше не були пов'язані між собою, або використовуючи Web знизити бар'єри для зв’язування даних, які в даний час пов'язані з використанням інших методів. Або більш конкретно, LD, це термін, який використовується для опису рекомендованих найкращих методів для виявлення, спільного використання та підключення частин даних, інформації та знань в Semantic Web, використовуючи URIs і RDF" [**9**]. В основу реалізації LD покладено модель даних RDF для опублікування даних в Інтернеті та використання RDF посилань для зв’язування ресурсів між собою.

7. Вибір словників для представлення інформації.

Для того, щоб здійснювати анотування, яке в подальшому може бути легко оброблене програмними додатками необхідно повторно використовувати терміни з відомих словників, де це можливо. Нові терміним повинні визначатися тільки тоді, якщо не знайдено необхідні терміни в існуючих словниках.

8. Повторне використання існуючих термінів

Необхідно перевірти чи можуть дані представлені з використанням термінів з поданих нижче словників, перш ніж визначити будь-які нові терміни:

- Friend-of-a-Friend (FOAF) , словниковий запас для опису людей.
- Dublin Core (DC) визначає загальні атрибути метаданих.
- Semantically-Interlinked Online Communities (SIOC) словник для представлення онлайнових співтовариств.
- Description of a Project (DOAP), словниковий запас для опису проектів.
- Simple Knowledge Organization System (SKOS), словник для представлення таксономії і слабко структурованих знань.
- Music Ontology забезпечує терміни для опису виконавців, альбомів та треків.
- Review Vocabulary, лексика для подання відгуків.

- Creative Commons (CC), словниковий запас для опису умови ліцензії.

Більш великий список відомих словників ведеться в [**10**].

Загальноприйнята практика змішувати терміни з різних словників. Особливо рекомендується використовувати rdfs:label та foaf:depiction властивостей, коли це можливо, оскільки ці терміни добре підтримується клієнтськими додатками.

Якщо потрібна URI посилання на географічні місця, напрямки досліджень, загальні теми, художники, книги і компакт-диски, необхідно використовувати Уніфіковані Ідентифікатори з джерел даних в рамках проекту W3C SWEO Linking Open Data [**11**], наприклад GeoNames, DBpedia, MusicBrainz, dbtune або RDF Book Mashup. Дві основних переваги використання Уніфіковані Ідентифікатори з цих джерел даних:

1. Уніфіковані Ідентифікатори роіменовуються, а це означає, що опис цієї концепції може бути отриманий з Інтернету. Наприклад, за допомогою URI DBpedia http://dbpedia.org/page/Doom можливо визначити широку інформацію про комп'ютерну гру Doom в тому числі опис на різних мовах і різних класифікацій.
2. Уніфіковані Ідентифікатори вже пов'язано з Уніфіковані Ідентифікатори з інших джерелами даних. Наприклад, можливо переходити від даних URI DBpedia http://dbpedia.org/resource/Berlin, до даних представлених на GeoNames і EuroStat.. Тому, використовуючи концепцію URI ці дані, з’єднуються з багатьма іншими даними утворюючи мережу зв’язаних даних.

Крім того сучасні алгоритми побудови взаємозв'язку в поточних наборах даних Linking Open Data в переважній більшості засновані на шаблонах. Це означає, що може бути створено величезна кількість взаємозв'язків, однак якість цих зв'язків з точки зору їх "семантичної сили" дещо обмежена. Добре відомо, що люди вміють асоціювати, пропонується дозволити людям робити деякі частини взаємозв'язків на основі асоціації.

9. Трансформація схем метаданих в рамках концепції LD.

В роботі [**12**] викладено методологію трансформації схем метаданих в формат RDF на прикладі схеми метаданих MARC21. Однак принципи які там викладені можуть бути перенесені на інші схеми метаданих.

Semantic Web, як веб пов'язаних даних з використанням URI, з доступністю по протоколу HTTP, дає можливість створення великих взаємопов'язаних наборів даних.

Типовий MARC21 запис буде містити назву видання, ім'я автора, і відомості про предметну класифікацію та інші бібліографічні дані. .

```
=LDR 00673nam a2200217 a 4504
=001 9cbbe7fc3a7346d99c281979d45b679c
=003 UK-BiTAL
=005 20050705133033.0
=008 990831s1999\\\\enk j\\\\\\000\||eng|d
```

```
=015 \\$aGB99Y5741$2bnb
=020 \\$a0747542155 :
=035 \\$a()0747542155
=040 \\$aStDuBDS$cStDuBDS$dUK-BiTAL
=082 04$a823.914$221
=100 1\$aRowling, J. K.
=245 00$aHarry Potter and the prisoner of
Azkaban /$cJ.K. Rowling.
=260 \\$aLondon :$bBloomsbury,$c1999.
=300 \\$a317p. ;$c21 cm.
=650 \0$aPotter, Harry (Fictitious character)
$vJuvenile fiction.
=650 \0$aWizards$vJuvenile fiction.
=655 \7$aChildren's stories.$2lcsh
```

Приклад запису MARC21.

Для відображення даних в форматі MARC21 до формату RDF в роботі [**13**] виконано базове і пряме представлення MARC в RDF. Нижче показано приклад такого відображення.

```
@base <http://example.com/a_marc_record> .
@prefix marc21: <http://example.com/marc21#> .
[]
marc21:LDR “00673nam a2200217 a 4504”;
marc21:001 “9cbbe7fc3a7346d99c281979d45b679c”;
marc21:003 “UK-BiTAL”;
marc21:005 “20050705133033.0”;
marc21:008 “990831s1999 enk j 000 ||eng|d”;
marc21:015 [
marc21:a “GB99Y5741”;
marc21:2 “bnb”
];
marc21:020 [
marc21:a “0747542155 :”
];
marc21:035 [
marc21:a “()0747542155”
];
marc21:040 [
marc21:a “StDuBDS”;
marc21:c “StDuBDS”;
marc21:d “UK-BiTAL”
];
marc21:082> [
marc21:ind1 “0”;
```

```
marc21:ind2 “4”;
marc21:a “823.914$221”
];
marc21:100 [
marc21:ind1 “1”;
marc21:a “Rowling, J. K.”
];
marc21:245> [
marc21:ind1 “0”;
marc21:ind2 “0”;
marc21:a “Harry Potter and the prisoner of Azkaban /”;
marc21:c “J.K. Rowling.”
];
marc21:260 [
marc21:a “London :”;
marc21:b “Bloomsbury,”;
marc21:c “1999.”
];
marc21:300 [
marc21:a “317p. ;”;
marc21:c “21 cm.”
];
marc21:650 [
marc21:ind2 “0”;
marc21:a “Potter, Harry (Fictitious character)”;
marc21:v “Juvenile fiction.”
], [
marc21:ind2 “0”;
marc21:a “Wizards”;
marc21:v “Juvenile fiction.”
];
marc21:655 [
marc21:ind2 “7”;
marc21:a “Children's stories.”;
marc21:2 “lcsh”
] .
```

Кореневим вузлом в цьому прикладі є самий запис. Ця модель є простою транслітерацією оригінального запису з використанням синтаксичної трансформації.

Основна проблема цього представлення є те, що семантика предиката зокрема залежить від даних суміжних даних, наприклад, останнє поле, 655, містить "дитячі казки.". Цей термін виводиться з класифікатору Бібліотеки Конгресу на основі на основі того що 2 є показником 7. Така складна

взаємодія показників для встановлення значення є притаманна стандарту MARC. Окрім того розробник також повинен орієнтуватися в кодах стандарту.

Більш зручним для читання є представлення без алфавітно-цифрових кодів. RDF дає зручну можливість узгодити відповідність кодів та їх значень. Після відображення кодів запис набуде наступного вигляду.

```
@base <http://example.com/a_marc_record> .
@prefix marc21: <http://example.com/marc21#> .
[]
marc21:controlNumber
“9cbbe7fc3a7346d99c281979d45b679c”;
#Following data comes from fixed positions in the
Leader
marc21:recordStatus “New”;
marc21:recordType “Language material”;
marc21:bibliographicLevel “Monograph/item”;
marc21:encodingLevel “Full”;
#Following data comes from fixed positions in 008
marc21:recordCreated
“1999-08-31”^^xsd:dateTime;
marc21:publicationStatus “Published”;
marc21:placeOfPublication “England”;
marc21:language “English”;
marc21:targetAudience “Juvenile”;
marc21:festschrift “No”;
#Following data comes from other control fields
marc21:controlNumberIdentifier “UK-BiTAL”;
marc21:recordUpdated
“2005-07-05T13:30:33Z”^^xsd:dateTime;
marc21:nationalBibliographyNumber [
marc21:number “GB99Y5741”;
marc21:sourceOfNumber “bnb”;
];
marc21:isbn “0747542155”;
marc21:deweyDecimalClassification “823.914”
marc21:associatedPersonalName “Rowling, J. K.”;
marc21:title “Harry Potter and the prisoner of
Azkaban”;
marc21:statementOfResponsibility “J.K.
Rowling.”;
marc21:placeOfPublication “London”;
marc21:dateOfPublication “1999”^^xsd:dateTime;
marc21:publisher “Bloomsbury”;
```

```
marc21:physicalExtent “317p.”;
marc21:physicalDimensions “21 cm”;
marc21:topicalTerm [
marc21:sourceOfTerm “LCSH”;
marc21:term “Potter, Harry (Fictitious
character)”;
marc21:formSubdivision “Juvenile fiction.”;
], [
marc21:sourceOfTerm “LCSH”;
marc21:term “Wizards”;
marc21:formSubdivision “Juvenile fiction.”
];
marc21:genre [
marc21:sourceOfTerm “LCSH”;
marc21:term “Children's stories.”;
] .
```

Для збагачення семантики слід звернути увагу на те, що кожний запис метаданих є відособленим і при класичному підході відсутнє повторне використання літералів запису. Для того щоб подолати цю проблему необхідно замінити літеральні значення на URIs представлення концептів. Причому літерали об’єднати в єдину онтологію. Такі онтології для MARC21 зроблено в роботі [**14**].

10. Семантичне представлення запису.

Оскільки множина записів метаданих зазвичай є великою тому доцільно застосовувати алгоритмічний підхід до побудови URI. Відповідно в розподіленій з паралельною обробкою системі слід враховувати, що у разі відсутності вже існуючого індефікатора його необхідно створювати та зареєструвати. Причому в момент реєстрації необхідно було б блокувати систему. Це призвело до ускладнення пошуку, централізованого керування та системи в загалому.

Окрім того слід забезпечувати механізми роботи з розподіленими даними та за умови забезпечення сторонніх агентів до цих даних. Зокрема в роботі [**12**] було виконано аналіз літералів на предмет встановлення еквівалентних відношень. Проте автори прийшли до висновку що при будь-якому алгоритмічному підході необхідне людське втручання, у випадку MARC21, наприклад, ім’я автора чи назва може відрізнятися в різних записах, хоча описуються одним і тим же URI.

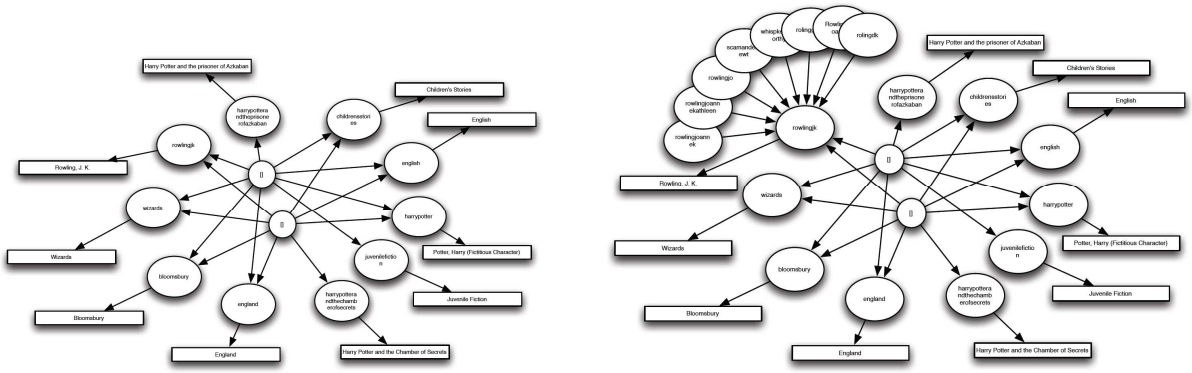
В такі предметній області як бібліотека вже давно існує вирішення такої проблеми завдяки Авторитетним записам.

Авторитетні записи містять різні форми запису імен авторів назв та предметних класифікаторів та інші дані. Такі дані легко перетворити до RDF. Такі авторитетні дані містять всі можливі варіанти написання прізвища та

ініціалів автора, назви або предметного класифікатора. Приклад таких даних показано на Мал. Нижче.

=100 1/$a Rowling, J. K.
=400 1/$a Rowling, Joanne K.$q (Joanne Kathleen)
=400 1/$a Rowling, Jo
=400 1/$a Scamander, Newt
=400 1/$a Whisp, Kennilworthy
=400 1/$a Roling, G’e. K e
=400 1/$a Rowlingova, Joanne K.
=400 1/$a Roling, Dz‿h‿. K.

В результаті було продемонстровано яким чином відрізняється використання авторизованих даних від даних які побудовано алгоритмічним способом.



а) б)

Приклади запису а) без використання авторитетних даних, б) з використанням них.

Проте навіть в такому випадку неможливо досягти точності в записах метаданих. Оскільки виникають випадки, наприклад, коли прізвище та ініціали авторів співпадають хоча на справді це різні автори які твори в різних галузях і навіть різних роках. Кращим випадком буде генерація узагальненого URI яке можна підв’язати до обох авторів (оскільки в них однакове ПІБ), водночас за необхідністю в подальшому уточнювати відомості про них. Не можливо створення більш точного посилання з менш інформативного, однак можливо створювати менш точне описове посилання коли ми маємо біль чітку інформацію.

11. Семантична аннотація

Для того щоб показати переваги семантичного підходу до електронної бібліотеки на відміну від класичної необхідно виділити ті структурні елементи які раніше не мали семантичної моделі, і до яких ми пропонуємо цю модель побудувати.

Основними двома компонентами електронної бібліотеки є її контент та набір програмного забезпечення для роботи з цим контентом. Для початку розглянемо контент ЕБ. Інформація в електронних бібліотеках описується в термінах електронні об’єкти (Digital objects - DO), які являють собою

мультимедійний контент і метадані [15]. Оскільки обсяги DO значні, то для спрощення пошуку та класифікації використовують анотування DO.

Формальна модель, яка запропонована в [16], виділяє два підходи до розуміння анотацій: анотації як метадані або анотації як контент.

У першому випадку ми маємо справу з різноманітними схемами метаданих (Dublin Core, MARC і ін.) які використовуються для опису інформаційних ресурсів. Ці анотації насамперед направлені на користувача.

У другому випадку анотації як представлення контенту призначені для автоматизованої машинної обробки. Ці анотації надають семантику документа. Семантична анотація - анотація написана формальною мовою з добре визначеною семантикою і заснована на онтологіях. Фактично ці анотації є формальною моделлю DO, з можливістю машинної обробки.

Семантика контенту в свою чергу, може визначатися на основі зовнішніх зв’язаних онтологій, що дозволить будувати семантичну модель документу, де зв'язок визначається між окремими сегментами DO, та на основі семантики зв’язків між структурними компонентами DO, де зв'язок визначається між логічними закінченими структурними компонентами Рис. 1.

Модель, яка представлятиме цифровий об'єкт повинна відображувати фактичний зміст даного об'єкту.

Розташування кожного цифрового об'єкту також як і анотації ідентифікується унікальним ідентифікатором – посиланням (link). Окрім цього посилання сполучає цифровий об'єкт і анотацію, і може відображувати відношення між об'єктами. Отже, можна виділити два типи посилань:

посилання анотації (Annotate link) - відображує відношення в середині цифрового об'єкту, який може бути як документом, так і анотацією;

посилання відношення (Relate-to link) - визначає відношення зовнішнього цифрового об'єкту до об'єкту, що анотується, це відношення конкретизується через онтологі, шляхом повторного використання термінів.

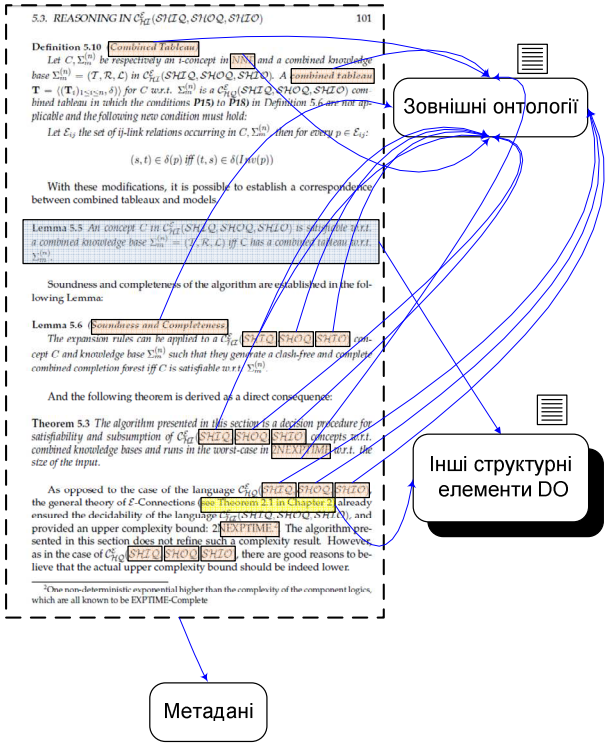


Рис. 1 Анотування документа шляхом визначення відношень між термінами, лемами та зовнішніми отологіями а також визначення зв'язків між логічно закінченими елементами з іншими частинами DO. Окрім цього DO описується метаданими.

Нехай LT множина типів посилань, тоді LT містить наступні типи посилань $LT = \{AnnotateLink, Relate-to\,Link\}$.

Посилання надаються у вигляді ідентифікаторів. Загальноприйнятими ідентифікаторами можуть виступати URI, DOI, OPENURL, Persistent URL (PURL), PURL-based Object Identifier. $H\left(k\right)$ множина ідентифікаторів цифрових об'єктів у момент часу k.

У цій моделі цифровий об'єкт надається у вигляді потоку. Потік sm це кінцева послідовність:

$$DO\left(\sum\right) \to sm : I = \left\{1, 2, ..., n\right\},\ n \in \mathbb{N}$$

де $\sum = (\varphi_1, \varphi_2, ..., \varphi_n)$ - алфавіт символів.

Якщо ми маємо потік sm:

$$DO\left(\sum\right) \to sm : I = \left\{1, 2, ..., n\right\},\ n \in \mathbb{N},$$

то в цьому потоці ми можемо виділити неперервний сегмент st_{sm} послідовності чисел a, b так що:

$$st_{sm} = [a, b],\ 1 \le a \le b \le n,\ n \in \mathbb{N}.$$

Безліч сегментів ми позначаємо ST, так що $\forall st_{sm_i} \in ST,\ i = 1, 2, ..., n$.

Якщо цифровий об'єкт DO має безліч унікальних ідентифікаторів H то функція hsm відображує унікальний ідентифікатор до сегментів які містяться в DO:

$$h \xrightarrow{hsm} st_{sm_i},\ n \in \mathbb{N}$$

Потік вимагає, щоб функція не мала властивостей сюрьїктивності і

інєктивності. Кожен цифровий об'єкт може мати принаймні один потік.

SM визначає множину потоків, так що $\forall sm_i \in SM$, $i = 1,2,...,n$.

Анотацію можна розглядати як процес розширення онтології O_i. Розглянемо найпростіший випадок, коли розширення відбувається шляхом додавання нових екземплярів онтології O_i. Кожний клас онтології O_i будемо позначати через kl_i, а множину класів через KL.

Анотація $a \in A(k)$ це кортеж:

$$a = \begin{pmatrix} h_a \in H\left(k\right), \\ A_\alpha \subseteq KL\left(k\right) \times LT \times ST\left(k\right) \times \\ \times SM\left(k-1\right) \times H\left(k-1\right) \end{pmatrix}$$

де h_a - унікальний власний ідентифікатор анотації a, тобто $h\left(h_a\right) = a$;

A_α множина n-арних відношень анотації a і визначається як добуток множин KL, LT, ST, SM та H.

У випадку анотування веб-документів, формальна модель зміниться. Нехай A множина всіх анотацій a, а D множина документів, відповідно, $DO = D \bigcup A$, причому підмножина множини DO позначатимемо do, тобто $do \in DO$.

Анотацією веб-документа називатимемо мічений граф:

$G := \left(\left(DO, E_{da} \subseteq A \times DO\right)\right)$,

де $DO = D \bigcup A$ вершини графа;

$E_{da} = \begin{Bmatrix} \left(a, do\right) \in A \times DO \mid \exists \alpha \in A_\alpha, \\ \alpha = \left(kl_i, sm_i, st_{sm_i}, hsm^{-1}, LT\right) \end{Bmatrix}$ - сторони графа

12. Практична реалізація.

Практичною реалізацію включення анотація до документів є використання мікроформатів. A microformat [**15**] являє собою веб підхід до семантичний розмітки, яке направлено на повторне використання існуючих XHTML і HTML-теги для передачі метаданих та інших атрибутів. Такий підхід дозволяє інформації, призначених для кінцевих користувачів (як, наприклад, контактну інформацію, географічні координати, календар подій, тощо), також буде автоматично обробленою програмним забезпеченням, таким чином одна і таж інформація повторно використовуються.

Незважаючи на те, що зміст веб-сторінок технічно можливо "автоматично обробити", таку обробки важко здійснювати тому, що традиційні теги розмітки використовуються для відображення інформації на сайтах, а не для опису інформації. Microformats призначені для подолання цього розриву шляхом приєднання семантики, і тим самим усунути інші, більш складні методи автоматизованої обробки, як, наприклад, обробки природної мови.

Поточні microformats дозволяють кодувати і видобувати структуровану інформацію про події, контактну інформацію, соціальних відносин і так далі. Сучасні браузери такі як Internet Explorer 8 підтримують розмітку microformats.

Одним з мікроформатів є RDFa. **RDFa** (або Описи Ресурсів - В - атрибути) представляє собою набір розширень для XHTML, і в даний час є рекомендацією W3C. RDFa використовує атрибути з XHTML в мета-елементах і елементах зв'язку. Це дозволяє додавати семантичну анотацію шляхом розмітки XHTML. Просте відображення визначається тим, що RDF трійки можуть бути вилучені.

The essence of RDFa is to provide a set of attributes that can be used to carry metadata in an XML language (hence the 'a' in RDFa). Суть RDFa полягає в тому, щоб представити набір атрибутів, які можуть бути використані для перенесення метаданих в XML (звідси й 'a' в RDFa).

Ці атрибути є:

about - URI або CURIE зазначенням ресурсів метаданих про поточний ресурс.

rel та rev - Визначення відносини або зворотних відносини з іншими ресурсами.

href , src та resource - зазначення партнеру ресурсу

property – визначає властивості контенту в елементі

content - необов'язковий атрибут перевизначає контент коли використовуються –атрибути властивостей

datatype - необов'язковий атрибут, задає ти даних тексту використовується з атрибута властивостей.

typeof - необов'язковий атрибут, задає RDF тип.

Приклад RDFa

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML+RDFa 1.0//EN"
   "http://www.w3.org/MarkUp/DTD/xhtml-rdfa-1.dtd">
<html xmlns="http://www.w3.org/1999/xhtml"
   xmlns:foaf="http://xmlns.com/foaf/0.1/"
   xmlns:dc="http://purl.org/dc/elements/1.1/"
   version="XHTML+RDFa 1.0" xml:lang="en">
 <head>
   <title>John's Home Page</title>
   <base href="http://example.org/john-d/" />
   <meta property="dc:creator" content="Jonathan Doe" />
 </head>
 <body>
   <h1>John's Home Page</h1>
   <p>My name is <span property="foaf:nick">John D</span> and I like
     <a href="http://www.neubauten.org/" rel="foaf:interest"
```

```
      xml:lang="de">Einstürzende Neubauten</a>.
    </p>
    <p>
     My <span rel="foaf:interest" resource="urn:ISBN:0752820907">favorite
     book</span> is the inspiring <span about="urn:ISBN:0752820907"><cite
     property="dc:title">Weaving the Web</cite> by
     <span property="dc:creator">Tim Berners-Lee</span></span>
    </p>
   </body>
  </html>
```

13. Список литературы

[1] Tim Berners-Lee. (2009, June) Design Issues. [Online]. http://www.w3.org/DesignIssues/LinkedData.html

[2] Tim O'Reilly. (2009, Mar.) Tom Heath's Displacement Activities. [Online]. http://tomheath.com/blog/2009/03/linked-data-web-of-data-semantic-web-wtf/

[3] Richard Cyganiak, Tom Heath Chris Bizer. (2007, July) Welcome to WWW4, the research application server of the Lehrstuhl für Wirtschaftsinformatik. [Online]. http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/

[4] Tim Bray, Dan Connolly, Paul Cotton, Roy Fielding, Mario Jeckle, Chris Lilley, Noah Mendelsohn, David Orchard, Norman Walsh, and Stuart Williams Tim Berners-Lee. (2004, Nov.) World Wide Web Consortium (W3C). [Online]. http://www.w3.org/TR/webarch/

[5] (2007, Oct.) World Wide Web Consortium (W3C). [Online]. http://www.w3.org/2001/tag/doc/httpRange-14/2007-05-31/HttpRange-14

[6] Roy T. Fielding. (2005, Jan.) [httpRange-14] Resolved. [Online]. http://lists.w3.org/Archives/Public/www-tag/2005Jun/0039.html

[7] Mark Birbeck, Shane McCarron, Steven Pemberton Ben Adida. (2008, Sep.) The World Wide Web Consortium (W3C). [Online]. http://www.w3.org/TR/rdfa-syntax/

[8] Mark Birbeck Ben Adida. (2009, Sep.) RDFa Primer. [Online]. http://www.w3.org/TR/xhtml-rdfa-primer/

[9] Linked Data community. (2009) Linked Data - Connect Distributed Data across the Web. [Online]. http://linkeddata.org/

[10] (2009) ESW Wiki. [Online]. http://esw.w3.org/topic/TaskForces/CommunityProjects/LinkingOpenData/CommonVocabularies

[11] (2009) ESW Wiki. [Online]. http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

[12] Danny Ayers, Nadeem Shabir Rob Styles, "SEMANTIC MARC, MARC21 AND THE SEMANTIC WEB," in *Linked Data on the Web (LDOW2008)*, Beijing, 2008.

[13] Ian Davis. (2005) MARC Transliteration. [Online]. http://iandavis.com/blog/2005/12/marc-transliteration

[14] Sebastian Ryszard Kruk Marcin Synak, "ESWC2005," in *MarcOnt Initiative the Ontology for the Librarian World*, 2005, 2005.

[15] Microformats. [Online]. http://microformats.org