УДК 681.3

РОЗРОБКА ОНТОЛОГІЧНИХ ТЕРМІНОСИСТЕМ ІНФОРМАЦІЙНИХ РЕСУРСІВ ІНТЕРНЕТ ТА ЇХ КОГНІТИВНИХ МОДЕЛЕЙ У НАУКОВИХ ДОСЛІДЖЕННЯХ

Ю.В. Рогушина, А.Я. Гладун, В.Н. Штонда

Інститут програмних систем НАН України, Київ-187, МСП, 03680, проспект Академіка Глушкова, 40,
email: _jjj_@ukr.net
Міжнародний науково-навчальний центр інформаційних технологій та систем НАН України і МОН України,
Київ-187, МСП, 03680, проспект Академіка Глушкова, 40, email: glanat@yahoo.com
Видавництво «Діалектика» – генеральний директор

Сьогодні багато інтелектуальних Веб-застосунків потребують використання розподілених баз знань, формалізованих у вигляді онтологій та тезаурусів. Тому важливо створити не тільки методики, певні застосовні евристичні правила, за якими потрібно створювати описи понять в онтологіях та тезаурусах. Це дозволить підвищити якість онтологій, що створюються, та забезпечити їх більшу інтероперабельність. Щоб коректно визначати відношення між різними термінами тезаурусу, пропонується використовувати основи мереологічного підходу для більш чіткого вибору відношень типу “частина-ціле”.

Now a great necessity in means and methods for creation of ontology and thesauri – interoperable models of domain knowledge representation that is used in distributed Web-applications – is exists. For correct definition of relations between the different thesaurus terms we propose to use some elements of ontological analysis and basic foundations of mereology. In addition some applied rules of thesaurus terms creation are proposed to build the knowledge with higher quality.

Ефективність науково-дослідних робіт безпосередньо залежить від якості їхнього інформаційного забезпечення, а саме пошук інформації є ключовим етапом будь-якого наукового дослідження. На сьогоднішній день глобальна мережа Інтернет – найважливіше джерело інформації для всіх областей знань, однак пошук спеціалізованої науково-технічної інформації виявляється іноді малоефективним.

Сучасні напрями розвитку інформаційних технологій (ІТ) пов'язані зі створенням інформаційних систем, що базуються на знаннях. У сучасних дослідженнях у сфері розподіленого керування знаннями застосовують термін *онтологія* для явної опису системи знань певної галузі або інформаційного ресурсу. Онтології забезпечують загальний словник певної сфери діяльності та визначають (з різними рівнями формалізації) значення термінів і відношення між ними. У найбільш загальному випадку вона являє собою угоду про спільне використання понять, що містить засоби представлення предметних знань і домовленості про методи розуміння.

Лексикографічний, онтологічний й тезаурусний опис контенту інформаційних ресурсів належать до методів формалізації знань суб'єкта про предметну область із використанням формалізованих знакових систем. Тезаурус покликаний відобразити семантичні відношення й зв'язки між термінами, що використовуються в тексті. Тезаурус можна розглядати як модель логіко-семантичної структури термінології, а також як модель структури відповідної науки. Тезаурус є окремим випадком онтології, що дозволяє представляити поняття так, що вони стають придатними для машинної обробки.

Незалежно від виду онтології, необхідно включити словник термінів і деякі специфікації їхніх значень, що дозволяє обмежити інтерпретацію цих термінів і відбити їхню взаємодію. За такого підходу поняття онтології перетинається з уже давно прийнятим в інформатиці і лінгвістиці поняттям *тезауруса*. Онтологія – це база знань, що описує факти, які передбачаються завжди істинними в рамках певного співтовариства на основі загальноприйнятого значення тезауруса.

Онтологія може використовуватися як посередник: між користувачем та інформаційною системою або між членами співтовариства, наприклад, між користувачами деякого корпоративного сховища даних.

Наявність онтології певного ІР дозволяє автоматизувати обробку семантики такого ІР (наприклад, шукати в Інтернеті саме ті ІР, що допомагають користувачеві розв’язати певну задачу). Побудова онтології є складною задачею, яка не може виконуватися повністю автоматично, але, використовуючи певні правила та технологічні прийоми, можна полегшити та пришвидшити цей процес.

Термінологічний словник-тезаурус відрізняється від інших словників тим, що він представляє когнітивну модель певної галузі знання або людської діяльності за допомогою декількох входів у словник: 1) від "концепту до концепту", тобто від одного поняття до іншого з поданням ієрархічних (вертикальних) і кореляційних (горизонтальних) зв'язків; 2) від "концепту до знака" через ідеографічну частину тезауруса з поданням понятєво-лексичних зв'язків; 3) від "знака до концепту" за допомогою алфавітного покажчика адреси слів у понятійних полях дескрипторів з поданням лексико-понятійних зв'язків; 4) "від знака до знака" за допомогою пермутаційного покажчика, що використається для пошуку словосполучень по одному зі складових компонентів. Таким чином, структура тезаурус задає лексичну систему підмови в чотирьох вимірах: "концепт-знак", "концепт-концепт", "знак-концепт", "знак-знак".

Когнітивна модель – термін, використовуваний як у лінгвістиці (концепція "мова є різновид когнітивного процесу"), так і в психології, де описуються механізми мислення й утворення концептуальної системи людської свідомості як тієї бази, на якій мислення протікає. В інформаційних системах він застосовується для опису структури знань, оброблюваних такою системою.

Онтологічний підхід до подання знань предметної області

Онтологія – угода про спільне використання понять (термінів), що містить засоби подання предметних знань і домовленості про методи логічного висновку (міркувань) [4]. Це формалізований опис погляду на світ у конкретній сфері інтересів, що складається з набору термінів і правил використання цих термінів, що обмежують їхнє значення в рамках конкретної галузі [5]. Знання в онтологіях формалізують, використовуючи п'ять видів компонентів: класи, відношення, функції, аксіоми та екземпляри [2]. Формальна модель онтології – упорядкована трійка O=<T, R, F>, де T – скінченна множина термінів ПрО, що описує онтологія O; R – скінченна множина відношень між термінами заданої ПрО; F – скінченна множина функцій інтерпретації, заданих на термінах і / чи відношеннях онтології O.

При створенні онтологій найбільшу складність становить формування множини F, тому що цей процес вимагає застосування спеціальних навичок з галузі інженерії знань і формальної логіки. У той же час стосовно трудомісткості основна робота з формування онтологій припадає на формування множини X, але слід зазначити, що ця робота доступна більшості фахівців довільної галузі. Складніше визначити множину відношень R, які треба використовувати для моделювання знань. У роботі [3] виділені найбільш загальні онтологічні відношення в реальних доменах – зв'язки еквівалентності, таксономічний, структурний, залежності, топологічний, причинно-наслідковий, функціональний, хронологічний, подоби, умовний і цільовий. Таксономія – це окремий випадок онтології, в якій присутні тільки ієрархічні зв’язки одного типу.

Одним із найпоширеніших відношень в онтології є відношення йменування. Воно є фундаментальним, тобто на його базі може бути побудована формальна система, що дозволяє виражати основні математичні поняття. Існує чотири фундаментальних відношення: приналежності (теорія множин ZF і NF), між функцією, її аргументом і результатом (теорія множин фон Неймана), йменування (онтологія Лесьневського) та "частина-ціле" (мереологія) [8].

Мереологія – це формальна теорія про частини і зв'язані з ними поняття, розроблена С. Лесьневським [9]. Об’єктом мереології є дослідження відношення "частина-ціле". Мереологія – частина тріади дедуктивних теорій, що включає також прототетику й онтологію (у Лесьневского онтологія розглядається тільки як система з єдиним відношенням «є» – “*is_a*”). Відношення "*частина-ціле*" є винятково важливим тому, що воно утворює основу поняття *системи,* яке є центральним у сучасному науковому пізнанні. Система являє собою структурне з'єднання своїх елементів. Її базовою формальною характеристикою є те, що елементи не просто входять у систему, а входять у неї внаслідок взаємодії з іншими елементами.

Інші аксіоми мереології описують взаємозв'язки між системою і її елементами. Приклад – аксіома системи і частин: якщо елемент зв'язаний в один бік, то він зв'язаний і в інший. Мереологія виходить за межі вивчення часткових відносин між елементами спільних систем. Вона також займається тими об'єктами, частини яких релевантні цілому. Такі об'єкти ідентифікуються як екземпляри. Серед мереологічних відношень можна виділити сім різних класів, і взагалі, транзитивність не прийнята серед екземплярів різних класів:

1) компонент-об'єкт: сторінка-книга;

2) член-колекція (наприклад, дерево-ліс);

3) частина-маса (наприклад, шматок-хліб);

4) матеріал-об'єкт (наприклад, алюміній-літак);

5) властивість-діяльність (наприклад, бачити-читати);

6) стадія-процес (наприклад, заварювання-готування чаю);

7) місцевість-область (наприклад, Закарпаття-Україна).

Більшість досліджень відношення "частина-ціле" присвячені вивченню частин, але можна також ідентифікувати різні типи цілого відповідно до таких властивостей: 1) чи віддільна частина від цілого (мелодія-пісня або вагон-потяг); 2) чи є частини просторовими або часовими (кімната-квартира або зима-рік); 3) чи відіграє частина певну функціональну роль стосовно цілого (двигун-автомобіль); 4) чи є частини неподільними (атом-молекула).

Знання цих теоретичних принципів допомагає більш точно визначити мереологічні відношення, що вводяться до онтології. Визначивши тип відношення за такою класифіацією, можна більш чітко визначити, чи можна використвувати для визнаачення зв’язків між поняттями одне або різні відношення.

Постановка задачі

Сьогодні багато інтелектуальних Веб-застосунків потребують використання розподілених баз знань, формалізованих у вигляді онтологій та тезаурусів. Тому важливо створити не тільки методики, певні застосовні евристичні правила, за якими потрібно створювати описи понять в онтологіях та тезаурусах. Це дозволить підвищити якість онтологій, що створюються, та забезпечити їх більшу інтероперабельність. Щоб коректно визначати відношення між різними термінами тезаурусу, пропонується використовувати основи мереологічного підходу для більш чіткого вибору відношень типу “частина-ціле”.

Формування тезауруса галузі, що цікавить користувача

Окремим випадком онтології, який простіше формувати та обробляти, є *тезаурус* – повний систематизований набір даних про будь-яку галузь знань, що дає змогу людині чи комп'ютеру в ній орієнтуватися. Можна досліджувати як тезауруси окремих фахівців, так і тезауруси галузей знань. Формальна модель тезауруса

$$Ts = \langle T, R \rangle,$$

де T – скінчена множина термінів; R – скінчена множина відношень між цими термінами; тезаурус можна розглядати як семантичну мережу, у вузлах якої знаходяться терміни, пов'язані відношеннями з обмеженого набору R. Основні технологічні фази створення тезауруса [10]:

- виділення лексичних одиниць, тобто формування словника (глосарія) T;
- розробка набору семантичних зв'язків;
- актуалізація зв'язків – установлення зв'язків між термінами.

При цьому дуже важливо сформулювати принципи, за якими буде здійснюватися кожна процедура. Для першого пункту визначальними є два аспекти – джерело лексичних одиниць та критерій їх добування. При розробці набору семантичних відношень можна знаходити їх у тексті, що описує дану галузь (намагатися вичленувати й уніфікувати ті відношення, що існують в текстах між термінами) або безпосередньо аналізувати знання. На практиці звичайно використовують поєднання обох методик. Для актуалізації семантичних зв'язків між термінами тезауруса можна використовувати знання експертів, а також документи, призначені як для фіксації структури знань (словники, класифікатори тощо), так і самі знання, що відображають ПрО (реферати, статті, монографії тощо).

Термін – це слово або словесний комплекс, що співвідноситься з поняттям визначеної організованої галузі пізнання (науки, техніки) та вступає у системні відношення з іншими словами, словесними комплексами й утворює разом з ними в будь-якому окремому випадку чи у певний час замкнуту систему, що відрізняється високою інформативністю, однозначністю, точністю й експресивною нейтральністю.

Для створення тезауруса можна скористатися методологією розробки онтологічних моделей – стандарт IDEF5 [11] сімейства IDEF, згідно з якою побудова тезауруса ПрО складається з п'яти основних дій:

- *вивчення і систематизація початкових умов* – мети і контексту розробки тезауруса, визначення меж ПрО, яка цікавить користувача;
- *збирання і накопичення даних* – підбір ІР, що належать до цієї ПрО;
- *аналіз даних* – вивчення відібраних ІР, формування словника термінів ПрО, що містяться у відібраних ІР;
- *початкова розробка тезауруса* – встановлення зв'язків між термінами ПрО [12], з якої потім витягуються базові терміни ПрО;
- *уточнення та затвердження тезауруса* – аналіз користувачем отриманого тезауруса та його коректування.

Під час формування тезауруса доцільно враховувати наступні рекомендації, які стосуються побудови визначень даних і метаданих та враховують вимоги, розроблені підкомітетом зі стандартизації ПК-6 „Телекомунікації та обмін інформацією між системами" з урахуванням ISO/IEC 11179.

Практичні рекомендації стосовно визначення термінів тезаурусу

Визначення термінів тезауруса може здійснюватися в автоматичному режимі (шляхом аналізу повнотекстових документів та інших інформаційних джерел), шляхом вилучення з інших баз знань (тезаурусів, онтологій тощо) або надаватися безпосередньо експертом ПрО [13].

Формальні вимоги до визначень термінів тезауруса:

1. Визначення має бути **викладене в однині.** Виняток становлять поняття, які самі є множинними. Наприклад, *"номер статті":* добре визначення – "номер посилання, що ідентифікує статтю"; погане визначення – "номер посилання для ідентифікації статей". У поганому визначенні використовується слово "статті", що може бути формою множини і це можна зрозуміти так, ніби один номер може посилатися на кілька статей.
2. Визначення повинне пояснювати**, чим є** наведене поняття**, а не тільки чим воно не є.** Наприклад, "розмір вартості фрахтування": добре визначення – "розмір витрат, які несе вантажовідправник для переміщення товарів з одного місця до іншого"; погане визначення –"розмір витрат, що не належать до витрат на пакування, документальне оформлення, завантаження, розвантаження та страхування". У поганому прикладі не вказано, що входить до поняття елемента даних.
3. Визначення повинне **мати вигляд описової фрази або речення.** Речення необхідне для формування точного визначення, яке містить важливі характеристики поняття. Просте наведення одного або кількох синонімів не є достатнім. Наприклад, "ім'я агента": добре визначення – "назва сторони, яка уповноважена діяти від імені іншої сторони"; погане визначення – "представник ". "Представник" є синонімом імені елемента даних, який не може бути адекватним визначенням.
4. Визначення повинне **містити лише широко відомі скорочення.** Розуміння значення скорочення, зокрема абревіатур та ініціалів, зазвичай обмежується певним середовищем. В іншому середовищі ті ж самі скорочення можуть викликати неправильне розуміння або непорозуміння. Таким чином, для запобігання неоднозначності, у визначеннях використовуються тільки повні слова без скорочень. Наприклад, "прилад для

вимірювання щільності": добре визначення – "прилад, який використовується для вимірювання концентрації рідини, в одиницях виміру маси до одиниці об'єму (м.д.о.) (тобто фунтів на кубічний фут; кілограмів на кубічний метр)"; погане визначення "прилад, який використовується для вимірювання концентрації рідини в термінах м.д.о. (тобто фунтів на кубічний фут; кілограмів на кубічний метр)". Проте м.д.о не є загальновідомим скороченням і його значення може бути незрозумілим для деяких користувачів. Скорочення має бути наведене повними словами.

5. Визначення має **бути викладене без використання визначень інших даних або базових понять.** Визначення термінів має наводитись у відповідному глосарії. Якщо потрібне інше визначення, воно має додаватись як примітка після тексту первинного визначення або як окремий запис у словнику. Пов'язані визначення можна отримати за допомогою атрибутів посилання (перехресних посилань). Наприклад: "код типу зразка": добре визначення – "код, який ідентифікує тип зразка"; погане визначення – "код, який ідентифікує тип обраного зразка. Зразок – це мала частка, вилучена для проведення експериментів. Він може бути як єдиним зразком для тестування, так і сурогатним зразком для контролю якості. Зразок для контролю якості – це сурогатний зразок, обраний для перевірки результатів тестування єдиних зразків". Погане визначення містить два додаткових визначення –"зразка" та "зразка для контролю якості".

Семантичні вимоги до визначень термінів тезауруса

1. Визначення має **відображати суттєвий зміст поняття.** Усі первинні характеристики поняття, мають бути відображені у визначенні з відповідним рівнем специфічності залежно від контексту. При цьому необхідно запобігати пояснення неважливих параметрів. Рівень деталізації залежить від потреб користувача системи та середовища.

Наприклад, "номер послідовності завантаження вантажу" (визначений контекст: будь-яка форма транспортування): добре визначення – «номер, що вказує на послідовність, в якій здійснюється завантаження до транспортного засобу або елемента транспортного середовища»; погане визначення – "номер, який відображає послідовність, в якій здійснюється завантаження до вантажівки" (у визначеному контексті вантажі можуть транспортуватись різними транспортними засобами:, вантажівками, кораблями, вантажними потягами і не обмежене лише вантажівками).

Інший приклад: "сума за рахунком-фактурою": добре визначення – "загальна сума, яку потрібно сплатити за рахунком-фактурою"; погане визначення – "загальна сума вартості всіх елементів, зазначених в рахунку-фактурі, включаючи усі відрахування, зокрема знижки та дисконти, додаткові платежі, зокрема страхові, транспортні та накладні витрати тощо". У поганому визначенні міститься зайва інформація.

2. Визначення має бути **точним та однозначним,** достатньо зрозумілим, щоб забезпечити його однозначну інтерпретацію. Наприклад, "дата отримання вантажу": добре визначення – «дата, на яку вантаж передається отримувачу»; погане визначення – «дата, на яку здійснюється доставка вантажу». У поганому визначенні не роз'яснюється, що таке "доставка". Під "доставкою" можна зрозуміти як момент розвантаження товару У певному місті, так і факт передачі товару кінцевому отримувачу. Не виключено, що кінцевий отримувач ніколи не отримає вантаж або його передача може здійснитися через кілька днів після розвантаження.

3. Визначення має **бути коротким.** Слід запобігати використання додаткових фраз описового характеру, подібних до "для забезпечення використання цього реєстру метаданих", "терміни, що мають бути описані". Наприклад, "ім'я набору символів": – добре визначення "ім'я, що присвоюється набору фонетичних або ідеографічних символів, в які зашифровані дані"; погане визначення – "ім'я, що присвоюється набору фонетичних або ідеографічних символів, в яких зашифровані дані для забезпечення використання цього реєстру метаданих або, якщо говорити про загальний вжиток, спроможність системного обладнання і програмного забезпечення обробляти дані, зашифровані одним або декількома шифрами". У поганому визначенні всі фрази після виразу "… в яких зашифровані дані" є зайвими.

4. Визначення повинне **мати можливість використовуватися окремо.** Зміст поняття має бути наочним у визначенні. Для розуміння поняття не потрібні додаткові роз'яснення. Наприклад, "назва міста розміщення школи": добре визначення – "назва міста, де знаходиться школа"; погане визначення – "див. сайт школи". Погане визначення не є самостійним, оскільки необхідно звернутися до додаткового джерела.

5. Визначення повинне **бути поданим без використання пояснювальної інформації, функціонального використання або процедурної інформації.** Пояснення не слід включати до визначень, тому що вони містять зайву інформацію. У разі потреби такі пояснення можуть бути розміщені в інших атрибутах метаданих. Припустимо додати кілька прикладів після визначення. Наприклад, "мітка поля даних": добре визначення – "ідентифікація поля в індексі, тезаурусі, базі даних тощо"; погане визначення – "ідентифікація поля в індексі, тезаурусі, базі даних тощо, яка застосовується для таких елементів інформації як примітки, колонки в таблицях". У поганому визначенні містяться примітки, що стосуються функціонального використання. Якщо інформація, що починається зі слів "яка застосовується…" є необхідною, то вона має бути розміщена в іншому атрибуті.

6. Визначення повинне **запобігати циклічних посилань.** Два поняття не слід розкривати одне через одне. Визначення одного поняття не може використовувати інше поняття як своє визначення, тому що це може призвести до ситуації, коли поняття визначається через інше поняття, яке, у свою чергу, визначається через перше поняття. Наприклад, два елементи даних з поганими визначеннями – "ідентифікаційний номер працівника – номер, що призначається працівнику; "працівник – людина, яка має відповідний ідентифікаційний номер працівника". Визначення посилаються одне на одне, але в жодному з них не наведено зміст поняття.

7. Визначення повинні **використовувати однакову термінологію та логічну структуру для пов'язаних визначень.** Для близьких або пов'язаних визначень має використовуватись одна й та ж сама термінологія та синтаксис. Наприклад, "дата відправлення товарів – дата, в яку товари були відправлені даній стороні", "дата отримання товарів – дата, в яку товари були отримані даною стороною". Використання єдиної термінології значно спрощує розуміння.

Використання онтологічних терміносистем електронних наукових інформаційних ресурсів для підтримки наукових досліджень

Сьогодні у наукових та освітніх установах активно ведуться роботи зі створення й розвитку електронних інформаційних ресурсів з використанням Інтернет-технологій для підтримки процесів проведення наукових досліджень й утворення. Одним з перспективних підходів до формування електронних наукових інформаційних ресурсів є створення й супровід їх метаописаний на семантичному рівні. При створенні таких метаописань виникає ряд важливих завдань. Необхідно розробити структуру інформаційного опису інформаційних об'єктів, характерних для даної предметної області, а також визначити джерела знань для формування відповідних електронних документів.

Щоб описати семантику інформаційних ресурсів, треба посилатися на відповідні онтології й тезауруси, забезпечити їхню якість та актуальність.

Висновки

Сьогодні для опису певної предметної галузі та створення інтелектуальних систем і мереж велике значення має розробка нових алгоритмів та методик формування тезаурусів та онтологій. Для адекватного відображення інтероперабельних моделей подання знань у роботі запропоновано методику, яка використовує мереологічний підхід для створення визначень термінів тезаурусу, що дозволило сформувати більш якісні знання.

1. *Gruber T. R.* A translation approach to portable ontology specifications // Knowledge Acquisition. – 1993. – V. 5. – P. 199–220.
2. *Гладун А.Я., Рогушина Ю.В.* Онтологический подход к поиску веб-сервисов в распределенной среде Интернета // Информатика. – Минск, 2006. – № 4. – С.116–127.
3. *Gómez-Pérez A., Moreno A., Pazos J., Sierra-Alonso A.*. Knowledge Maps: An essential technique for conceptualisation // Data & Knowledge Engineering. – 2000. – V.33(2). – P. 169–190.
4. *Гладун А.Я., Рогушина Ю.В.* Онтологии и мультилингвистические тезаурусы как основа семантического поиска информационных ресурсов Интернет // The Proc. of XII-th Intern. Conf. KDS'2006, Varna, Bulgaria. - P.115-121.
5. *Нариньяни А.С.* Кентавр по имени ТЕОН: Тезаурус + Онтология. – http://www.artint.ru/articles/narin/teon.htm.
6. *Musen M.* Domain Ontologies in Software Engineering: Use of Protege with the EON Architecture // Methods of Inform. in Medicine, 1998. – P. 540–550.
7. *Андреев А.М., Березкин Д.В., Симаков К.В.* Особенности проектирования модели и онтологии предметной области для поиска противоречий в правовых электронных библиотеках. – http://www.inteltec.ru/publish/articles/textan/ RCDL2004.shtml
8. *Непейвода Н.Н.* Мереология. – http://www.logic.ru/Russian/events/ifras/nnn.pdf
9. *Лесьневский С.* Об основаниях математики. – http://www.philosophy.ru/library/logic/lesnewxi.html.
10. *Рогушина Ю.В., Гладун А.Я.* Мереологические аспекты онтологического анализа интеллектуальных Web-сервисов // Зб.праць VII Міжнар. конф. „Інтелектуальний аналіз інформації" IAI-2007. – Київ, 2007.– С. 312–321.
11. IDEF5 - Ontology Description Capture Method - (www.idef.com/IDEF5.html)
12. *Браславский П. И., Гольдштейн С. Л., Ткаченко Т. Я.* Тезаурус как средство описания систем знаний // Информационные процессы и системы. – 1997. – № 11 (Серия 2). – С. 16–22.
13. *Гладун А.Я., Рогушина Ю.В.* Основи методології формування тезаурусів з використанням онтологічного та мереологічного аналізу // Штучний інтелект. –2008. –№ 4. – С. 53–61.