УДК 004.415

В.А Резніченко, О.В. Новицкий, Г.Ю. Проскудіна

ІНТЕГРАЦІЯ НАУКОВИХ ЕЛЕКТРОННИХ БІБЛІОТЕК НА ОСНОВІ ПРОТОКОЛУ OAI-PMH

Розглядається підхід до інтеграції електронних бібліотек на основі технології Ініціативи відкритих архівів, а саме за допомогою протоколу OAI-PMH. Описано досвід побудови інтегрованих бібліотечних систем, а також інструменти для їх реалізації.

Вступ

Університети та дослідницькі інститути всього світу активно планують та реалізують репозиторії (архіви, електронні бібліотеки) своєї наукової продукції. Крім того, веб-механізм, поширення швидкісних мереж надають нові можливості для своєчасного поширення наукової інформації. У процесі роботи з електронними ресурсами кожна організація стикається з цілим рядом проблем, щодо одержання доступу до інформаційних ресурсів інших організацій, так і виробництва власної електронної інформації та її поширення. При цьому все більш актуальними стають питання інтеграції інформаційних ресурсів. За останні роки створення систем інтеграції даних стало важливим напрямком практичних розробок інформаційних систем різного призначення, в тому числі й електронних бібліотек (ЕБ) або архівів.

Під інтеграцією даних в електронних системах ми розуміємо забезпечення єдиного уніфікованого інтерфейсу для доступу користувачів до сукупності автономних джерел, які як правило, мають неоднорідність щодо деяких їх властивостей [1]. Своєрідний клас систем інтеграції представляють системи, в яких за основу прийнято технологію Ініціативи відкритих архівів (Open Archive Initiative – OAI) [2]. У більшості відомих систем цієї категорії їх інформаційні ресурси представляють собою колекції текстових документів, передусім наукових публікацій, які автономно формуються у вузлах глобальної мережі, підтримуються та адмініструються їх власниками.

Згідно з технологією OAI, передбачається матеріалізована інтеграція у єдиному репозиторії не самих інформаційних ресурсів, що цікавлять користувачів системи інтеграції, а представлених деяким стандартним чином метаданих, що описують колекції інформаційних ресурсів джерел даного архіву і окремі елементи цих колекцій. Збір таких метаданих для репозиторія здійснюється згідно зі спеціально розробленим протоколом Open Archives Initiative – Protocol for Metadata Harvesting (OAI-PMH) [3], що забезпечує глобальні послуги доступу та пошуку.

Робота містить більш докладний опис такого підходу, а також двох програмних системи для його реалізації, що виконують роль провайдера даних (EPrints) та сервіс провайдера (PKP Open Archives Harvester).

Підхід продемонстровано на прикладі реалізації електронної бібліотеки Житомирського державного університету ім. Івана Франка з використанням програмного забезпечення (ПЗ) EPrints. Даний проект був зареєстрований в якості провайдера даних в OAI та був вибраний сервіс провайдером The University of Illinois OAI-PMH Data Provider Registry (http://gita.grainger.uiuc.edu/registry/searchform.asp), також подано заявку на реєстрацію у OAIster (http://oaister.umdl.umich.edu). Крім того був створений власний сервіс провайдер для збору метаданих з цього архіву на основі ПЗ PKP Open Archives Harvester.

Ця робота містить: огляд шляхів вирішення задачі інтеграції (розділ 1); розгляд ініціативи “Вікриті архиві” – як шлях до створення середовища з високою сту-

пінню інтероперабельності (розділ 2); опис інструменьтальних засобів, що реалізують запропонований підхід та приклади реалізації (розділ 3) та висновки.

1. Шляхи вирішення задачі інтеграції

Існує кілька підходів до вирішення проблеми створення електронних бібліотек з інтегрованими інформаційними ресурсами. Серед них можна виділити наступні два класи таких систем: з інтегрованим веденням ресурсів та з розподіленим веденням ресурсів.

Підхід з інтегрованим веденням ресурсів передбачає збір, збереження й обробку інформаційних ресурсів у єдиному репозиторії. Такий підхід видається доцільним у випадку, коли інформаційні ресурси організаційно породжуються в одному місці і безпосередньо належать одному власнику. Прикладом таких організаційних структур можуть бути науково-дослідні інститути НАН України. Будь-який інститут може створити інтегровану наукову електронну бібліотеку (НЕБ), яка б містила колекції її різноманітних наукових ресурсів (наукові статті, звіти, дисертації, учбово-методичні посібники, матеріали конференцій) і вирішувала всі проблеми з веденням таких ресурсів та наданням доступу до них. Однак, наприклад, для створення НЕБ НАН України, яка б припускала централізацію всіх ресурсів, такий підхід не зовсім підходить, оскільки він потребує вирішення цілого ряду складних організаційно-технічних задач і, насамперед, тих, що спрямовані на збір вихідних інформаційних ресурсів з інститутів. Крім того, він потребує створення розвинутої структури ведення такої НЕБ.

Підхід з розподіленим веденням ресурсів припускає, що існує багато організацій, які здійснюють самостійне створення і ведення електронних бібліотек і надають можливість доступу до цих ресурсів, включаючи також і організацію пошуку необхідних ресурсів. Крім того, існує "надбудова" над ними, що дозволяє робити пошук за цими ресурсами і, при наявності відповідних умов, надавати доступ до самих ресурсів. У цьому випадку інститути НАН України створюють свої НЕБ, а інтегрована служба, що діє на рівні НАН України, забезпечує пошук і видачу відповідних ресурсів.

На сьогодні існує два концептуальних рішення цього підходу. Перше припускає існування механізму *перехресного пошуку* за багатьма архівами (НЕБ), коли всі ресурси, бібліографічні описи та пошуковий сервіс знаходиться в організації. Так працюють системи з використанням протоколу Z39.50 [4]. При цьому пошук здійснюється шляхом безпосереднього звернення до всіх або до вибраних користувачем електронних бібліотек з наступним зведенням одержаних результатів у єдиний список. Друге – пропонує здійснювати *збір (харвестинг, harvesting) метаданих*, що описують інформаційні ресурси "на місцях" для того, щоб можна було надати централізований пошук в одному місці на основі зібраних метаданих. За суттю це є деякий аналог інтегрованого електронного каталогу. Далі в роботі розглядається другий підхід.

2. Ініціатива відкритих архівів

Ініціатива відкритих архівів (OAI – Open Archive Initiativee) виникла в зв'язку з тим, що багато організацій, де створюються електронні інформаційні ресурси (e-print-співтовариство), і, насамперед, наукові ресурси, вирішили надати відкритий доступ до них. У зв’язку з цим виникла проблема надання інтегрованого доступу до неоднорідних гетерогенних репозиторіїв. OAI націлена саме на те, щоб розробляти та сприяти розвитку й поширенню середовища і відповідних стандартів, які б дозволили об'єднати зусилля e-print-співтовариства з інтегрованого доступу до їхнших ресурсів [2].

Суть підходу відкритих архівів, полягає у тому, щоб дозволити здійснювати веб-доступ до інформаційних ресурсів, розташованих у інтероперабельних репозиторіях, за допомогою організації спільного використання, публікації й архівування метаданих таких ресурсів.

2.1. Протокол OAI для збору метаданих (OAI-PMH). Протокол OAI для збору (хагвестінгу) даних (OAI-PMH) визначає механізм збору записів, що містять

метадані з репозиторіїв. Протокол OAI-PMH надає провайдерам даних простий спосіб такого представлення їх метаданих, який робить їх доступними для провайдерів сервісів. При цьому для обміну метаданих використовуються технології HTTP (Hypertext Transport Protocol) і XML (Extensible Markup Language). Зібрані в такий спосіб метадані можуть бути представлені в будь-якому форматі, обраному співтовариством організацій, що вирішили об'єднати свої зусилля для створення інтегрованої федеративної ЕБ. Проте в протоколі OAI-PMH для забезпечення базового рівня інтероперабельності специфіковано формат Дублінського ядра [5]. Таким чином, метадані з різних неоднорідних джерел поєднуються в єдиній базі даних для того, щоб надати множину сервісів на основі таких агрегованих метаданих. Зв'язки між такими об'єднаними метаданими і відповідними інформаційними ресурсами (тобто з контентом інформаційних ресурсів) не визначаються в цьому протоколі, таким чином він не надає можливість робити повнотекстовий пошук за інформаційними ресурсами, а тільки за їхніми метаданими. Він просто дозволяє об'єднати інформаційні ресурси на рівні метаданих і саме на цьому рівні виконувати пошук.

Хоча концепція протоколу OAI-PMH досить проста, однак побудова на її основі відповідного набору сервісів, які б задовольнили потреби користувачів, залишається досить складною задачею. Ця задача цілком лежить на "плечах" провайдера сервісів, своєрідної пошукової системи, що дозволяє користувачам знаходити інформацію та досліджувати декілька репозиторіїв одночасно.

2.2. Інформаційна модель OAI-PMH. Виходячи з того факту, що джерелом виникнення протоколу була електронна публікація, модель даних OAI-PMH у загальному випадку інтерпретується в термінах бібліографічних даних, що описують академічний ресурс, хоча можливі й інші інтерпретації [1]. OAI-PMH має просту й гнучку інформаційну модель (рис. 1).

У верхній частині моделі – описуваний *ресурс* (resource). Це може бути як традиційний бібліотечний об'єкт (наприклад, книга, стаття), так і інші сутності (наприклад, зображення, поняття). Далі іде *елемент* (item) – складова репозиторія, за допомогою якої поширюються метадані про ресурс. Елемент концептуально являє собою контейнер, що зберігає або динамічно генерує метадані про окремий ресурс у

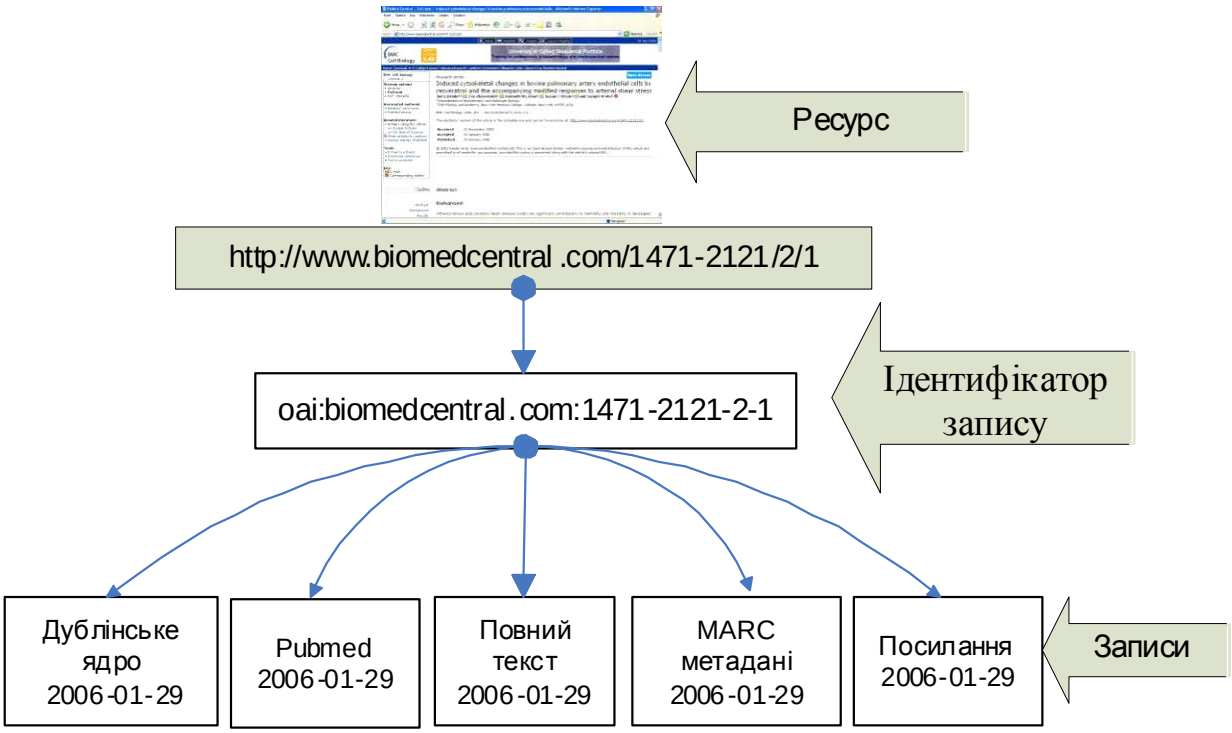


Рис. 1. Інформаційна модель OAI-PMH

декількох форматах, кожен з яких може бути зібраний у вигляді *запису* (*record*) через OAI-PMH. Кожний елемент має *ідентифікатор запису* (OAI identifier) або унікальний ідентифікатор, який однозначно визначає елемент у репозиторії, він використовується у запитах OAI-PMH для добування метаданих з елементів. Елемент може містити метадані у декількох форматах. Унікальний ідентифікатор вказує на елемент, і всі можливі записи, що є в одному елементі, разом використовують один і той самий унікальний ідентифікатор. Записи описують ресурс у довільному форматі метаданих, що може бути виражений в XML Schema. В ідею протоколу OAI-PMH закладена підтримка будь-яких схем опису метаданих, але він потребує обов'язкового включення в опис ресурсу набору метаданих Дублінського ядра (Dublin Core - DC). Також бажано включати в опис і більш розширені набори метаданих (наприклад, MARC).

Необхідно підкреслити, що ідентифікатор запису не є ідентифікатор документа (об'єкта). Очевидно, що багато користувачів захочуть одержати доступ до повного тексту ресурсу, описаному записом метаданих. Протокол рекомендує, щоб архіви використовували елемент запису метаданих для зв'язування запису з ідентифікатором (URL, URN, DOI та ін.) асоційованого документа (об'єкта). Для цієї мети обов'язковий формат DC надає елемент «ідентифікатор» (DC.Identifier).

2.3. Провайдери даних і сервісів. Концепція протоколу OAI-PMH виділяє дві ролі: провайдера даних та провайдера сервісів [3].

Провайдер даних – це служба, що підтримує створення і ведення одного чи більше репозиторіїв (бази документів, архівів, електронних бібліотек), здійснює публікацію своїх ресурсів, а також надає можливість доступу до своїх метаданих для їхнього використання в інших системах. Як правило, провайдер даних надає вільний доступ до своїх метаданих і, можливо але не обов'язково, надає вільний доступ до повних текстів своїх документів з НЕБ чи з інших інформаційних ресурсів. Провайдер даних може мати самостійний веб-інтерфейс для організації пошуку, перегляду і доступу до своїх ресурсів, а також інші сервіси, що надаються кінцевим користувачам. Провайдер даних самостійно вирішує питання про відкритість своїх інформаційних ресурсів і доступність до них. Зокрема, провайдер даних може прийняти рішення про інтеграцію усіх або частини своїх інформаційних ресурсів на рівні метаданих у провайдера сервісів і для цього організує експорт відповідних метаданих у форматі протоколу OAI-PMH.

Провайдер сервісів здійснює збір і збереження метаданих, наданих провайдерами даних, для надання кінцевим користувачам різних сервісів. До таких сервісів, зокрема, відноситься збереження й індексування метаданих з метою організації пошуку на їхній підставі необхідних документів. Зазначимо, що при організації пошуку необхідної інформації не відбувається звертання до провайдерів даних, тому що він здійснюється на основі збережених метаданих у провайдера сервісів. Провайдер сервісів може збирати не всі метадані, надані провайдером даних, а тільки ті, котрі доцільно збирати на основі тих чи інших критеріїв. Схема такої взаємодії показана на рис. 2.

Зазначимо, що в цій схемі той самий провайдер може грати обидві ролі, тобто забезпечувати як створення і ведення репозиторіїв інформаційних

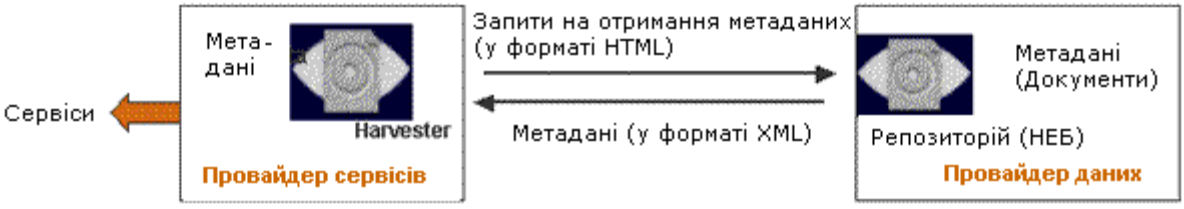


Рис. 2. Загальна схема взаємодії провайдера даних і провайдера сервісів

ресурсів, так і здійснювати збір метаданих від інших провайдерів та надавати на їхній основі необхідні сервіси. Запропоновані рішення є гнучкими для їхнього подальшого розвитку. Так, наприклад, на рис. 3 показано варіант розвитку цього підходу з використанням багатьох провайдерів сервісів. Таке рішення дозволяє, наприклад, створювати проблемно-орієнтовані чи функціонально-орієнтовані провайдери сервісів, а також, при необхідності, знижувати навантаження на провайдер сервісів у випадку істотного збільшення навантаження на нього.

полягає в обробці зібраних метаданих, наприклад, їх нормалізацію або перетворення в інший формат метаданих.

Нарешті, можна використовувати множину провайдерів сервісів, агрегаторів з одночасним комбінуванням з пошуком, як це показано на рис. 5.

2.4 Архітектурні рішення. З огляду на вищеописану концепцію OAI PMH, пропонується наступна архітектура розподіленої НЕБ (рис. 6). Організації, виходячи зі своїх можливостей і потреб, створюють, збирають і підтримують в елек-

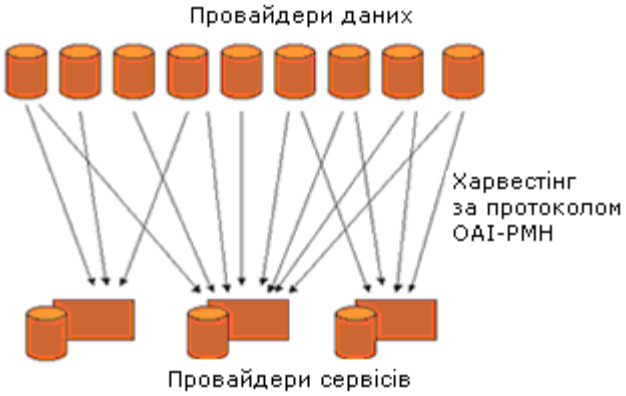


Рис. 3. Архітектура з багатьма провайдерами сервісів

На рис. 4 показано розширення з включенням агрегатора. Агрегатор виконує обидві функції – провайдера даних і сервісів. Як провайдер сервісів він збирає метадані від інших провайдерів даних і потім, виконуючи функцію провайдера даних, робить зібрані метадані доступними для збору іншими провайдерами сервісів. Завдання агрегатора

ктронному вигляді свої власні інформаційні ресурси у вигляді електронних колекцій або ЕБ.

У загальному випадку склад інформаційних ресурсів може бути довільним. Перелік пропонованих цими ЕБ сервісів також визначається в організаціях. Природно, що мінімальними сервісами є перегляд і пошук необхідної інформації та на-

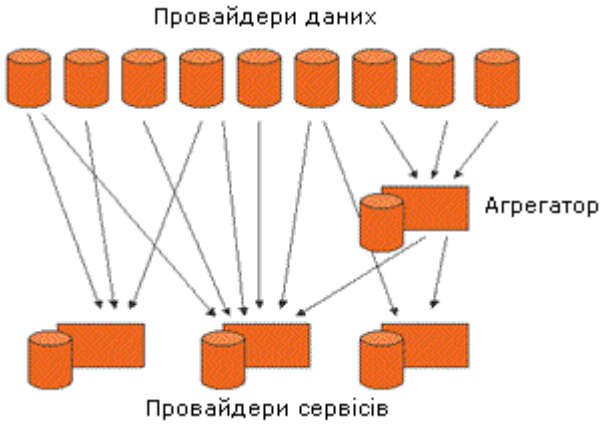


Рис. 4. Архітектура з агрегаторами

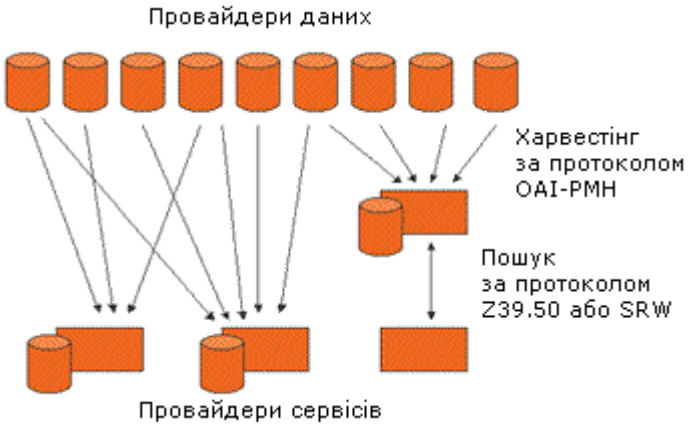


Рис. 5. Сполучення багатьох провайдерів сервісів, харвестингу і пошуку

дання результатів пошуку кінцевим користувачам. Крім того, можуть включатися інші сервіси, які є характерними для ЕБ [6]. Організації самостійно приймають рішення щодо функціональних можливостей створюваних ними ЕБ, включаючи і механізми обмеження прав доступу до тих чи інших ресурсів.

Виходячи з розроблених концепцій функціонування НЕБ, організації вибирають програмне забезпечення, здатне задовольнити необхідні (передбачувані) потреби. У розглянутій архітектурі не накладається ніяких принципових обмежень щодо обраного програмного забезпечення за винятком одного – воно повинно підтримувати механізм експорту метаданих у форматі OAI-PMH. У даний час існує багато інструментальних засобів створення ЕБ, що підтримують протокол OAI PMH, включаючи і безкоштовні програмні продукти. Такі НЕБ виконують у розглянутій архітектурі функції провайдерів даних.

Для інтеграції ЕБ створюється служба, яка буде виконувати функції провайдера сервісів. В її обов'язки входить:

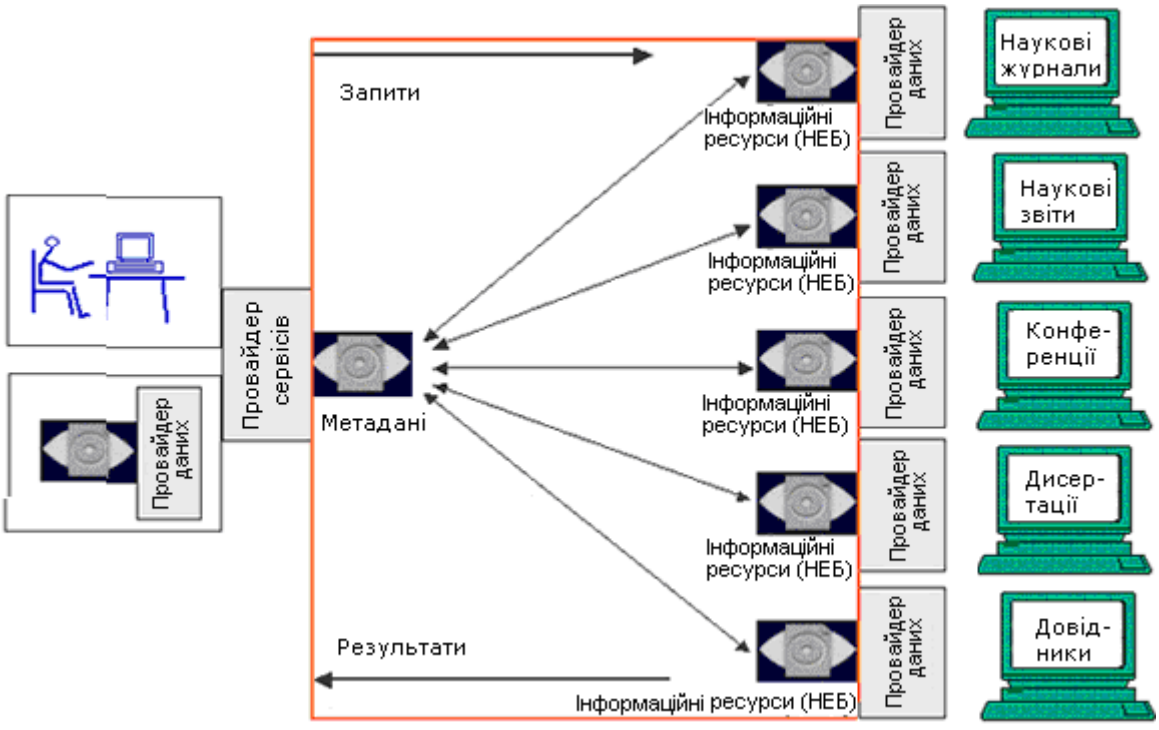


Рис. 6. Архітектура інтеграції НЕБ

– збір метаданих від провайдерів даних, що прийняли рішення про надання своїх інформаційних ресурсів в інтегровану пошукову службу провайдера сервісів;

– індексування метаданих з метою наступного надання сервісу з організації пошуку інформаційних ресурсів;

– надання сервісів з перегляду і пошуку інформаційних ресурсів провайдерів даних з можливим наданням посилань до повних текстів документів та інших сервісів.

Суттєвою перевагою такої архітектури є те, що пошук стає цілком незалежним від організації репозиторію документів. Пошук в одній великій базі даних або в тисячі мілких – для користувача є одним, оскільки цим займаються харвестери, які здійснюють пошук у відомих їм базах даних.

Архітектура в цьому плані нічого революційно нового не має. Термін "метапошуковик" та запит по декількох базах даних існує вже багато років. Новим тут є те, що використовуємий протокол організації OAI є відкритий, він не застосовується десь тільки в одній організації і служить для взаємодії різних баз. Одна з суттєвих місій OAI полягає у тому, щоб поширювати засоби, що дозволяють установам та колективам, навіть самим маленьким, розміщувати в мережі документи, які можна знайти за запитами стандарту OAI-PMH.

2.5. Проблеми інтеграції. Крім застосування єдиних протоколів доступу, проблеми інтероперабельності інформаційних бібліотечних систем включають формати метаданих та їх розширення (наприклад, класифікатори предметних областей), моделі документів, впорядковане пойменування та інш. Зупинимось на деяких з них більш докладно.

2.5.1. Уніфікація схем метаданих. Запропонований підхід передбачає вирішення проблеми уніфікації схем метаданих, з якими працюють провайдери даних. Така уніфікація істотно спрощує процес збору й індексації метаданих. В якості такої схеми пропонується набір метаданих Дублінського ядра (DC). У зв'язку з цим доцільно розробити єдину схему опису метаданих на основі DC для представлення описової інформації щодо всіх інформаційних ресурсів.

Проте, у загальному випадку підхід допускає існування різних схем метаданих. А на провайдер сервісів лягає відповідальність за вибір однієї зі схем як стандартної й ототожнення (відображення) інших схем з обраною. У такий спосіб буде надаватися гнучкість у підтримці різних способів опису інформаційних ресурсів. Втім в даний час відомі провайдери даних теж досить вдало вирішують ту ж проблему. Серед яких провідне місце посідає система програмного забезпечення для побудови провайдерів даних Eprints 3.0.

2.5.2. Проблема надання сервісів. Запропонований підхід вирішує проблему надання сервісу інтегрованого пошуку необхідних ресурсів. Проте, при відповідному розширенні функціональних можливостей провайдера сервісів можна охоплювати також й інші сервісні функції.

2.5.3. Проблема валідності та захисту інтелектуальної власності. Використання OAI забезпечить належний рівень правдивості та відповідності інформації, оскільки будуть існувати гарантії що подана інформація є остаточною та дані дослідження є офіційним оприлюдненням. Водночас не буде порушуватися право інтелектуальної власності, оскільки інформація що знаходиться на дата провайдері уже зазвичай пройшла перевірку на цю норму.

3. Інструментарій реалізації розглянутого підходу

3.1. Програмне забезпечення провайдерів даних. Як було вищезазначено провайдер даних підтримує створення і ведення одного чи більше репозиторіїв, здійснює публікацію своїх ресурсів, а також надає можливість доступу до своїх метаданих для їхнього використання в інших системах. Набір платформ програмного забезпечення для побудови таких репозиторіїв постійно змінюється. Найбільш відомі та поширені на сьогодні:

– Archimede (http://www1.bibl.ulaval.ca/archimede/index.en.html);
– CDSware (http://cdsware.cern.ch);
– DSpace (http://www.dspace.org);
– Eprints (http://software.eprints.org);
– Fedora (http://www.fedora.info/index.shtml);
– Greenstone (http://www.greenstone.org/cgi-bin/library).

Цей список не є вичерпним. Існує цілий ряд комерційних продуктів. Добрий огляд та аналіз цих платформ можна знайти в [7–8]. Всі перераховані системи мають набір спільних характеристик: це – відкрите програмне забезпечення, що поширюється під ліцензією GNU; репозитарії які побудовані на цих платформах OAI-сумісні, тобто підтримують протокол збору метаданих OAI-PMH (Greenstone поки що підтримує цей протокол лише частково [9]); системи підтримують повнотекстовий пошук для ресурсів визначених форматів; а також обов'язково використовують стандартний набір метаданих DC для опису своїх ресурсів. Як приклад стисло зупинимось на найбільш поширеній системі EPrints.

3.1.1. Програмне забезпечення EPrints, загальний опис. EPrints – вільно розповсюджуване програмне забезпечення під ліцензією GNU, що використовується для формування й керування Відкритими Архівами. На сьогодні у світі створено з використанням EPrints більше 200 архівів з більш ніж 170 000 записами. ПЗ EPrints може використовуватися для створення архівів робіт наукових досліджень, зображень, даних і інших видів цифрової інформації.

ПЗ EPrints розроблено в Школі електроніки й інформатики Університету Саутгемптона (Великобританія). Зі створенням системи EPrints тісно зв'язаний проект TARDis (Targeting Academic Research for Deposit and Disclosure) [10], основним завданням якого було дослідження всіх сторін створення електронного архіву з метою розробки типового архіву для академічних установ.

Основними системними вимогами для EPrints версії 3.0 є: ОС Unix, мова програмування Perl 5.8.x, сервер баз даних MySQL 4.1.x, веб-сервер Apache 2.x. Апаратні вимоги – сервер з обсягом оперативної пам'яті 1 Гб і процесором з тактовою частотою 1 Ггц і більше та відповідним дисковим простором для зберігання повнотекстових документів, при великому навантаженні на сервер бажано використовувати жорсткий диск з підтримкою SCSI (Small Computer Systems Interface).

3.1.2. Функції та можливості EPrints. ПЗ EPrints надає наступні можливості [6]:

– створення електронних архівів;
– підтримка файлів різного формату;
– індексація файлів PDF, ASCII, Microsoft Word, HTML;
– перегляд формул в документах, створених на мові LaTex;
– виконання повнотекстового та розширеного пошуку (по метаданим);
– гибке адміністрування прав доступу;
– гибка інтеграція з основним сайтом (з використанням основного стилю оформленния Веб-сайту організації).

EPrints має докладну документацію за всіма аспектами проекту. Сайт демонстрації demoprints.EPrints.org представляє різноманітну інтерактивну допомогу. Крім того, документація використовує технологію Wiki (http://www.wiki.org), де користувачі EPrints розміщають практичні поради, сценарії та іншу корисну інформацію. Приклади застосування та більш докладний опис цієї системи можна знайти у [11–15].

3.1.3. Експорт метаданих у формат OAI PMH. У розглядуваній системі реалізовано підхід підтримки багатьох наборів метаданих, серед яких є і DC, що декларує протокол OAI-PMH як обов'язкового. На рис. 7 показані можливості системи Eprints 3.0 щодо експорту своїх метаданих. Eprints виставляє метадані у форматі DC своїх ресурсів, які знаходяться у відкритому доступі, тобто є загальнодоступними. Якщо OAI-сервіс затребує інші формати метаданих, наприклад, MODS, система надає і таку можливість.

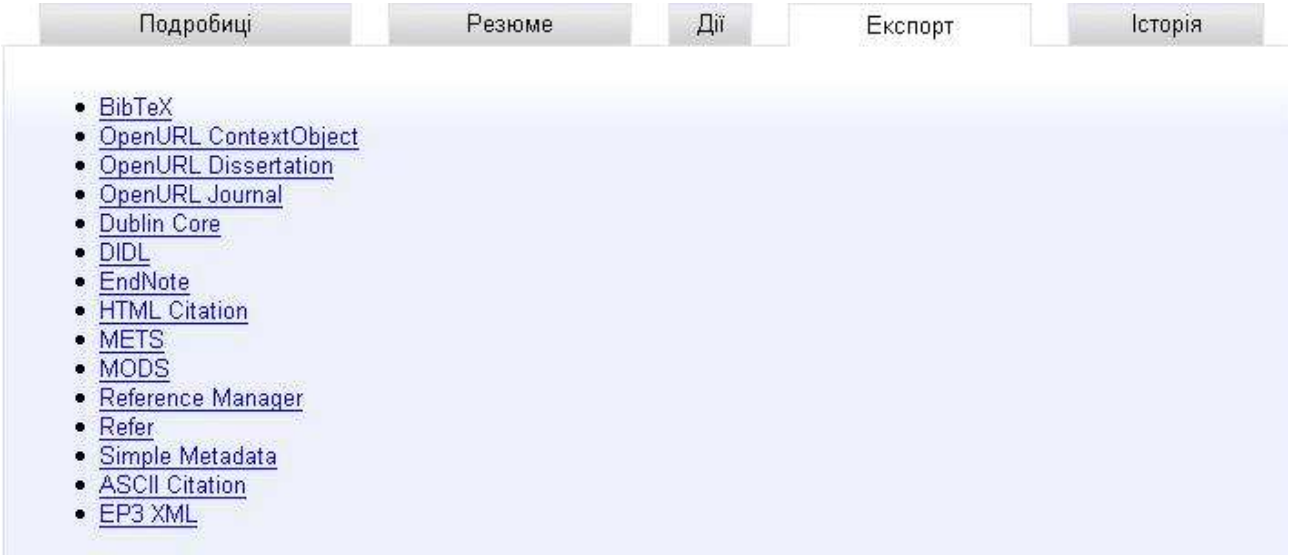


Рис. 7. Перелік форматів конвертації метаданих у Eprints 3.0

Набір форматів метаданих в які можливо експортувати дані з Eprints 3.0 (рис. 7):

– BibTeX- бібліографічний формат метаданих;

– OpenURL ContextObject – стандарт метаданих ANSI/NISO Z39.88-2004 [16] для контекстнозалежних сервісів, зазвичай повнотекстового пошуку;

– OpenURL Dissertation – попередній стандарт, спеціалізований для ресурсів типу дисертації;

– OpenURL Journal – попередній стандарт, спеціалізований для ресурсів типу журнал (рис.8);

– Dublin Core − Дублінське ядро стандарт метаданих ANSI/NISO Z39.85-2001 (а також стандарт ISO 15836-2003) [5];

– DIDL - Digital Item Declaration Language, за допомогою якого в MPEG-21 описуються складні електронні об'єкти. Це описання вводить ряд абстрактних понять, які формують модель даних [17];

– EndNote – поширений у науковому співтоваристві бібліографічний формат посилань цитування (http://www.ecst.csuchico.edu/~jacobsd/bib/formats/endnote.html), використовується в однойменному комерційному продукті [18];

– HTML Citation – HTML-формат цитування для документів, що використовується для перегляду або пошуку документів у системі Eprints 3.0;

– METS – стандарт кодування та передачі метаданих [17];

– MODS – схема метаданих опису об'єкта [17];

– Reference Manager – формат метаданих для створення та управління архівами та бібліографічними описами, експорт у цей формат дозволить використовувати метадані Eprints 3.0 у системі Reference Manager, що є системою того ж класу що і EndNote. Порівняльний аналіз цих систем можна знайти на http://thomsonresearchsoft.com/compare/;

– Refer – формат файлу побудований у відповідності до спеціально відформатованого документу (troff) [19], він може використовуватися практично будь-якою програмою і є доволі узагальненим форматом бібліографій (http://www.ecst.csuchico.edu/~jacobsd/bib/formats/refer.html);

– Simple Metadata (SimpleMDE) – цей набір метаданих є підмножиною повного можливого набору метаданих і використовується, коли виконується швидка анотація [20];

– ASCII Citation – звичайний текстовий формат;

– EP3 XML – експорт до XML.

Як приклад одного з запропонованих форматів представлення метаданих на рис. 8 наведено приклад запису у форматі OpenURL Journal.

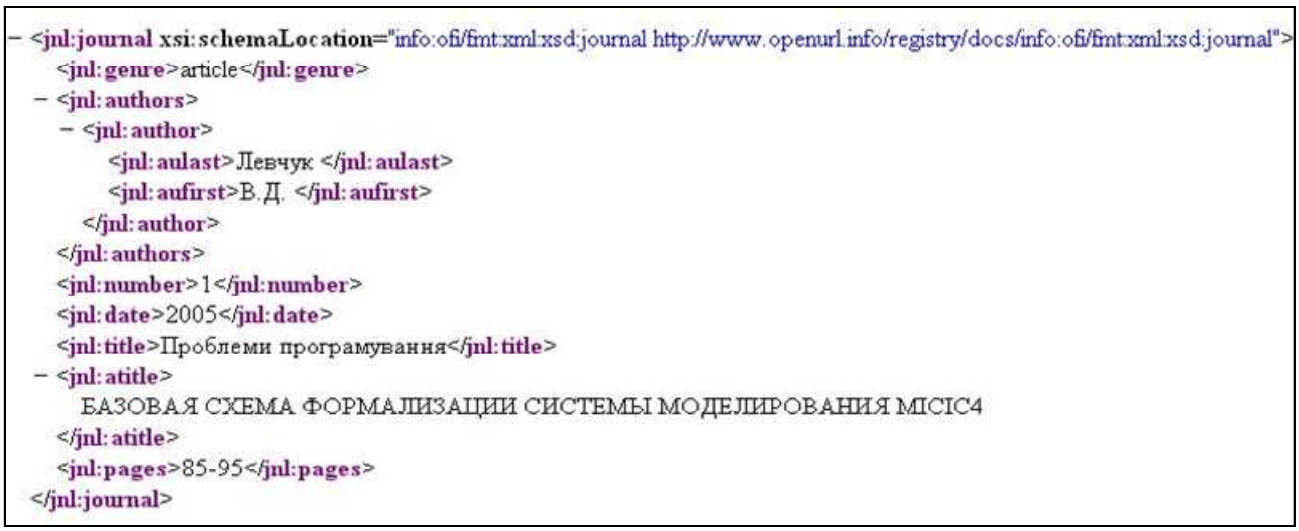


Рис. 8. Приклад запису метаданих у форматі OpenURL Journal

3.2. ПЗ провайдерів сервісів. На сьогоднішній момент існує кілька проектів по створенню ПЗ для реалізації сервіс-провайдеру:

– IWF Metadata Harvester (http://ftp.gwdg.de/pub/gnu2/iwfmdh/). Недоліком цього програмного забезпечення є використання Microsoft Windows Server 2003, Visual Studio, Visual C++ .NET і Microsoft SQL Server 2000, що усікає кросплатформні можливості ПЗ, хоча саме ПЗ поширюється по ліцензії GNU;

– SilvaOAI Extension (http://www.infrae.com/products/oaipack). Це розширення до системи контент-менеджера Silva Content Management System, що дозволяє працювати з провайдерами даних за протоколом OAI-PMH. Це ПЗ працює тільки в системі Silva CMS;

– Thumbgrabber 2.0 (http://prdownloads.sourceforge.net/uilib-oai/). Це також ПЗ для реалізації функцій сервіс-провайдеру протоколу OAI-PMH. Основні системні вимоги це Windows 2000 або Windows XP, Microsoft Active COM (наприклад DSO OLE Document Properties Reader 2.0, CAPICOM v2.0 Type Library і ін.). До недоліків системи варто віднести відсутність кросплатформності, проблеми з локалізацією.

Далі більш докладно зупинимось на найбільш поширеній на сьогодні, безкоштовній системі того ж класу, що й попередні – PKP Open Archives Harvester.

3.2.1. PKP Open Archives Harvester призначена для індексування метаданих за протоколом OAI-PMH. PKP Open Archives Harvester 2.0.0 (http://pkp.sfu.ca/?q=harvester, далі для скорочення будемо називати Harvester), це програма яка повністю реалізована як веб-застосування і є кросплатформною, а її модульний підхід дозволяє нарощувати функціональність. Система має докладну документацію користувача і розроблювача. Дата останньої версії – січень 2007.

Даний сервіс провайдер підтримує розробку системи відкритих журналів OJS, що розроблена в рамках проекту Public Knowledge Project [21] і призначена для керування відкритими журналами й публікаціями із ціллю спрощення доступу наукового суспільства до їхнього змісту.

Програмний продукт реалізований на PHP версії 4.2 і працює із сервером баз даних MySQL версії 3.23.23 і вище або PostgreSQL, у якості веб-сервера можуть виступати як Apache від 2.0.4, так і Microsoft IIS 6.

Harvester версії 2 має наступні властивості:

– Можливість збирати OAI метадані в різних схемах (форматах), наприклад, некваліфікованого DC, розширеного DC системи PKP (Open Journal Systems/Open Conference Systems, OSJ/OCS, http://pkp.sfu.ca/ojs), MODS та MARCXML. Інші схеми підтримуються за допомогою плагинів.

– Підтримка простого й розшире-

ного пошуку, використовуючи поля переходів (crosswalked fields) репозиторіїв, з яких збирається інформація. Можливо здійснювати розширений пошук для репозиторіїв з спільною схемою метаданих, коли використовуються поля з цієї схеми, або якщо схеми різні, але встановлені поля переходів.

– Можливість виконання гранулярного (агрегованого) збору метаданих.

– Легкість налаштування інтерфейсу користувача на основі шаблонів HTML і CSS.

– Масштабованість.

Установка Harvester здійснюється за допомогою веб-інтерфейсу. Після установки система не вимагає ніяких додаткових настроювань і готова до роботи.

3.2.2. Додавання нового архіву в PKP Open Archives Harvester. Після успішної установки Eprints і Harvester наступним кроком є додавання створеного архіву, або провайдеру даних, в середовище сервіс провайдеру. Для цього у форму (рис. 9) слід внести:

– назву провайдеру даних;

– короткий опис;

– URL провайдеру даних, за яким доступний інтерфейс користувача;

– тип протоколу за яким збираються метадані (за замовчуванням OAI);

– базовий URL для збору метаданих за протоколом OAI;

– метод індексування, ListRecords чи ListIdentifiers (за замовчуванням ListRecords);

– схему метаданих (за замовчуванням DC).

При виборі методу індексування архіву сервіс провайдер надає два значення ListIdentifiers або ListRecords. Команда ListIdentifiers щодо протоколу OAI [3, 22] використовується для збору заголовків ідентифікаторів записів репозиторію. Додаткові аргументи дозволяють шукати ідентифікатори селективно − ґрунтуючись на їхній приналежності до певного набору в

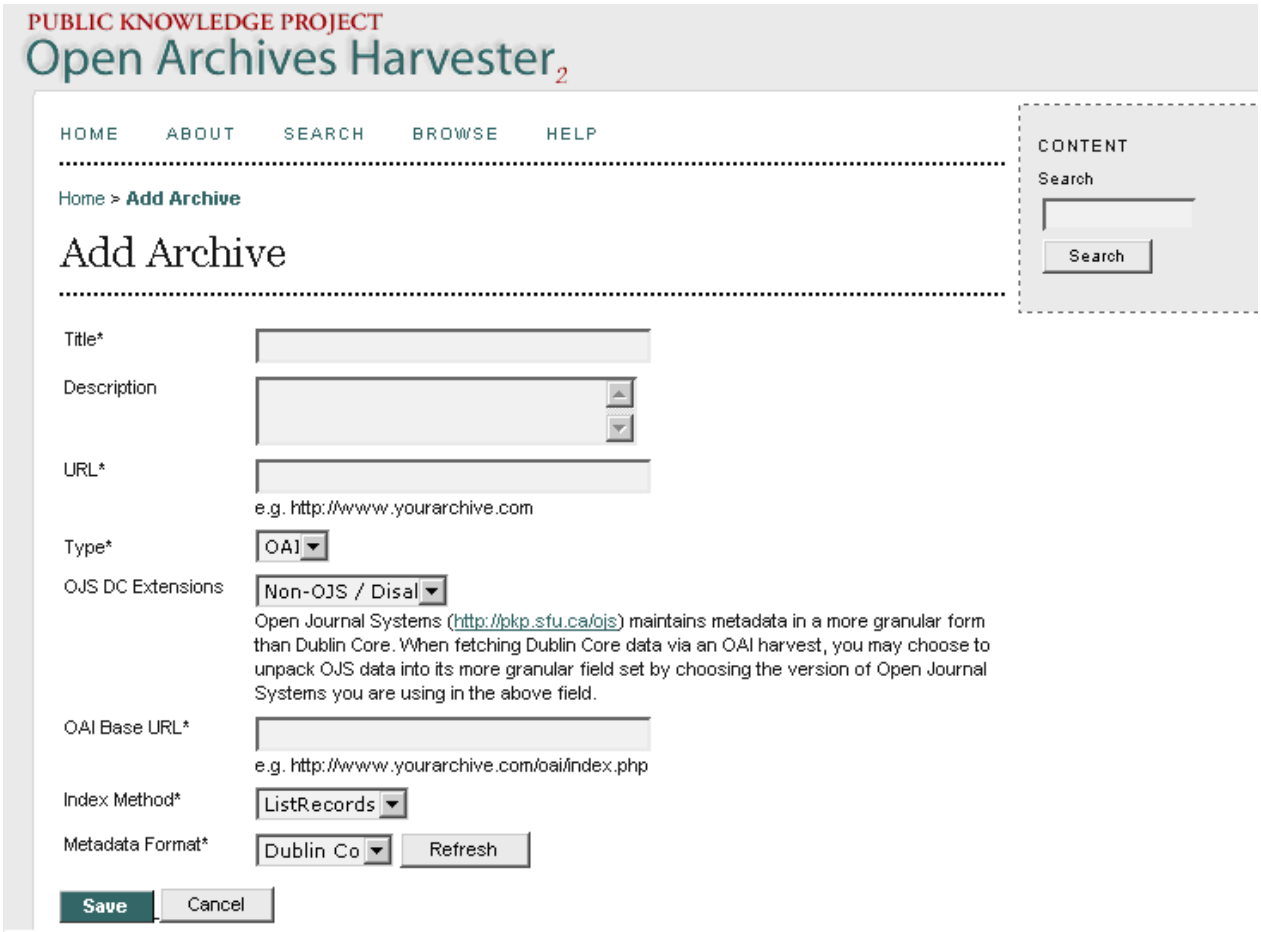
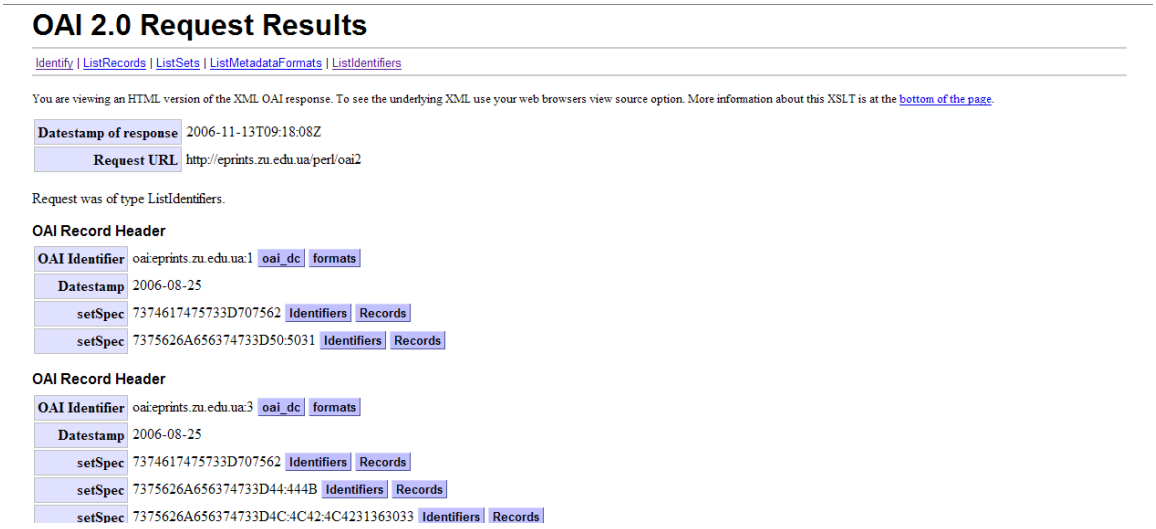


Рис. 9. Приклад форми для занесення даних про новий архів

репозиторії або на часових параметрах (модифікація, створення або видалення в зазначений період). На рис. 10 показано приклад XML-файлу відповіді на запит ListIdentifiers, до якого застосовані стилі оформлення.

Команда ListRecords призначена для збору записів з архіву, що включають повні метадані. Використовуючи додаткові параметри setSpec, setName, setDescription можлива вибіркова індексація архіву, основуючись на приналежності запису до конкретного набору в архіві або на часових параметрах. Приклад XML-файлу відповіді на запит ListRecords, до якого застосовані стилі оформлення (рис. 11).



OAI 2.0 Request Results

Identify | ListRecords | ListSets | ListMetadataFormats | ListIdentifiers

You are viewing an HTML version of the XML OAI response. To see the underlying XML use your web browsers view source option. More information about this XSLT is at the bottom of the page.

Datestamp of response	2006-11-13T09:18:08Z
Request URL	http://eprints.zu.edu.ua/perl/oai2

Request was of type ListIdentifiers.

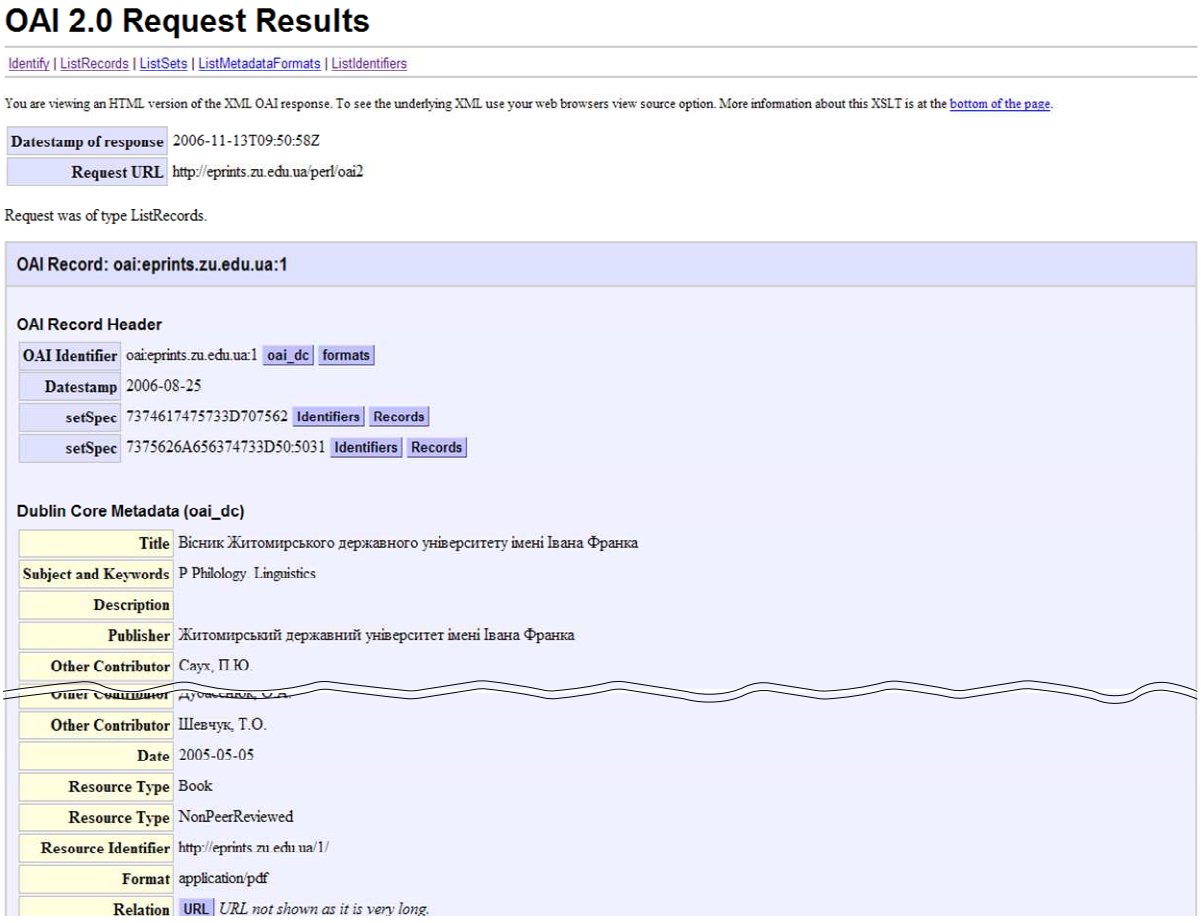
OAI Record Header

OAI Identifier	oai:eprints.zu.edu.ua:1 oai_dc formats
Datestamp	2006-08-25
setSpec	7374617475733D707562 Identifiers Records
setSpec	7375626A656374733D50:5031 Identifiers Records

OAI Record Header

OAI Identifier	oai:eprints.zu.edu.ua:3 oai_dc formats
Datestamp	2006-08-25
setSpec	7374617475733D707562 Identifiers Records
setSpec	7375626A656374733D44:444B Identifiers Records
setSpec	7375626A656374733D4C:4C42:4C4231363033 Identifiers Records

Рис. 10. Результат на запит ListIdentifiers



OAI 2.0 Request Results

Identify | ListRecords | ListSets | ListMetadataFormats | ListIdentifiers

You are viewing an HTML version of the XML OAI response. To see the underlying XML use your web browsers view source option. More information about this XSLT is at the bottom of the page.

Datestamp of response	2006-11-13T09:50:58Z
Request URL	http://eprints.zu.edu.ua/perl/oai2

Request was of type ListRecords.

OAI Record: oai:eprints.zu.edu.ua:1

OAI Record Header

OAI Identifier	oai:eprints.zu.edu.ua:1 oai_dc formats
Datestamp	2006-08-25
setSpec	7374617475733D707562 Identifiers Records
setSpec	7375626A656374733D50:5031 Identifiers Records

Dublin Core Metadata (oai_dc)

Title	Вісник Житомирського державного університету імені Івана Франка
Subject and Keywords	P Philology. Linguistics
Description	
Publisher	Житомирський державний університет імені Івана Франка
Other Contributor	Саух, П.Ю.
Other Contributor	Шевчук, Т.О.
Date	2005-05-05
Resource Type	Book
Resource Type	NonPeerReviewed
Resource Identifier	http://eprints.zu.edu.ua/1/
Format	application/pdf
Relation	URL *URL not shown as it is very long.*

Рис. 11. Результат запиту ListRecords

3.2.3. Управління репозиторіями в PKP Open Archives Harvester. Користувач із правами адміністратора може в повному обсязі керувати репозиторіями системи. При індексуванні репозиторію можна здійснювати вибіркове індексування за такими характеристиками, як статус публікації (опублікована, в друці, представлена на розгляд, не публікована), предметний класифікатор, що прийнятий у даному архіві та ін. Потім здійснюється індексування метаданих вибраного віддаленого репозиторію. Можлива вказівка часових умов, коли будуть вибиратися записи на віддалених репозиторіях.

Окремої уваги заслуговує агрегатор сервіс-провайдеру. Він реалізований у вигляді переходів (crosswalks) пошукових полів метаданих у різних схемах. Переходи використовуються для визначення подібних полів у різних схемах метаданих. Тобто якщо будуть встановлені поля-переходи, то можливо виконати перехресний пошук в репозиторіях з різними схемами метаданих. Іноді для такої задачі використовується термін відображення (mapping). На рис. 12 показано співставлення поля “Title” для різних схем DC, MARC, MODS. Як видно на рис.12, поле “Title” у схемі DC має один запис, водночас, як у MARC чи MODS до поля “Title” належить група записів.

3.2.4. Сервіс пошуку інформації (Reading Tools). Для кожного репозиторію у системі є можливість вказівки предметного напрямку. У такий спосіб при перегляді конкретного запису провайдеру сервісів можливо відправляти запити, які містять метадані даного запису на сторонні пошукові машини, бібліотеки або словники. Вибір набору пошукових сервісів залежить від предметного спрямування репозиторію. Наприклад, якщо зазначено, що репозиторій спрямований на комп'ютерні науки, то запити ми можемо відправити на Free On-Line Dictionary of Computing або на більш загальні словники. Наприклад, нехай нас зацікавив запис "Enhancing OAI Metadata for Eprint Services: two proposals" рис. 13. При перег-

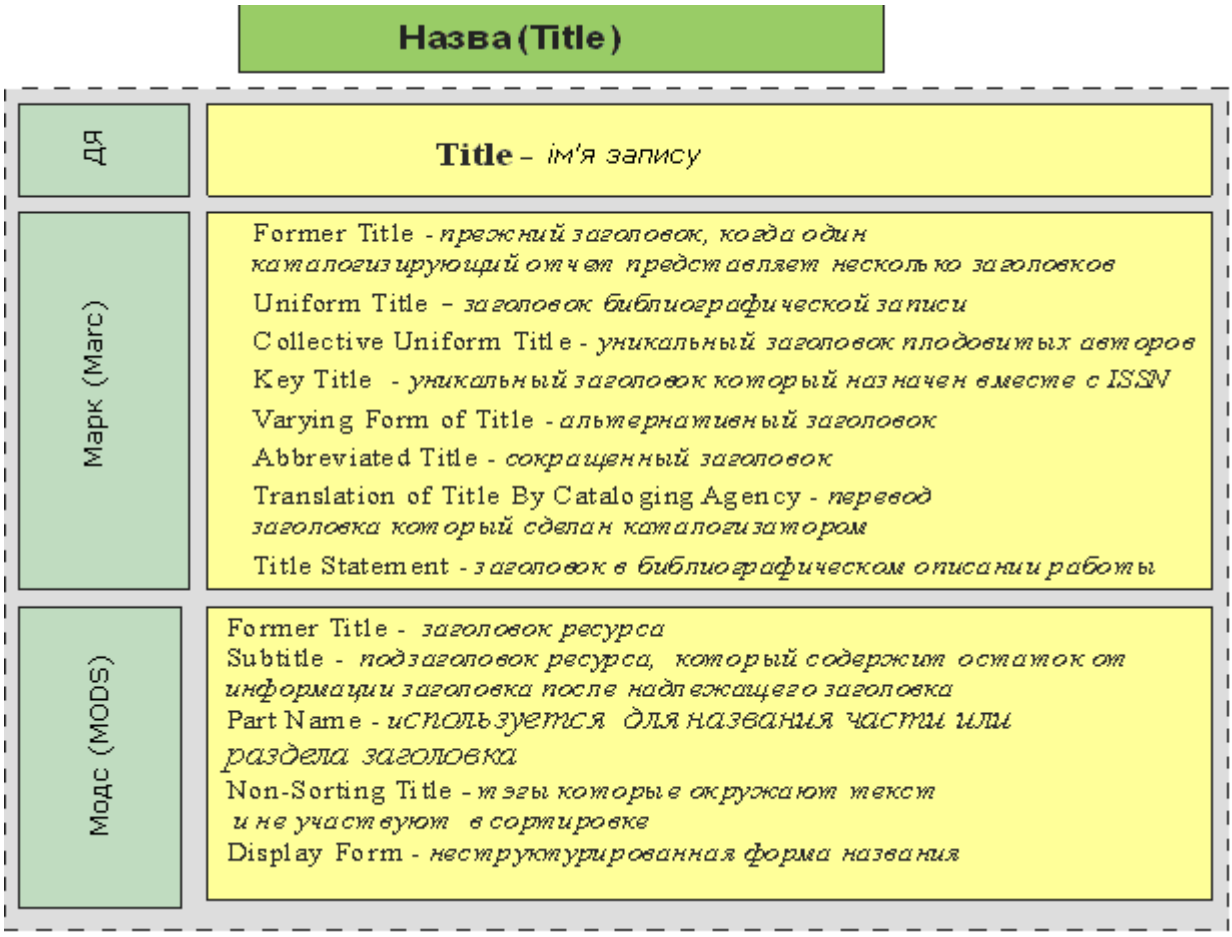


Рис. 12. Відповідності поля “Назва” у різних схемах метаданих

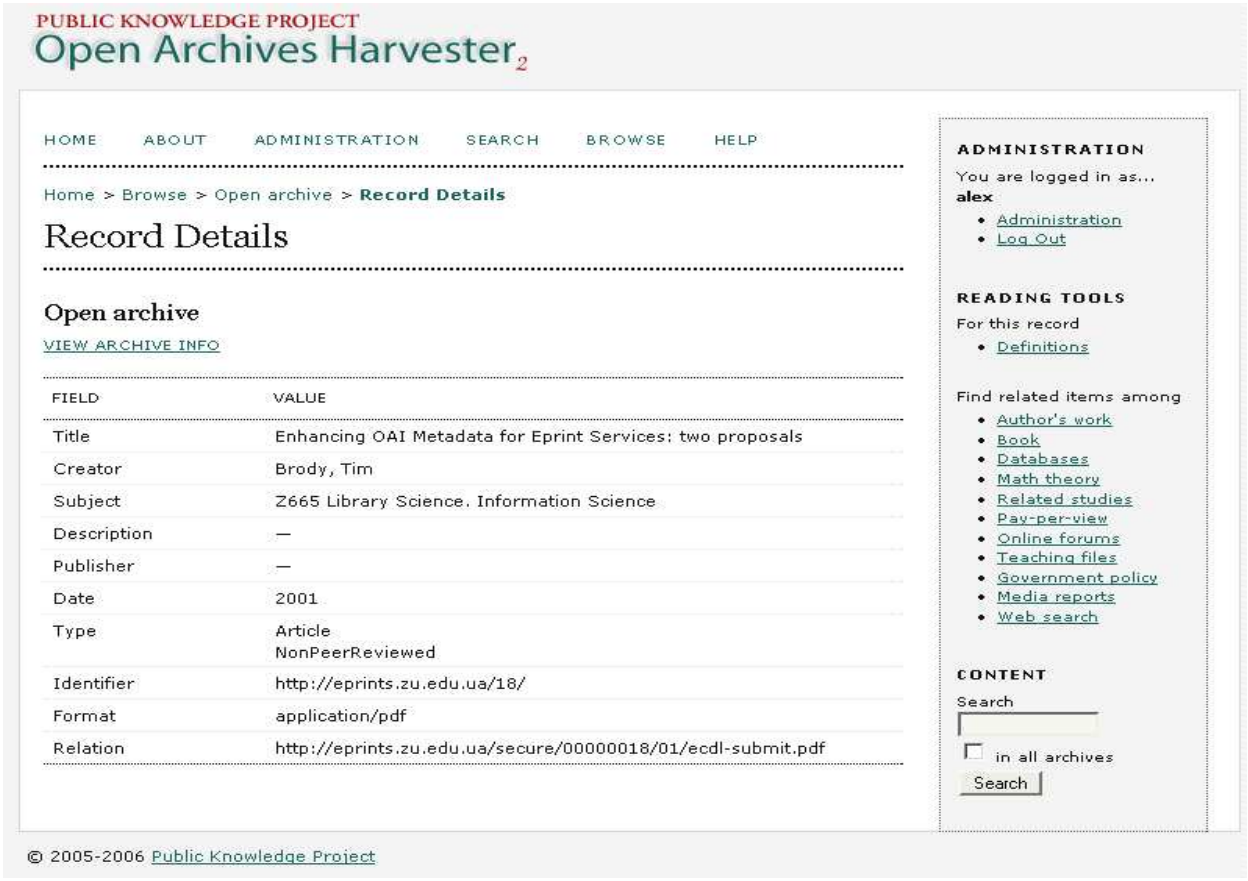


Рис. 13. Перегляд детальних метаданих

ляді метаданих для уточнення значення поняття OAI, подвійне клацання миші по слову "OAI" дасть нам сторінку з можливістю вибору інформаційного сервісу рис.14.

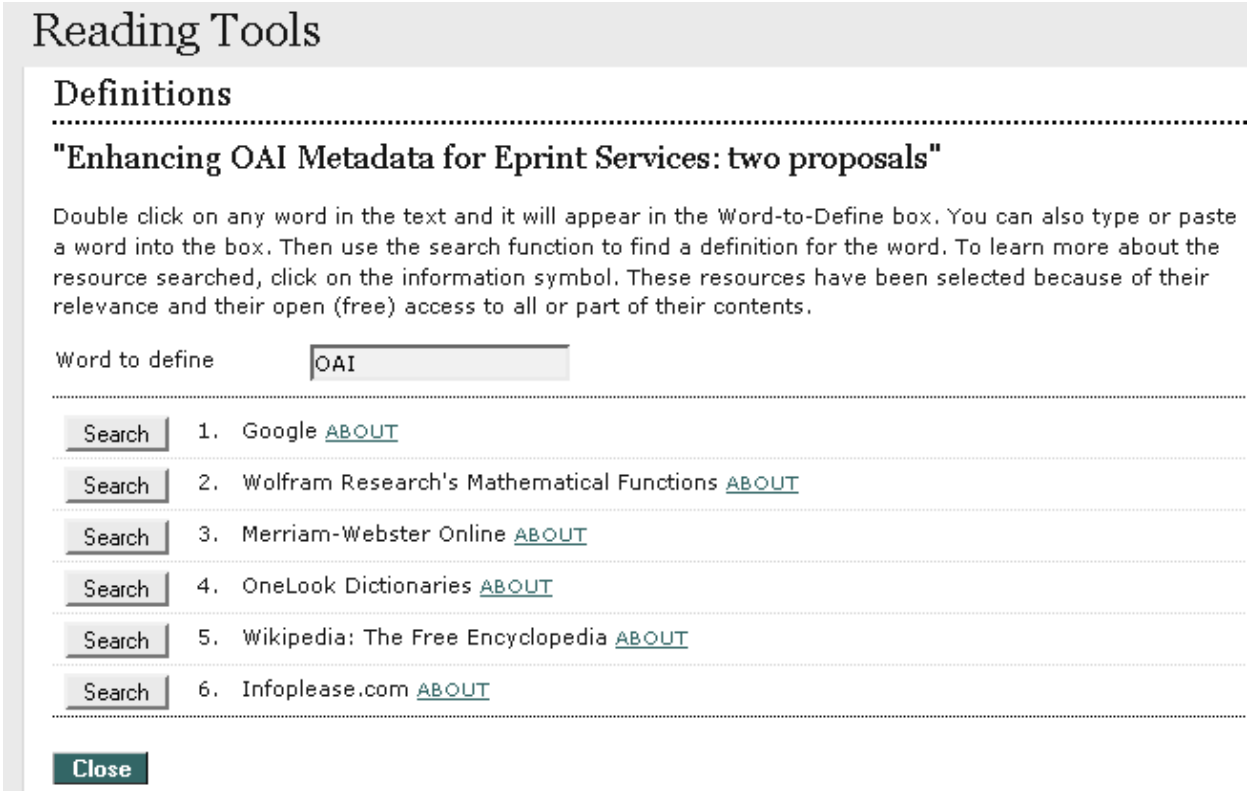


Рис. 14. Сторінка вибору інформаційного сервісу для одержання додаткової інформації про поняття "OAI"

Висновки

Існує ряд підходів до вирішення проблеми інтеграції наукових репозиторіїв, вони відрізняються ступінню централізації ресурсів, метаінформації та пошукових сервісів. Підхід, який реалізований у системах, що підтримують протокол OAI-PMH, є одним із них. Він полягає у тому, що ресурси залишаються в організації, де створені наукові архіви (репозиторії, бібліотеки). Такі репозиторії можуть бути побудовані, наприклад, за допомогою вільно розповсюдженого програмного забезпечення Eprints для створення відкритих архівів. Вони виконують роль провайдерів даних. Для об'єднання таких репозитріїв за предметним чи галузевим принципом на центральний сервер, що виконує роль провайдеру сервісів, копіюється метаінформація, з архівів які працюють в інтегрованій системі. Пошук здійснюється на центальному сервері, а за повним текстом користувач звертається до відповідного архіву. Провайдер сервісів можна реалізувати, наприклад на основі програмного забезпечення PKP Open Archives Harvester. Серед переваг такого підходу можна виділити простоту, сучасність, забезпечення високої якості сервісів, можливості розвитку, масштабованість, можливість інтегрувати ресурси з багатьма іншими відкритими ресурсами.

Протокол OAI-PMH побудовано на основі XML і для забезпечення сумісності потребує обов'язкової підтримки схеми опису метаданих Dublin Core. Водночас рекомендується додаткова підтримка інших більш складних форматів метаданих. Якщо будь-яка група організацій домовляється про використання додаткового формату, вони можуть легко розширити можливості своєї взаємодії для вирішення будь-яких специфічних задач, але залишаючись при цьому в полі ресурсів, що доступні через протокол OAI.

Нами був створений демонстраційний сервіс провайдер на основі PKP Open Archives Harvester який збирав метадані з створеного раніше архіву [12] й доступний за адресою http://oai.zu.edu.ua:8080/. Таким чином було побудовано макетний приклад системи OAI-PMH обміну метаданих, що дає можливість виконувати розподілений пошук, зручно нарощувати систему та надавати додаткові сервіси обслуговування користувачів.

1. *Когаловский М.Р.* Тенденции развития технологий управления информационными ресурсами в электронных библиотеках // Тр. VIII Всероссийской научн. конф. Электронные библиотеки: перспективные методы и технологи. – Суздаль, Россия. – 2006. – С. 46 – 55.
2. *Лагозе К., Ван де Зомпель Г.*. Инициатива «Открытые архивы»: создание среды с высокой степенью интероперабельности. Электронные библиотеки. – 2001. – Т. 4. Вып. 6. http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2001/part6/LS
3. *The* Open Archives Initiative Protocol for Metadata Harvesting Protocol Version 2.0 of 2002-06-14. http://www.openarchives.org /OAI/2.0/openarchivesprotocol.htm
4. *Жижимов О.Л.* Введение в Z39.50. – Новосибирск, 2003. – http://z3950.uigm.nsc.ru:210/introduction/Part_tit.htm
5. *ANSI/NISO* Z39.85-2001. The Dublin Core Metadata Element Set. – National Information Standards Organization. - 2001. http://www.techstreet.com/cgi-bin/pdf/free/335284/z39.85-2001.pdf
6. *Резніченко В.А. Захарова О.В., Захарова Е.Г.* Електронні бібліотеки: інформаційні ресурси та сервіси // Проблеми програмування. − 2005. − № 4 – С. 60–72.
7. *A Guide* to Institutional Repository Software, 3rd Edition, Open Society Institute, August 2004.
8. *Резниченко В.А., Зарова О.В., Захарова Е.Г.* Каталог програмних засобів створення електронних бібліотек. Інститут програмних систем НАН України. – Київ, 2006. - 32 с. - бібл., 20 назв. Укр. - Деп. В ДНТБ України.
9. *Резниченко В.А., Проскудина Г.Ю., Овдий О.М.* Создание цифровой библиотеки коллекций периодических изданий на основе Greenstone // Электронные библиотеки.-2005. - 8. - Вып. 6. http:// www.elbib.ru/index.phtml?page=elbib/rus/journal/2005/part6.

10. *Gutteridge C., Hitchcock S., Simpson P., Hey J.* Report on the technical issues of using GNU EPrints software for the development of an institutional e-Print repository at the University of Southampton: TARDis deliverable D.2.3.2. 2003. http://tardis.EPrints.org/
11. *Gutteridge C.* EPrints 2.3 Documentation. October12, 2005. http://www.EPrints.org/documentation/tech/EPrints-docs.pdf
12. *Новицкий А.В., Резниченко В.А., Проскудина Г.Ю.* Пример построения научных архивов с помощью EPrints. RCDL'2006, 17–19 Октября, Суздаль, Россия, С. 154−161.
13. *Новицкий А.В., Резниченко В.А., Проскудина Г.Ю.* Создание научных архивов с помощью системы EPrints. Электронные библиотеки. –2006. – **9**. – Вып. 4. http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2006/part4/Novitski
14. *EPrints* Self-Archiving FAQ, http://www.EPrints.org/openaccess/self-faq/
15. *Sale A.* Eprint website for the University of Tasmania. August 2004. http://EPrints.comp.utas.edu.au:81/archive/00000011/
16. ANSI/NISO Z39.88-2004. The OpenURL Framework forContext-Sensitive Services. – National Information Standards Organization. – 2005. http://www.niso.org/standards/resources/Z39_88_2004.pdf
17. *Understanding* Metadata. National Information Standards Organization. – 2004. – http://www.niso.org
18. *EndNote* ® …Bibliographies Made Easy. Getting Started Guide. – Thomson. – 2006. - 86 p. http://scientific.thomson.com/media/pdfs/ENXGettingStartedGuide.pdf
19. *Joseph F. Ossanna, Brian W. Kernighan, Gunnar Ritter*. Heirloom Documentation Tools. Nroff/Troff User's Manual. – 2007. – http://heirloom.sourceforge.net/doctools/troff.pdf
20. *Simple* Metadata Annotation Specification. - Version 6.2 – Linguistic Data Consortium February 3, 2004. – http://projects.ldc.upenn.edu/MDE/Guidelines/SimpleMDE_V6.2.pdf
21. *EPrints* Open Access. http://www.EPrints.org/openaccess/
22. *The* Effect of Open Access on Citation Impact / T. Brody, H. Stamerjohanns, F. Vallieres, S. Harnad, Y. Gingras, C. Oppenheim http://www.ecs.soton.ac.uk/ ~harnad/Temp/OATAnew.pdf

Отримано 02.02.2007

Про авторів:

Резніченко Валерій Анатолійович,
кандидат фізико-математичних наук,
старший науковий співробітник,

Проскудіна Галина Юріївна,
науковий співробітник,

Новицкий Олександр Вадимович,
аспірант.

Місце роботи авторів:

Інститут програмних систем НАН України,
03187, Київ-187, проспект Академіка Глушкова, 40.
Тел. (044) 526 5139, 526 6033
Email: reznich@isofts.kiev.ua
gupros@isofts.kiev.ua
alex@zu.edu.ua