УДК 004.4

ПОБУДОВА ІНТЕГРОВАНОЇ Е-ІНФРАСТРУКТУРИ ПІДТРИМКИ НАУКОВИХ ДОСЛІДЖЕНЬ В ГРІД-СЕРЕДОВИЩІ

Є.М. Твердохліб, П.І. Перконос

Інститут програмних систем НАН України,
03680, Київ, Україна, проспект Академіка Глушкова, 40,
Тел. +38(044) 526 3068, e-mail: eugine@nas.gov.ua, perkonos@nas.gov.ua

У статті висвітлені проблеми побудови інтегрованої е-Інфраструктури інформаційної підтримки наукових досліджень у грід-середовищі. Описані найвідоміший проект створення електронної грід-інфраструктури D4Science і архітектура програмної основи даного проекту – системи gCube. Наведені основні результати розробки подібної е-інфраструктури в Українському національному гріду.

The article deals with problems of building an integrated e-Infrastructure of information support of scientific research in the Grid environment. We describe the most famous project is creating an electronic grid infrastructure in Europe D4Science and the architecture of the program framework of this project – gCube system. The main results of the development of e-Infrastructures in the Ukrainian national grid are reported.

Вступ

Складність і масштабність наукових задач, що вирішуються в теперішній час, супроводжується постійно зростаючими потребами у все більших інформаційних ресурсах для накопичення величезних обсягів даних експериментальних досліджень, спостережень і виконання величезного обсягу обчислень в процесі обробки експериментів і розрахунків теоретичних моделей. Грід-технології, що бурхливо розвиваються останніми роками, направлені в першу чергу на ефективне використання власне обчислювальних ресурсів грід-інфраструктури (так званий «Computing Grid»). Водночас, потреби накопичення і використання інформаційних ресурсів, які створюються при рішенні наукових задач в грід-інфраструктурі (так званий «Data Grid»), пропонованими технологіями і програмними засобами задовольняються в значно меншому ступені.

Сучасні тенденції розвитку грід-технологій ставлять завдання створення єдиної інтегрованої інфраструктури (е-інфраструктури), яка об'єднала б обчислювальні і інформаційні можливості грід-систем в наукових дослідженнях [1]. Подібна е-інфраструктура призначена для накопичення в розподілених сховищах різноманітної інформації (структурованої і неструктурованої), простого, гнучкого і такого, що настроюється, доступу до неї різними категоріями дослідників з використанням єдиних уніфікованих механізмів, інструментів і ефективного використання цієї інформації у процесі складних обчислень. Парадигма віртуалізації ресурсів, що активно розвивається в інформаційних технологіях, представляє всі обчислювальні, програмні й інформаційні ресурси в гріду в формі віртуальних ресурсів. Сама ж подібна е-інфраструктура представляється у формі множини віртуальних дослідницьких середовищ (Virtual Research Environments, VREs) [2], які надають дослідникам у кожній науковій області всю сукупність потрібних ним ресурсів, використання яких проводиться за допомогою віртуалізованих сервісів.

У даній роботі досліджуються шляхи вирішення подібної проблеми створення віртуальних дослідницьких середовищ в е-інфраструктурі гріда у напрямі інформаційної підтримки наукових досліджень.

Е-Інфраструктура гріду в проекті D4Science

Проект D4Science [3, 4], що розробляється в Сьомій Рамковій Програмі Європейської комісії, направлений на створення інтегрованої електронної грід-інфраструктури для підвищення ефективності наукових і міждисциплінарних досліджень, обслуговування розширеної мережі наукових співтовариств в різних областях знань шляхом створення і використання спеціалізованих віртуальних дослідницьких середовищ для них (рис. 1). Основні віртуальні дослідницькі середовища, що функціонують в рамках проекту D4Science, наступні:

- **AquaMaps** об'єднує інформацію з біологічного моніторингу світового океану, що поступає з різних джерел в режимі реального часу і представляє її у формі безлічі комп'ютерних карт;
- **Fishery Country Profiles Production System (FCPPS)** описує різні аспекти рибного господарства, його ресурси, промисловість, виробничі і соціальні показники, правові аспекти й інше;
- **Vessel Transmitted Information (VTI)** містить промислову статистику з рибальства і рибних уловів у межах виняткових економічних зон і за їх межами в розрізах риболовецьких судів, простору і часу, інтегровану з екологічними даними;
- **Integrated Capture Information System (ICIS)** містить різні статистичні дані з рибальства в світовому океані;

- **DRIVER** інтегрує можливості взаємодії е-інфраструктури і D-Net систем у цілях використання ресурсів, розташованих зовні меж свого адміністративного домену і функціональних можливостей, об'єднує більше 2 мільйонів різних документів з 260 сховищ у 36 країнах;
- **INSPIRE** включає більше мільйона документів в області фізики високих енергій з підтримкою ефективних перехресних посилань і семантичних інформаційних зв'язків;
- **4D4Life (Distributed Dynamic Diversity Databases for Life)** містить так званий «Каталог життя»: класифікації і таблиці різновидів контрольних перевірок усесвітніх заводів, тварин, грибків і мікробів у цілях доступу до фундаментальної інформації про біологічну біосферу Землі та інші середовища.

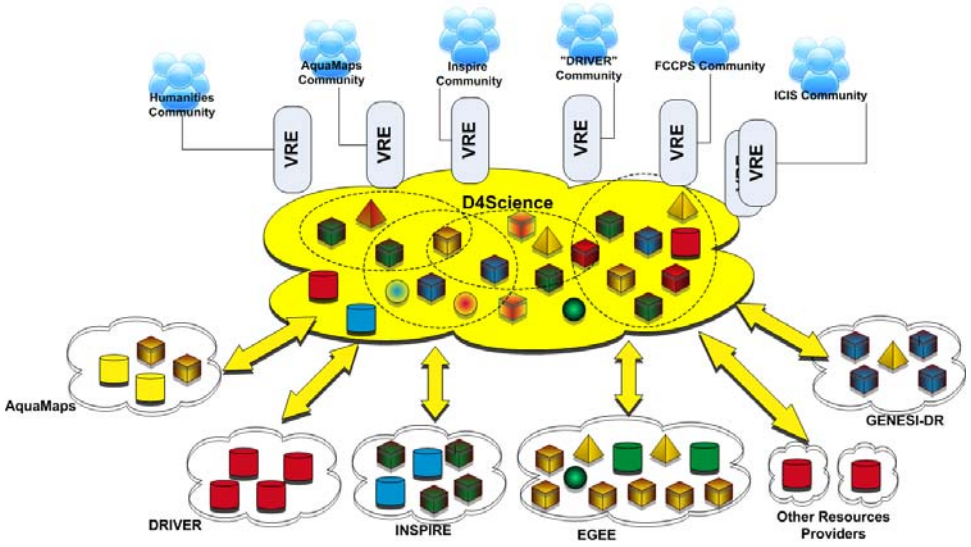

Рис. 1. Взаємодія віртуальних дослідницьких середовищ в е-інфраструктурі

Програмною основою проекту D4Science є система gCube, призначена для створення інтегрованої е-інфраструктури в грід-середовищі, яка характеризується наступним:

- різнорідні ресурси є фактично доступними в загальній інфраструктурі ресурсів, не дивлячись на їх розташування, технології і використовувані протоколи;
- різні співтовариства користувачів мають доступ до різних видів інформації згідно явно вираженим умовам, у відповідність з якими може відбуватися їх сумісне використання;
- кожне співтовариство користувачів може визначити своє власне віртуальне дослідницьке середовище на обмежений період і без додаткових витрат для постачальників ресурсу;
- декілька віртуальних дослідницьких середовищ може співіснувати без втручання одне в інше, навіть при умові конкуренції для таких же ресурсів.

Використовуючи віртуальне дослідницьке середовище, співтовариство користувачів має можливість управляти доступом до розподілених даних, сервісів, сховищ інформації і обчислювальних ресурсів, інтегрованих в єдину індивідуально настроєну інфраструктуру. Кожне віртуальне дослідницьке середовище підтримує загальні метадані, призначені для очищення, збагачення і трансформації інформації. Всі процеси виконуються на основі опису потоків робіт з використанням шаблонів перетворень, словників, тезаурусів і онтологій з широкими можливостями формування вітрин і публікацій даних, здійснення оцінки релевантності даних, створення користувачами нової інформації, що виробляється за допомогою обробки даних або різного моделювання.

Архітектура системи gCube

На даний час система gCube [5, 6] стала могутньою інфраструктурою, призначеною для забезпечення скоординованої роботи великої кількості програмних компонентів. Система gCube включає велику кількість стандартів специфікацій, технологій, моделей і технологій, об’єднаних в єдине ціле щоб гарантувати високу якість сервіс-орієнтованої інфраструктури (надійність, безпечність, автономність, пружність і дефекто захищеність) при побудові і підтримці розвитку віртуальних дослідницьких середовищ.

Система gCube має багатошарову архітектуру [7], показану на рис. 2.

Серцевину системи складають Сервіси інфраструктури виконання додатків (gCore Application Framework, gCF), призначені для реалізації абстракції технологій нижчих веб-сервісів (WSRF, WS Notification, WS Addressing й інші) і пропонують розвинуті можливості для управління станом, межами, подіями, безпекою, конфігуруванням, усуненням збоїв, публікаціями і доступом до інформації:

- управління всім життєвим циклом сервісів gCube, беручи участь у взаємодії з інфраструктурою і автономним навколишнім середовищем, і дозволяючи налаштування вимогам замовника, базуючись на змінах станів, зокрема, розгортання, ініціалізація, активація і невдача;

- забезпечення охоплення і правила безпеки, пов'язаних з ресурсами, що розділяються, управління придбанням та оновленням облікових даних служби, делегуючи повноваження та облікові дані у сфері розповсюдження з вхідних замовлень послуг в вихідні;
- реалізація стандартів WSRF для публікації, доступу та повідомлення про зміни у службових станах, пропонуючи багатий набір абстракцій для моделювання, прозорого управління змінами даних на інших ресурсах зберігання в процесі життєвого циклу, зокрема його оновлення від віддалених ресурсів зберігання в службових межах;
- стандартизація виправлення системних недоліків у службі інтерфейсів і реалізацій, виключення повторів того ж самого і повторів у випадках з еквівалентною семантикою недоліку, перетворення недоліків в еквівалентні полегшені виключення на службі кордонів;
- забезпечення посередництва при доступі до конфігураційних ресурсів на віддалених і локальних файлових системах, перенаправляючи невдачі читань до резервних копій, і створенні об'єктних закріплень для всіх аспектів обслуговування;
- спрощення відкриття ресурсу через об'єктні закріплення, шаблони, інспекції XPath для ряду запитів до інформаційних послуг інфраструктури;
- спрощення паралельного програмування через довільні комбінації, засновані на події синхронізації, планування, паралелізації і встановлення послідовності місцевих процесів.

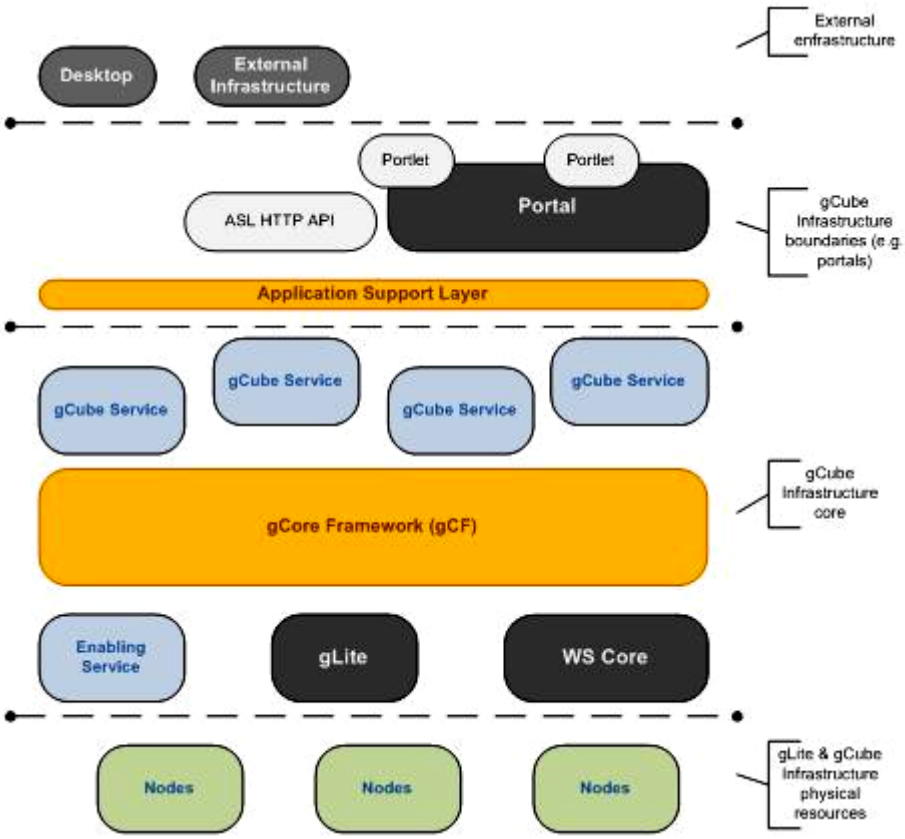

Рис. 2. Архітектура системи gCube

Іншою складовою частиною системи є Сервіси підтримки додатків (Application Support Layer, ASL), призначені для побудови локальних і мережевих додатків кінцевих користувачів, з використанням сучасних веб-технологій. Використання технології ASL не тільки ховає складність системи gCube від прикладного розробника, але і створює нові можливості для сумісності gCube і зовнішніх інфраструктур, завдяки використанню протоколів WSRF, як єдиного засобу доступу до сервісів gCube. Призначенням сервісів є:

- погоджувальне зберігання, пошук і перетворення інформаційних ресурсів;
- публікація інформаційного ресурсу, його відкриття і доступ;
- розповсюдження і пошук інформації;
- безпека.

Система gCube спроектована і створена відповідно до принципів компонентно-орієнтованого програмного забезпечення, має повністю сервіс-орієнтовану архітектуру. В результаті, ряд підсистем логічно згруповані в таких доменах.

1. Сервіси оперування інфраструктурою і управління VRE (Infrastructure Operation and VRE Management) забезпечують: (i) організацію і використання віртуальних дослідницьких середовищ з гарантією оптимального споживання доступних ресурсів; (ii) реєстрацію складових інфраструктури; (iii) аутентифікацію і авторизацію

політик, що дозволяє жорстко контрольований обмін між компонентами інфраструктури; (iv) визначення і оркестровку складних процесів, гарантуючи оптимальне споживання доступних ресурсів. Сервіси інфраструктури вирішують задачі визначення VO/VRE і забезпечують динамічне розгортання наступних ресурсів через інфраструктуру*:*

- сховище програмного забезпечення, обслуговування сервісів (встановлення і оновлення,) і менеджер управління вузлами gCube;
- фреймворк сервісів Інформаційної системи;
- управління VO, Ідентифікація і сервіси Уповноважень.

2. Сервіси організації інформації (Information Organisation) забезпечують: (i) зберігання і організацію різноманітних інформаційних об'єктів (складені об'єкти, що охоплюють частини багаторазового використання і альтернативні уявлення) і їх компонентів, що включають керування статистичними даними і об'єктами метаданих подібно до *Time Series* (дані, прив'язані до часових інтервалів); (ii) керування об'єктами метаданих в багаторазових форматах, що обкладають кожен інформаційний об'єкт; (iii) керування об'єктами анотацій, що потенційно збагачують усі інформаційні об'єкти:

- сервіси керування контентом і зберіганням;
- сервіси керування метаданими та індексатор XML;
- система анотації (внутрішній і зовнішній інтерфейси);
- сервіси керування онтологіями;
- сервіси трансформації даних;
- сервіси керування Time Series.

3. Сервіси пошуку інформації (Information Retrieval) забезпечують багатий набір засобів пошуку інформації, що дозволяють здійснювати пошук даних і інформації широким рядом методів і включають: (i) пошуковий фреймворк, що компонується з компонентів оркестрування методів пошуку, набору пошукових операторів, компонентів процесора запиту і механізму трансформації даних; (ii) фреймворк індексного управління, що підтримує всі аспекти життєвого циклу індексів, і забезпечує пошукові здібності для багатого набору індексів, як, наприклад, повнотекстові індекси, опереджаючи індекси, гео-індекси індекси; (iii) розподілений фреймворк, що забезпечує виконання пошукових операцій вищого рівня, які включають погоджувальне ранжирування контенту, вибір джерел інформації, об'єднання результатних наборів інформації (класифіковане злиття різних наборів даних):

- пошукові сервіси і пошукові оператори;
- сервіси індексування (Forward/FullText/GeoSpatial);
- опис джерел даних і їх вибору;
- служба персоналізації.

4. Сервіси уявлень (Presentation Services) реалізують завдання найвищого шару gCube-посиленої інфраструктури з подвійним завданням: (i) забезпечити засобами для побудови інтерфейсів користувача для забезпечення взаємодії з системою і інфраструктурою; (ii) забезпечити повний набір інтерфейсів користувача для досягнення взаємодії з системою. Вищезазначений Шар Прикладної Підтримки є засобом для забезпечення простоти експлуатації інфраструктури gCube, тоді як стандартний портальний двигун призначений для реалізації рідних інтерфейсів користувача.

Оперування інформацією і її інтеграція в системі gCube здійснюється за допомогою колекцій даних. Колекція даних є елементарним об'єктом, що описує набір інформаційних об'єктів зовнішнього джерела інформації.

Інформаційний об'єкт (в подальшому – документ) може бути комплексним**,** мультимедійним чи багатопрофільним. Інформаційним об'єктом може бути:

- файл у будь-якому форматі;
- мультимедійний об'єкт (аудіо чи відео);
- файл у форматі HTML чи XML (сторінка сайту);
- геофізичні та картографічні дані;
- зріз бази даних.

Набір документів описується з використанням документарної моделі системи gCube. В моделі кожен документ може бути складеним з декількох складових частин, кожна з яких виступає документом. Кожна складова частина документу може мати декілька альтернативних уявлень.

Колекція дозволяє маніпулювати наборами комплексних документів. Кожна колекція включає кількість її документів, а кожен документ може виступати лише складовою частиною даної колекції. Колекція в свою чергу може бути складовою частиною іншої колекції.

Колекція є базовою структурою організації інформації у сервісах керування інформацією.

Побудова е-інфраструктури в Українському національному гріду

Побудова сучасної е-інфраструктури наукових досліджень в українському грід-середовищі для забезпечення інтеграції інформаційних ресурсів різних тематичних областей з можливостями підтримки міждисциплінарних зв'язків виконувалась у рамках відповідного проекту у Державної цільової науково-технічної програми впровадження і застосування грід-технологій на 2009–2013 роки.

На даний час були отримані наступні основні результати [8]:

- розроблений типовий web-портал науково-дослідницького проекту в грід-інфраструктурі, що інтегрує виконання розрахунків, накопичення і аналіз отриманих результатів;
- розроблений опис інформації, що використовується в проекті, відповідно до її структури, місць знаходження на вузлах грід-середовища для цілей анотування, індексування, пошуку і представлення;
- розроблені згідно концепції WorkFlow описи процесів наукового дослідження, що виконується в рамках проекту, з накопичення, використання і аналізу інформації;
- розроблена система моніторингу створеного віртуального дослідницького середовища.

Система gCube, як базове програмне середовище е-інфраструктури, розміщена на кластері Інституту програмних систем НАН України. Архітектура розміщення системи на кластері показана на рис. 3. Для розміщення системи gCube використано два обчислювальних сервера, на яких розміщено 12 віртуальних машин. Всього було інстальовано 235 сервісів і допоміжних програмних підсистем, згрупованих у 24 класи.

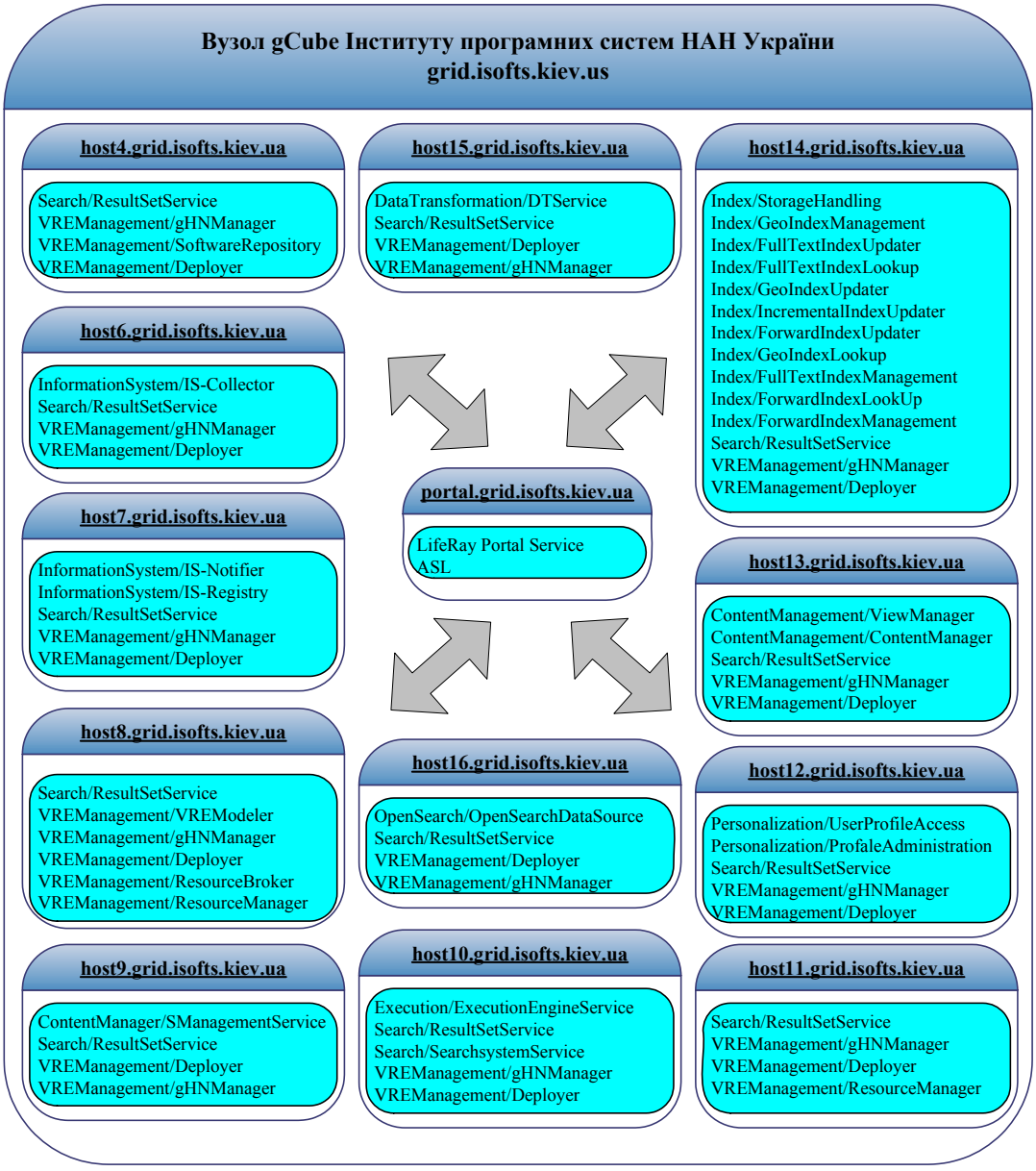

Рис. 3. Архітектура системи на кластері ІПС НАН України

Створення колекцій даних. Як приклад колекції були використані бібліотечні сховища Інституту програмних систем НАН України і Житомирського державного університету ім. Івана Франка, що загалом складають більше п'яти тисяч документів. Структура документів відповідає Dublin Core. Інтерфейс сховищ підтримує протокол OAI-PMH 2.0, що є одним з більш ніж 15 протоколів, що підтримуються системою gCube.

Процедура створення даної колекції виконується з допомогою портлету ResourceManager (рис. 5) віртуальної організації, до якої буде належати VRE з даною колекцією і включає наступні етапи:

- створення загального ресурсу опису колекції;

- створення запису активації колекції з описом інтерфейсу доступу до інформаційного ресурсу, його структури і форматів даних;
- створення запису активації уявлення документів колекції з описом процедури трансформації даних з зовнішнього репозиторію (джерела) в структуру системи, структури уявлення документів при їх публікації на вітрині порталу;
- формування процедури побудови індексів для інформаційного наповнення колекції візуально виконується з допомогою портлету IRBootstrapperPortlet (рис. 4) і включає опис процедур для створення набору індексів;
- побудова кожного індексу виконується завданням параметрів відповідних процедур (у нашому випадку – в системі є велика кількість процедур підтримки різноманітних форматів даних, формат, що використовується в бібліотечному репозиторії – OAI_DC, включений в базову систему, у випадку використання нового формату має бути написаний набір відповідних процедур на мові ASL);
- опис полів колекції виконується за допомогою портлету SearchManager. Для кожного поля вказуються параметри пошуку інформації і її представлення.

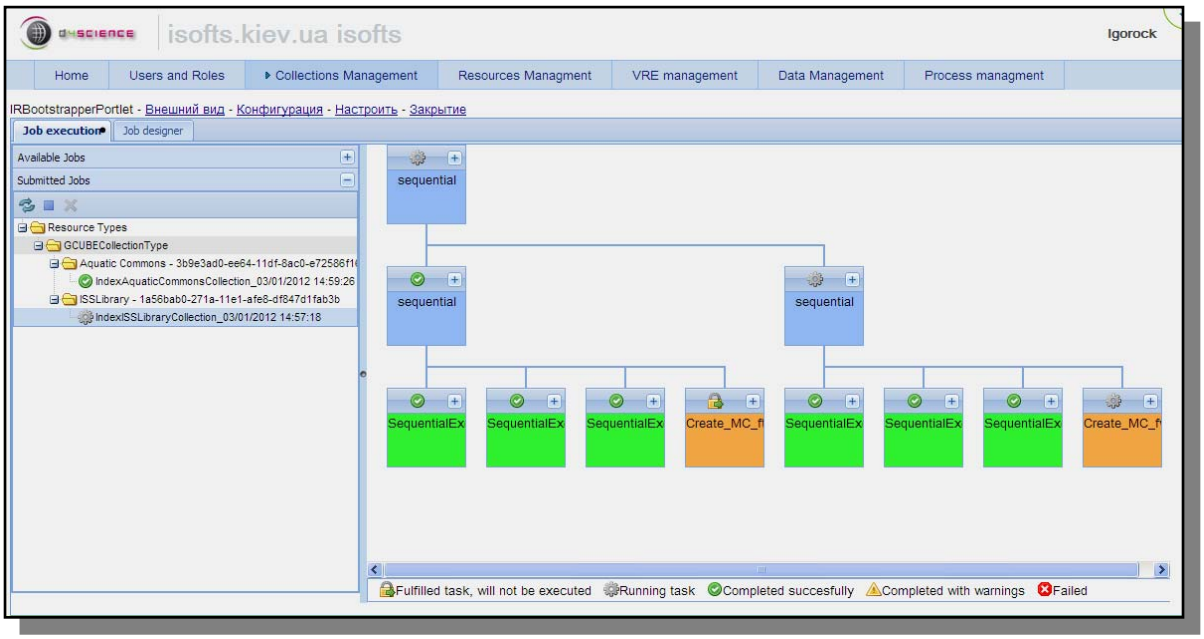

Рис. 4. Візуальний вигляд процедури побудови індексів

Головне призначення колекцій даних полягає в можливостях пошуку інформації з використанням розвинутих пошукових запитів на основі опису метаданих, отримання необхідних результатів за інформацією, що розміщена в розподілених сховищах даних в грід-інфраструктурі, перегляду інформації.

У системі gCube інтерфейсні функції пошуку інформації реалізовані у системі пошукових портлетів. При формуванні пошукового запиту вказується:

- перелік колекцій, в яких здійснюється пошук;
- критерії пошуку, що задаються у вигляді контекстного запиту за полями документів чи документу в цілому.

В результаті формується перелік документів, що задовольняють критеріям пошукового запиту. Для кожного знайденого документу є можливість переглянути:

- адресу знаходження документа;
- формат зберігання;
- анотацію;
- зміст документа відповідно до формату зберігання.

Документообіг в віртуальному дослідницькому середовищі. Засоби управління документообігом в системі gCube призначені для реалізації процесів підготовки документів для публікації у віртуальних дослідницьких середовищах з використанням технології WorkFlow.

Першим етапом опису документообігу є проектування його схеми. На рис. 5 показано зразок подібної схеми для підготовки документу для включення в колекцію даних.

У схемі вказуються:

- етапи обробки документа, наприклад, первинна публікація, обговорення, виправлення й інше;
- ролі користувачів, що виконують кожний етап обробки документів.

Другим етапом опису документообігу є розробка шаблону звітів, що мають оброблятися за певною схемою WorkFlow.

Використання документообігу здійснюється наступним чином.

1. На основі розроблених схеми документообігу і шаблону документа створюється конкретний звіт – документ, що має бути створений і оброблений.

2. Для даного документу для всіх ролей вказуються користувачі, що мають виконувати операції, що вказані на схемі. В порталі користувач має перелік документів, що потребують його втручання в процес документообігу. Для кожного документа користувач має можливість відповідно до його повноважень:

- переглянути документ;
- виправити документ;
- створити коментарій до документа чи його складовій частині;
- виконати поточну операцію обробки документа відповідно до своєї ролі;
- виконати експорт документа в формат DOC для обробки в текстовому редакторі чи HTML для публікації на порталі.

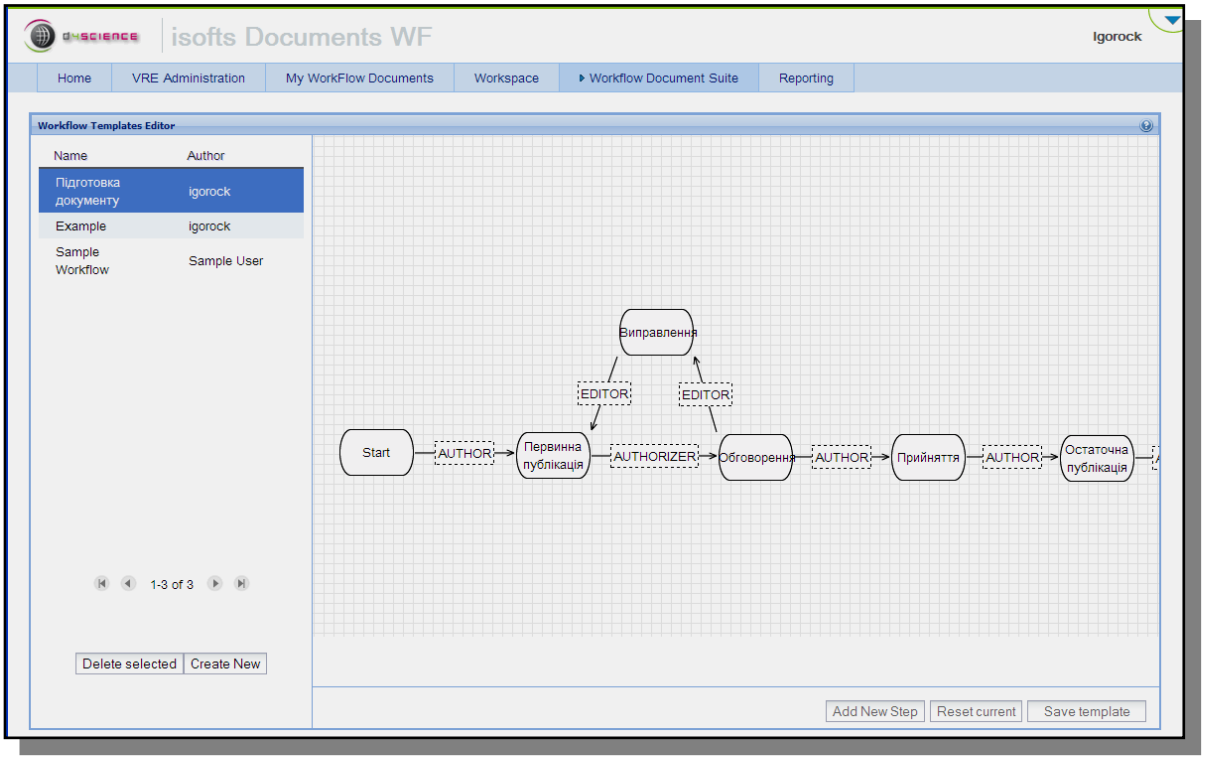

Рис. 5. Схема WorkFlow формування документа

Висновки

У роботі розглянуті практичні питання побудови інтегрованої е-інфраструктури в грід-середовищі, призначеної для об'єднання обчислювальних і інформаційних можливостей грід-систем у наукових дослідженнях.

Описана архітектура системи gCube, яка є програмною основою європейської е-інфраструктури D4Science, що інтегрує різноманітні інформаційні ресурси різних галузей науки і економіки.

Висвітлені основні результати, отримані при побудові е-інфраструктури в Українському національному гріду.

1. *Andrade P., Pagano P., Manzi A.* e-Infrastructures Integration with gCube: https://www.egi.eu/indico/contributionDisplay.py?contribId=40&sessionId=18&confId=207
2. *Blankel T., Candela L., Hedges M., Priddy M., Simeoni F.* Deploying general-purpose virtual research environments for humanities research: http://rsta.royalsocietypublishing.org/content/368/1925/3813.abstract
3. *D4SCIENCE-II* :Data infrastructure ecosystem for science: http://www.ist-world.org/ProjectDetails.aspx?ProjectId=a78b23ef5c0040c5b00a3c2344fd2e28&SourceDatabaseId=018774364ea94468b3f4dec24aa1ee53
4. *www*.d4science.eu/
5. *www*.gcube-system.org
6. *https*://gcube.wiki.gcube-system.org
7. *Candela L., Castelli D., Pagano P.* gCube: A Service-Oriented Application Framework on the Grid: http://ercim-news.ercim.eu/gcube-a-service-oriented-application-framework-on-the-grid
8. *http:*//portal.grid.isofts.kiev.ua:8080