

Создание прототипа системы управления информационными ресурсами*

Ю. Леонова, А. Федотов

Институт вычислительных технологий Сибирского отделения РАН,
Новосибирск, Россия
{Juli,Fedotov}@ict.nsc.ru

Аннотация В данном докладе дается определение научной информационной системы (НИС) с точки зрения научной коммуникации. Определены задача, предметная область, субъекты, объекты, основная функциональность НИС, приводится список основных видов информационных ресурсов. В работе выполнен анализ функциональных требований предъявляемых к таким системам.

Ключевые слова: научные информационные системы, функциональные требования, информационный ресурс, информационная модель

Введение

Особенность развития современного общества характеризуется возрастающим объемом и быстрым устареванием научной информации. Для повышения эффективности научных исследований ученым необходим доступ к информации о результатах исследований, выполняемых в интересующей области. Поэтому любое научное исследование обычно начинается с поиска научной информации об исследованиях в данной области, но поиск необходимой информации в постоянно возрастающем объеме статей, книг, монографий, отчетов, патентов становится все сложнее. Ученым приходится много времени тратить на поиск и обработку информации, позволяющей быстро ознакомиться с результатами других исследований и исключить их дублирование.

Таким образом, в современном научном сообществе большое значение имеют темп и качество научно-информационного обмена. Основой коммуникации в научном сообществе является научная публикация. «Публикация выступает как первичный источник сведений о научном знании, отношениях между учеными, строения и динамике научных объединений и т.п. Для науковеда, философа, логика, методолога, специалиста по информатике, социолога науки той конечной реальностью, из которой исследователь черпает свои представления о науке, выступают публикации. Отличающиеся друг от друга изображения науки в различных исследовательских традициях ...

* Работа выполнялась при финансовой поддержке РФФИ (Проект № 18-07-01457).

становятся объектами изучения лишь постольку, поскольку сведения о них имеются в научной публикации» [1]. Главной задачей создания информационной системы поддержки научных исследований является ускорение темпа и повышение качества информационного обмена в научном сообществе.

Работы по унификации данных и систематизации информационных ресурсов являются актуальными для многих профессиональных областей. Информационные системы поддержки научных исследований оперируют с различного рода информацией. Это могут быть публикации, электронные документы, электронные коллекции, онтологические описания, массивы данных, логические описания и др. Эти ресурсы, востребованные разными группами исследователей, могут оказаться недоступными из-за проблем с их поиском и идентификацией. Семантические связи между информационными ресурсами повышают их ценность и предоставляют дополнительные возможности для информационного поиска и идентификации. Сложность использования ресурсов обусловлена проблемами обеспечения актуальности версий ресурсов, а также отсутствие единых механизмов преобразования при разнородности представления и хранения ресурсов (электронные таблицы, базы данных, электронные документы, каталоги, файлы – бинарные, текстовые, XML-разметка и т.п.).

Интеграция информационных ресурсов в единую информационную среду и организация доступа к информационным ресурсам позволяет унифицировать процесс обмена результатами научных исследований и повышает эффективность взаимодействия отдельных групп исследователей.

Актуальной задачей является создание модели информационной системы поддержки научных исследований позволяющей:

- унифицировать процесс обмена результатами научных исследований;
- оперировать с данными и документами, интегрированными в открытое семантическое пространство;
- предоставлять сервисы преобразования разнородных ресурсов, реализующих средства описания, представления, автоматического связывания ресурсов, а также взаимодействие с поисковыми и классификационными механизмами в соответствии с потребностями пользователей.

Модель должна обеспечить следующую функциональность:

- публикация ресурсов, включающая процедуры регистрации, именования, аннотирования и определение формата;
- аналитическая обработка ресурсов;
- доступ к опубликованным ресурсам, включая функции динамического формирования;
- для автоматизированного функционирования необходимы функция мониторинга ресурсов и актуализации их метаописаний, функции уведомления пользователей о появлении новых ресурсов и обновлении существующих, функция диспетчеризации.

Предметная область научной информационной системы

Научная информационная система (НИС) – система, которая автоматизирует процессы научного информационного обмена. В данной статье мы рассматриваем НИС с позиции информационного обмена. В англоязычной литературе существует альтернативное определение данного термина как CRIS (Current Research Information System)[13]. Информационными ресурсами НИС являются научные публикации. Научная публикация формально является основным результатом работы ученого. Публикуя материал, ученый знакомит научную общественность с результатами своих исследований, их анализом и выводами. Но помимо донесения информации о проведенной работе, у публикации в НИС есть еще одна полезная функция – обеспечение поиска публикаций по аналогии, что существенно облегчает поиск научной информации об исследованиях в данной области.

По объему и жанру публикации результатов научных исследований можно разделить на пять основных типов:

1. краткие сообщения, включая тезисы докладов на конференциях, симпозиумах, совещаниях и т.п. ;
2. научные статьи – основная форма полноценного представления результатов научных исследований;
3. монографии – наиболее развернутое изложение результатов исследования какой-либо научной проблемы;
4. публикации на правах рукописей, например, диссертации и их авторефераты;
5. отчеты о научной работе, которые могут быть зарегистрированы во ВНИТИЦ (Всероссийский научно-технический информационный центр).

Наиболее важной из них считается статья, которая может быть опубликована в специализированном периодическом или непериодическом издании. Другие формы научных публикаций либо отличаются большим объемом и поэтому печатаются редко (например, монографии), либо недостаточно информативны, как тезисы докладов на конференции.

Обозначим основные категории пользователей, с которыми взаимодействует НИС:

1. научные работники (ученые, исследователи, аналитики);
2. работники образования и учащиеся (преподаватели, учителя, аспиранты, студенты);
3. управленческий персонал (руководители организаций, отделов, лабораторий, проектов);
4. инвесторы (финансирующие фонды и организации, физические лица);
5. прочие (журналисты, статистики, обозреватели).

Категории пользователей и их потребности определяют основные виды сервисов, которые предоставляют НИС: поиск и распространение научной информации, службы уведомления, архивное хранение информации, аналитические службы.

Требования, предъявляемые к информационным системам поддержки научных исследований

Требования можно разделить на следующие виды:

- пользовательские требования;
- функциональные требования;
- модели и словари данных;
- требования к производительности;
- требования к удобству использования;
- требования к безопасности.

В работе [13] представлен набор наиболее общих функциональных требований к НИС.

1. Сбор информационных ресурсов. Для сбора информации необходимо использовать различные варианты ввода данных:
 - ввод данных пользователями;
 - сбор данных в Internet посредством специальных программных агентов (“пауков”);
 - обмен данными с другими ИС.
2. Релевантность документов. При автоматическом сборе информации в Internet может накапливаться и нерелевантная или малорелевантная информация для данной НИС. Решение проблемы возможно следующими способами.
 - Создание подробных форматов представления метаданных о ресурсах и структурированных справочников для тематической классификации ресурсов. НИС должна вкладывать описания в метаданные на веб-страницы и предоставлять интерактивные средства пользователям для создания метаданных определенного формата при размещении ресурсов.
 - Разделение информационных ресурсов в зависимости от варианта поступления в систему (размещенные экспертами/пользователями и “пауком”, а также указание степени достоверности информации с учетом ее источника.
 - Указание поисковым средствам пространства поиска и классификации информации, а также задание критериев оценки качества введенной информации
 - Использование схем классификации ресурсов согласно потребностям пользователей и классификация ресурсов в соответствии с этими схемами
3. Актуальность, полнота, достоверность происхождения документов. Способы решения проблем актуальности и полноты аналогичны способам решения проблемы охвата ресурсов. Способы определения достоверности происхождения информации следующие:
 - для интерактивного ввода – информация вводится только аутентифицированными пользователями;

- для автоматизированных систем сбора – накладывание ограничений на область действия агента, выполняющего сбор информации;
 - для обмена информацией с другими ИС – задание фильтров на импортируемые информационные ресурсы;
 - для всех способов ввода – должна выполняться проверка и классификация всей введенной информации.
4. Использование интеллектуальных служб обработки запросов пользователя. Службы обработки запросов пользователей должны обеспечивать поиск по атрибутам, полнотекстовый поиск, просмотр ресурсов по категориям, семантический поиск (необязателен).
 5. Извлечение знаний. Использование частичной автоматизации извлечения знаний. В основе подхода лежит представление смысла текста в виде семантической сети, принцип построения которой основывается на использовании частоты совместной встречаемости понятий в тексте. Пользователю сеть представляется в виде тематического дерева (дерево ключевых терминов и связанных с ними понятий), что позволяет выполнять навигацию и существенно облегчает процесс исследования текста и поиска требуемой информации. Данный подход также используется для решения таких задач, как автоматическое реферирование, тематическая классификация и кластеризация текстов, семантический поиск и т.д. Кроме того к НИС, работающие с разными типами информационных ресурсов, предъявляются следующие требования
 6. Поддержка не централизованных архитектур информационных систем. Это требование является необходимым условием для полноты, аутентичности и актуальности информации. Опыт эксплуатации НИС показал сложность создания централизованных научных систем, охватывающих научную информацию в какой-то области науки, или в какой-то стране.
 7. Структурированность информационного пространства. Для поддержки сложных функций поиска и классификации информации помимо хранения полнотекстового описания, необходимо реализовывать поиск по атрибутам, полнотекстовый поиск, просмотр ресурсов по категориям и словарям-классификаторам. При этом выбор классификаторов определяется степенью специализации системы.
 8. Адаптивное представления информации. Для повышения скорости поиска и точности подбора информации пользователем без потери качества поиска НИС должна учитывать запросы пользователей, их компетентность при работе с НИС, ограничения по времени. НИС должна обеспечивать возможность пользователю получать различные уровни абстракции при представлении информации от кратких описаний для максимального быстрого поиска, до очень подробных описаний информационных объектов.
 9. Историчность информации. Спецификой научной информации является ее быстрое устаревание и потеря актуальности. Для многих типов информационных ресурсов важно хранить всю информацию о всех изменениях и иметь возможность восстановить состояние ресурса на любой момент времени. Например, информация об авторах может меняться со

временем при смене персоной фамилии, места работы. Также необходимо учитывать переформирования и переименования организаций, наименования географических объектов также могут меняться. Поэтому необходимо учитывать временной фактор и использовать актуальную информацию для сущностей, связанную с промежутками времени. При распознавании сущностей необходимо обеспечивать выполнение запросов на какой-нибудь момент времени в прошлом, то есть создание среза истинности информации о сущностях на произвольную дату.

10. Архив. Как было отмечено выше, большая часть научной информации быстро устаревает. Но существуют информационные ресурсы, к которым необходимо обеспечивать доступ длительное время. К таковым, например, относятся документы, имеющие длительную юридическую силу, патенты или мультимедийная информация об исторических событиях, которая может быть востребована через любой период времени. Кроме того, научные отчеты институтов, речи ученых могут также иметь огромную историческую ценность, становясь только еще ценнее со временем. Поэтому системы должна поддерживать возможность длительного хранения информационных ресурсов с возможностью восстановления их.

В условиях работы в распределенной среде к НИС предъявляются требования:

- поддержка принятых стандартов метаданных для экспорта и импорта данных;
- поддержка протоколов обмена информации с другими информационными системами;
- поддержка возможности ссылки на внутренние ресурсы как в интерфейсах пользователей, так и на системном уровне.

Основные подсистемы к НИС и подходы к их реализации

Рассмотрим технологию построения прототипа информационной системы поддержки научной деятельности в соответствии с приведенными выше требованиями.

К основным функциям информационных систем относятся функции сбора и регистрации информационных ресурсов, их сохранение, обработка, актуализация, обработка запросов пользователей [14].

Сбор и регистрация информационных ресурсов. Сбор и регистрация информационных ресурсов могут выполняться одновременно или последовательно. Возможны различные варианты создания электронных ресурсов:

- ввод данных в систему вручную, например, путем измерения или наблюдения фактов реального мира;

- полуавтоматический – ввод с носителей информации (тексты на бумажных носителях);
- автоматический – ввод с различных датчиков, обмен данными с другими информационными системами.

В рассматриваемой авторами НИС ввод данных осуществляется полуавтоматическим и автоматическим способами, информационными ресурсами являются бумажные или электронные публикации.

При реализации этих функций возникает необходимо решать задачи очистки, верификации, сжатия данных, преобразования данных из одного формата в другой.

Функционирование среды основывается на использовании протоколов Z39.50 и LDAP. При этом предусмотрены механизмы преобразования данных из предметных схем в абстрактную схему протокола Z39.50. Виртуальная среда состоит реестра объектов и ресурсов, основного сервера Z39.50, нескольких функциональных модулей, а также web-интерфейса с публичным и административным разделами для доступа к различным функциям среды. Для каждого источника устанавливается отдельный сервер Z39.50, который осуществляет преобразование данных из схемы источника в абстрактную схему данных.

Хранение информационных ресурсов. Система должна обеспечивать управление и хранение распределенных данных в базах данных, поддержку различной структуры хранимых данных, их целостности и эффективного доступа к ним. Обработка информационных ресурсов. Обработка информационных ресурсов в информационных системах позволяет представлять пользователям продуцированную информацию, являющейся производной данных ранее введенных в систему и хранящихся в базе данных. Обработка также осуществляется при выполнении некоторых системных функций – при проверке ограничений целостности данных, при поиске в словарях и индексах.

Примером информационных систем без обработки информационных ресурсов являются простые системы текстового поиска, выдающие пользователям документы, удовлетворяющие условиям запроса, без какой-либо трансформации. Примером информационных систем, которых обеспечивают продуцирование производных данных, являются системы, выполняющие автоматическую генерацию рефератов и аннотаций хранимых документов. Одним из возможных вариантов эффективного доступа к информационным ресурсам является классификация документов. Проблема заключается в сложности ориентирования в информационных массивах, что не позволяет получить наиболее актуальную и полную информацию по определенной теме. При использовании тематической классификации для систематизации ресурсов необходимо определить состав логико-семантических категорий (фасетов) и ключевых терминов (понятий), покрывающих избранную достаточно узкую предметную область, интересующую пользователя.

Актуализация информационных ресурсов. Поддержка информационных ресурсов в ИС позволяет моделировать состояние и поведение

предметной области. Для поддержки динамической модели ИС необходимо актуализировать информационную модель – своевременно отслеживать в ней изменения состояния предметной области. Для этой цели необходимо актуализировать информационные ресурсы НИС. Актуализация информационных ресурсов осуществляется путем ввода или удаления документов, модификации ссылок, связывающих документы. При изменении структуры предметной области ИС для актуализации информационных ресурсов выполняется изменение схемы базы данных.

Предоставление информационных ресурсов пользователям. Основной целью создания ИС является удовлетворение информационных потребностей пользователей посредством предоставления необходимой им информации на основе хранимых информационных ресурсов. Предоставление ресурсов может осуществляться с помощью pull- и push-технологий:

- Pull-технология – предоставление ресурсов по инициативе пользователя через пользовательские интерфейсы, реализуется в ИС при использовании механизмов поиска и навигации.
- Push-технология – инициатором предоставления является сама система, в соответствии с определенным регламентом и для определенного круга пользователей. Применяется для распространения различного рода информации среди пользователей, стандартное сообщение рассылается по списку рассылки всем зарегистрированным в нем пользователям. В НИС таким образом организуется информирование пользователей о поступлении новых документов.

Помимо перечисленных основных функций ИС, видимых пользователям, существуют дополнительные функции, часть из которых возлагается на персонал ИС и ПО:

- управление распределенными информационными ресурсами, например фрагментация баз данных, тиражирование данных, синхронизация копий;
- защита физической целостности информационных ресурсов и их восстановление при разрушениях;
- обеспечение информационной безопасности в системе;
- управление метаданными;
- администрирование информационными ресурсами;
- обеспечение адаптации системы к изменениям требований к ней и к изменениям в предметной области.

Список литературы

1. *Мирский Э. М.* Массив публикаций и система научной дисциплины. «Системные исследования», 1977 г.
2. *Вигурский К. В.* Что такое электронная библиотека? // Доклад на конференции «Информационные технологии в образовании - 2005». — электронный ресурс <http://rd.feb-web.ru/library.htm> (дата обращения: 28.01.2013)

3. ГОСТ Р ИСО/МЭК 12207-99. Информационная технология. Процессы жизненного цикла программных средств. ГОССТАНДАРТ РОССИИ. Москва, 1999.
4. ГОСТ 34.003-90. Информационная технология. Комплекс стандартов и руководящих документов на автоматизированные системы. Термины и определения.
5. Жижимов О. Л., Мазов Н. А., Федотов А. М. Некоторые заметки об эволюции цифровых репозитариев традиционных библиотек к полнофункциональным электронным библиотекам // Вестник Владивостокского государственного университета экономики и сервиса. Территория новых возможностей. — № 3 (7). — 2010. — С. 55–63.
6. Козаловский М. Р. Метаданные, их свойства, функции, классификация и средства представления // Труды 14-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL2012, Переславль-Залесский, Россия, 15-18 октября 2012 г.
7. Лукашевич Н. В. Тезаурусы в задачах информационного поиска / Москва: Издательство МГУ. — 2011. — С. 512. — ISBN 978-5-211-05926-9.
8. Ляпунов А. А. О соотношении понятий материя, энергия и информация // В кн.: Ляпунов А. А. Проблемы теоретической и прикладной кибернетики. — Новосибирск: Наука, 1980. — С. 320–323.
9. Отле П. Библиотека, библиография, документация: Избранные труды пионера информатики / Поль Отле. Пер. с англ. и фр. — Москва: ФАИР-ПРЕСС: Пашков Дом, 2004. — 348, [1] с. — (Специальный издательский проект для библиотек). — Библиогр.: с. 312–327. — Имен. указ.: с. 340–342. — ISBN 5-8183-0624-0.
10. Резниченко В. А., Проскудина Г. Ю., Кудим К. А. Концептуальная модель электронной библиотеки [Электронный текст] / В. А. Резниченко, Г. Ю. Проскудина, К. А. Кудим // Труды XI Всероссийской научной конференции RCDL'2009», Россия, г. Петрозаводск (Карелия), 17–21 сентября 2009 г. — С. 23–31.
11. Советов Б.Я. Информационные технологии : учебник для вузов / Б.Я. Советов, В.В. Цехановский. — М.: Высшая школа, 2005.
12. Федотов А. М. Методологии построения распределенных систем // Вычислительные технологии. 2006, Т. 11, Избранные доклады X Российской конференции «Распределенные информационно-вычислительные ресурсы» (DICR-2005), Новосибирск 6-8 октября 2005 г. С. 3–16.
13. Кулагин М. В., Лопатенко А. С. Научные информационные системы и электронные библиотеки. Потребность в интеграции. // Сб. трудов Третьей Всероссийской конф. по электронным библиотекам. RCDL'2001 Петрозаводск, 11–13 сентября 2001 г., с. 14-19.
14. Лекции по Автоматизированным информационным системам и СУБД. Лекция 2 - функции информационных систем. «Бауманки.НЕТ» [Электронный ресурс]. Режим доступа: <http://baumanki.net/lectures/10-informatika-i-programmirovanie/320-lekcii-po-avtomatizirovannym-informacionnym-sistemam-i-subd/4274-2-funkcii-informacionnyh-sistem.html>
15. Федотов А. М., Шожкин Ю. И. Электронная библиотека Сибирского отделения РАН // Информационное общество. — 2000. — № 2. — С.22–31.
16. Шожкин Ю. И., Федотов А. М., Жижимов О. Л., Гуськов А. Е., Столяров С. В. Электронные библиотеки — путь интеграции информационных ресурсов Сибирского отделения РАН // Вестник КазНУ, специальный выпуск. — г. Алматы, Казахстан, Казахский национальный университет им. аль-Фараби. — 2005. — № 2. — С. 115–127.

17. Шожин Ю. И., Федотов А. М., Жижимов О. Л., Федотова О. А. Эволюция информационных систем: от Web-сайтов до систем управления информационными ресурсами // Вестник НГУ Серия: Информационные технологии. 2015. Том 13, Выпуск № 1. С. 117–134. — ISSN 1818-7900.
18. D3.2b The Digital Library Reference Model // Funded under the Seventh Framework Programme, ICT Programme — “Cultural Heritage and Technology Enhanced Learning” Project Number: 231551. April 2011. (<http://www.dlorg.eu>) — Full Text: http://db4.sbras.ru/elbib/data/show_page.phtml?22+269.
19. DCMI — Dublin Core Metadata Initiative // (<http://www.dublincore.org/>).
20. DSpace [Электронный ресурс]: an open source solution for accessing, managing and preserving scholarly works // dspace.org [web-сайт] / MIT Libraries; HP Labs. — 2007. (<http://www.dspace.org/>).
21. ISO 14721:2012 Reference model for an Open archival information system (RM OAIS) / Recommended Practicle: CCSDS 650.0-M-2 (Magenta Book). June 2012. Available at: <http://public.ccsds.org/publications/archive/650x0m2.pdf>
22. Mooers C.N. «Mooers» law, or why some retrieval systems are used and other are not // American Documentation. - 1960. - Vol.11, N.3.
23. Otlet P. Traite de documentation. // Bruxelles: Ed. Mundaneum, 1934.