Представление и хранение научных публикаций в веб-системах

И. А. Тарханов

Аннотация. В век повсеместного распространения Интернета любопытно проследить, как наука и современные технологии влияют на веб-системы, оперирующие научными публикациями. В таких системах основной формой хранения и представления публикаций является файл. В данной статье предлагается более естественная форма хранения и представления научных знаний, рассматривается проблематика применения данного подхода и перспективы развития веб-систем научных публикаций.

Ключевые слова: *научная публикация, веб-система, электронная библиотека, издательство, граф, PII, URI, ГОСТ, Semantic Web, e-learning, SCORM.*

1. Современные веб-системы научных публикаций

В статье речь пойдет о системах, работающих через Интернет, и оперирующих научными публикациями. Это электронные библиотеки, научные журналы и сайты издательств. На сегодняшний момент крупнейшими мировыми издателями электронных научных публикаций являются Springer, Wiley, Elsevier [1–3]. В России самая крупная веб-система, агрегирующая научные публикации, это электронная библиотека eLibrary [4]. Рассмотрим, как оперируют научными знаниями известные электронные издательства и библиотеки.

В большинстве известных систем основная форма публикаций — это файл, как правило, в формате pdf. В систему вводятся файлы, подготовленные авторами, либо обработанные редакторами. Далее они индексируются и хранятся в БД. При просмотре используется выгрузка того же файла пользователем или показ его содержимого в браузере. Тем самым, файл является основной формой хранения и распространения публикаций. Такой подход мало чем отличается от переноса и хранения текстов в бумажном виде.

Преимущества такого подхода в простоте его реализации. Не нужно разбирать структуру публикаций, писать модули экспорта и импорта данных и их представления в браузерах. Хранение файлов в таких системах хорошо изучено и предоставляет ряд апробированных программных решений, большинство которых относят к классу CMS, а для хранения используются реляционные СУБД.

Однако традиционный подход не позволяет гибко работать с содержимым публикации внутри системы: вносить изменения в публикацию в любое время (динамичное содержимое), делать сравнение версий публикаций, представлять материалы в разном виде (например, для индексации в других системах — в виде XML файла, для показа на сайте — в виде HTML документа с форматированием). Еще одним существенным недостатком является необходимость иметь на компьютере программы для просмотра файлов, редактирования примечаний, работы ссылок, навигации по содержимому. Использование сторонних программ ставит под сомнение основное преимущество веб-приложений — мобильность. Оперирование текстом публикации, как файлом, лишает читателей интерактивности, которая обеспечивается современными веб-технологиями (Ajax, Silverlight, HTML 5) и не требует установки на компьютер редакторов. Современные интернет-технологии дают возможность создавать веб-приложения, ничем не отличающиеся по функциональности от rich-приложений, а по времени отклика часто и превосходящие их.

Для решения вышеописанных проблем предлагается отказаться от традиционной формы хранения файлов и разработать форму, более естественную для представления научных знаний. Перейдя от уровня документов и термина «научная публикация» к уровню знаний и термину «научное знание», веб-системы с новой моделью представления можно называть базами научных знаний [5].

Очевидно, что во всех системах кроме основного текста публикации существует еще и структурированные метаданные, которые вводятся в системы дополнительно. Необходимо отметить, что вопро-

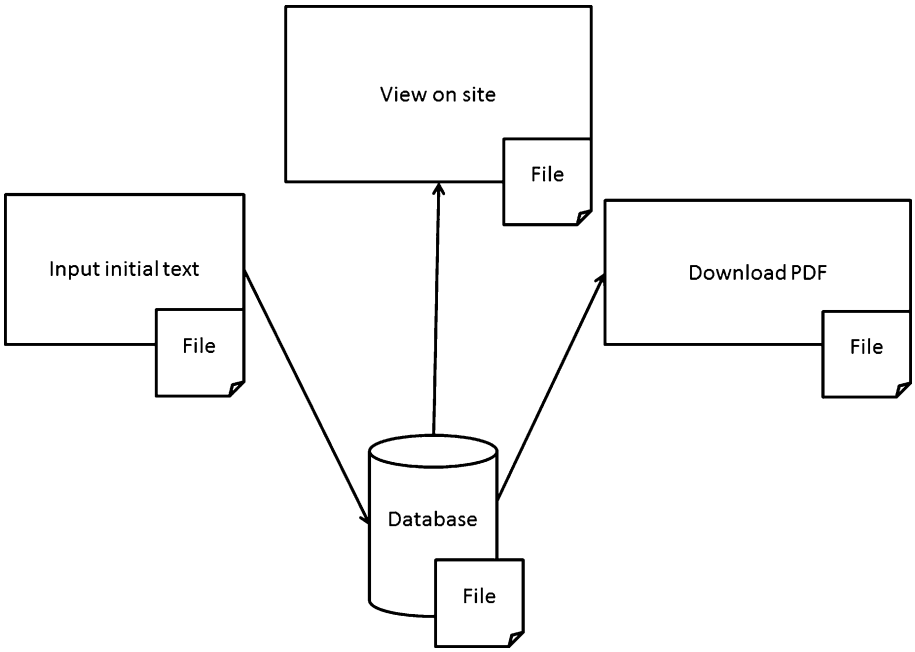

Рис. 1. Традиционный поход представления публикаций

сы хранения в веб-системах метаданных достаточно изучен и в рамках этой статьи не рассматривается.

2. Модель представления научного знания в веб-системах

Проблематика представления научных знаний в информационных системах широко исследуется [5–9], результаты применяются в системах KBS (Knowledge-Based Systems), которые, в свою очередь, успешно решают разного рода прикладные задачи. Здесь мы сосредоточимся на специфике «научности» знаний и на решении проблем описанных выше (гибкость изменения, версионность, простота представления данных). Рассмотрим определения научного знания, которыми оперируют большинство исследователей.

Структура научного знания является сложно организованной системой, и в ее структуре выделяют два уровня, или стадии исследования: эмпирический и теоретический. Первый начинается с анализа первичных данных, полученных в ходе проведения наблюдения или эксперимента. После обработки полученных сведений, информация получает статус научного факт и на теоретическом уровне познания происходит исследование всего процесса, начиная с отдельных суждений и заканчивая построением теоретических гипотез (т. е. предположений). Теоретические знания опираются на исследуемый эмпирический материал (научное обоснование), а эмпирические исследования определяются задачами и целями, поставленными на теоретическом уровне (системность).

Следовательно, научное знание можно представить как множество достоверных эмпирических знаний (научных фактов, авторитетных источников, доказанных ранее законов и принципов) $F_1, F_2, \ldots, F_n$ и полученной на их основе некоторой теории, вывода, гипотезы H. На основе полученного теоретического знания можно строить дальнейшие теоретические выводы, подкрепляемые новыми эмпирическими данными, либо теорией, полученной из достоверных источников $(H_1, H_2, \ldots, H_n)$.

Любое подмножество полученного графа представляет собой структуру научной публикации. При этом рамки самой публикации достаточно условны. Для структуры тезисов это достаточно простой граф. Для монографии или диссертации более сложный, с гораздо большим количеством ребер и узлов. Очевидно, что граф структуры научного знания должен быть связным, не иметь циклов и является ориентированным. Элементы множества узлов могут принимать разную форму — текста, формул, графики, видео. Они не ограничены в размере.

Рассмотрим подробнее, что представляют собой ребра данного графа. Отметим, что обоснованность одно из главных свойств научного знания, а связь между знаниями в публикации имеет разный характер. В рамках одной публикации эта логическая обосновывающая связь – internal. В случае, если связь выходит за рамки текущей публикации, то она ин-

терпретируется как ссылка на источник: на электронный или печатный – external (см. рис. 2).

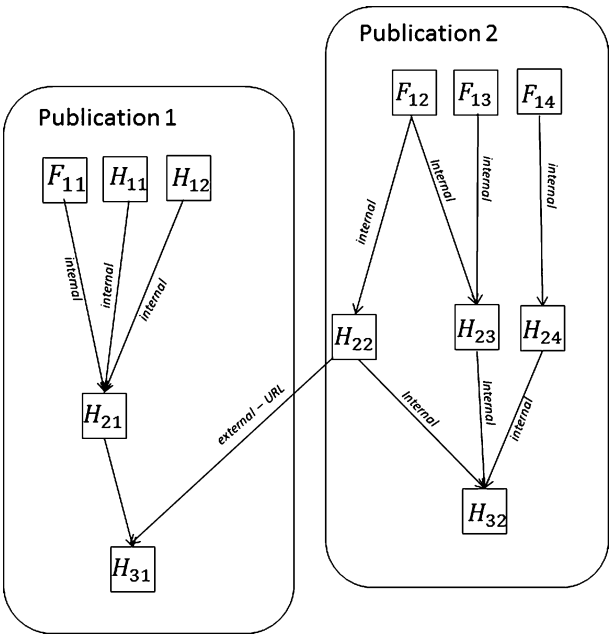

Рис. 2. Модель научных публикаций

В современном мире количество электронных источников растет, и они уже занимают гораздо боле существенную долю по сравнению с печатными источниками. Наиболее логичным вариантом идентификации внешнего источника является использование URL, однако существует ряд препятствий. В российском ГОСТ «5.0.7. Библиографическая ссылка» нет четких указаний по оформлению электронных источников [10]. Например, в нем есть требование к обязательному указанию страниц в журнале, что явно противоречит самой сути электронных публикаций. В ходе анализа представления ссылок в ряде научных электронных изданий (научная электронная библиотека eLibrary [4], электронный журнал «История» [11], электронный журнал «Корпоративные финансы» [12]) в электронном виде можно выделить следующие проблемы:

1. Не всегда веб-сайты можно считать источниками достоверных научных знаний. Как минимум, такой источник должен иметь статус СМИ. На сегодняшний день отсутствует механизм проверки правильности внешних ссылок. Выборочная проверка ряда журналов из перечня ВАК показала, что даже в крупнейшей электронной научной библиотеке eLibrary почти 50 % интернет-ссылок не являются действительными или корректными.
2. Сложность сопоставления ссылок на разных языках. Многие исследователи считают эту проблему одной из основных, препятствующей интеграции России в международное научное сообщество. Наблюдаются существенные различия в требованиях к оформлению ссылок в России и за рубежом. По многим аспектам эти требования не совместимы [10].

Данные проблемы не первый год пытаются решать зарубежные издатели. Крупнейшая в мире вебсистема SCOPUS [1] предлагает использовать коды DOI (Digital Object Identifier) или PII (Publisher Item Identifier), по которому можно легко найти свою статью, указав его прямо в адресе интернет-ресурса [13]. С помощью этих кодов можно однозначно определить саму публикацию, ссылки на нее и идентифицировать ее издателя. Очевидно, что такой же уникальный идентификатор (URI) может иметь каждое научное знание, узел графа в модели.

3. Результаты внедрения

Основные идеи представленной модели реализованы в платформе «е-НОЖ» (электронный научно-образовательный журнал) и апробированы на материалах журнала «История» [11] Института всеобщей истории РАН.

При апробации данной модели элементарной единицей представления научных знаний был выбран абзац, который помимо текста, формул, таблиц и изображений, может содержать и видеоматериалы (см. рис. 3).

Использованная модель структуры публикации позволяет производить быстрые точечные изменения в публикациях, добавлять новые связи без пересчета всей структуры. Результат изменения можно сохранить в виде новой версии, создав ее на основе уже имеющейся структуры с заменой только измененных абзацев. К примеру, автор нашел новое обоснование выводов в опубликованной более года назад статье. Он может легко добавить это в статью, редакторы проверят его правки и опубликуют новую версию без существенных изменений всего опубликованного материала.

Отображение связных графов на реляционные БД хорошо изучено и позволяет не только записывать и читать структуру публикации любой сложности, но и относительно легко масштабировать хранилище. Для этого содержимое знаний и ссылок на них выделено в более быстрое хранилище «ключ-значение», либо используется распределенная СУБД [14].

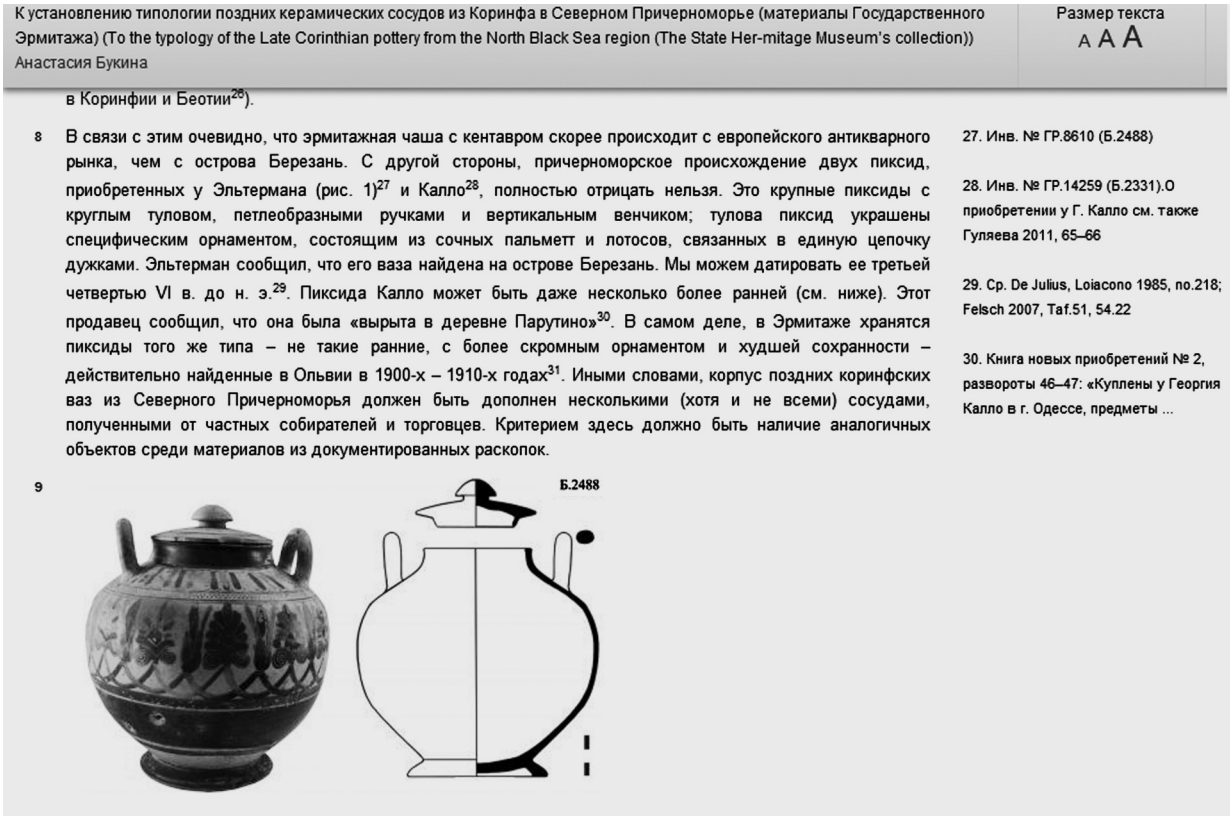

Рис. 3. Представление научных публикаций в электронном журнале «История»

В платформе поддерживаются средства расширенного редактирования, позволяющие редактировать текст и таблицы в формате HTML. В перспективе возможна поддержка редактирования онлайн формул в известном формате [15], элементов векторной графики и т. д. Тем самым, научное знание в терминах платформы — это универсальный контейнер, содержимое которого может принимать любой электронный вид. Теоретически каждый абзац может быть частью нескольких публикаций разных версий и легко идентифицируется и извлекается с учетом уникальности ссылки.

4. Перспективы развития веб-систем научных публикаций

Изложенная модель имеет много общего с идеями развития Semantic Web [16]. Граф научных знаний имеет много общего с RDF-S схемой. Но онтологические языки, в частности OWL, слишком сложны в силу своей ориентации на системы логики, и не имеет смысла использовать их лишь для описания объектной схемы систем. Предложенная модель в силу своей простоты для этого подходит лучше. Напротив многие онтологии могут быть легко адаптированы путем их упрощения для совместимости с данной моделью [17, 18]. Следовательно, в случае распространения Semantic Web, хранилища, построенные по данному принципу, могут быть легко интегрированы с другими семантическими инструментами и ресурсами. В этом случае поиск и извлечение знаний выйдут на принципиально новый уровень, рамки представления знаний в виде документов станут более условными. Это направление активно пропагандируется исследователями Semantic Web и называется «Web of Data». С распространением Semantic Web адрес (URL) научных публикаций, возможно, эволюционирует в уникальные идентификаторы знаний. Интернет объединит веб-системы научных публикаций в единую распределенную базу знаний, если этому не будут противоречить экономические и юридические факторы.

Веб-системы научных публикаций еще не сформировались в единый класс систем, но уже сейчас можно сказать, что решаемые задачи и функционал таких систем схож с функционалом систем дистанционного обучения (e-learning), достаточно сформировавшейся отраслью, обладающей собственным рынком [19]. Рассмотрим основные тенденции рынка систем дистанционного обучения в последние десятилетия.

Основной причиной развития систем e-learning стала необходимость интеграции и обмена учебными материалами между всеми участниками рынка. Результат этой интеграции — формат SCORM, созданный лабораторией ADL (Advanced Distributed Learning). Если внимательно изучить спецификацию SCORM, то видно, что манифест учебного материала в формате SCORM — это xml-файл, содержащий структуру, схожую с представленным в этой статье графом научного знания, в котором узлы — это универсальные контейнеры разнородной информации [20].

По аналогии с e-learning в ближайшее десятилетие разрозненные издатели и библиотеки электронной научной литературы будут стремиться к объединению в единое информационное пространство на основе понятной и естественной модели представления научных знаний. Создание единого формата электронных научных публикаций, аналогичного SCORM — первый весомый шаг на этом пути.

Выводы

1. Традиционному способу представления и хранения материалов в виде файлов не хватает гибкости. Файлы удобны как способ распространения публикаций, но не пригодны для изменения, повторного использования, представления с помощью гипертекста.
2. Представление структуры научных публикаций в виде ориентированного графа более естественно и предоставляет новые возможности: версионность и гибкость изменения, неограниченные возможности визуального представления с помощью современных веб-технологий, масштабируемость.
3. Представленная структура научного знания требует от текущих форматов представления знаний более строгой структуризации и изменения правил оформления ссылок внутри научных публикаций. В частности, это касается российских стандартов.
4. Предложенная модель может быть адаптирована в онтологические модели, что позволит системам, ее использующим, в будущем статью частью Semantic Web. Веб-системы, использующие новую модель, эволюционируют в распределенные базы научных знаний (эволюция Интернет от «Web of Documents» к «Web of Data»).
5. Системы класса e-learning уже используют гибкие подходы к представлению учебных материалов. Результатом стал единый формат для создания и распространения учебных материалов SCORM. Вероятно, по этому же пути будут развиваться системы электронных научных публикаций.

Литература

1. Электронный научный Интернет-портал издательства Elsevier http://www.sciencedirect.com, http://www.scopus.com
2. Портал международной издательской компании «Springer Science+Business Media» http://www.springerlink.com
3. Электронная библиотека издательства «John Wiley & Sons» http://onlinelibrary.wiley.com
4. Научная электронная библиотека eLibrary http://elibrary.ru
5. *Арлазаров В. Л.*, *Емельянов Н. Е.* От баз данных к базам знаний (объекты, формы, содержание) // Труды ИСА РАН. 2006. Т. 23. С. 6–17.
6. *Загорулько Ю. А.*, *Боровикова О. И.* Подход к построения порталов научных знаний // Автометрия. 2008. Т. 44. № 1. С. 100–110.
7. *Глухих И. Н.* Представление знаний и вывод решений в ситуационных базах знаний // Вестник Тюменского государственного университета. 2006. № 5. С. 265–270.
8. *Курганская Г. С.* Модель представления знаний и система дифференцированного обучения через Интернет на его основе // Известия Челябинского научного центра УрО РАН. 2000. № 2. С. 171–180.
9. *Krishnamurthy M. V.*, *Smith F. J.* Integration of scientific data and formulae in an object-oriented knowledge-based system Knowledge-Based Systems. June 1994. V. 7. Iss. 2. P. 135–141.
10. *Кириллова О. В.* Подготовка российских журналов для зарубежной аналитической базы данных SCOPUS: рекомендации и комментарии. http://elsevierscience.ru/info/add-journal-to-scopus
11. Электронный научно-образовательный журнал «История» http://history.jes.su
12. Электронный журнал Корпоративные финансы http://ecsocman.hse.ru/mags/cfjournal
13. http://en.wikipedia.org/wiki/Publisher_Item_Identifier
14. *Кузнецов С.* Будущее транзакционных систем. Открытые системы // СУБД. 2011. № 4. ISSN: 1028–7493.
15. ru.wikipedia.org/wiki/TeX

16. *Хорошевский В. Ф.* Пространства знаний в сети Интернет и Semantic Web (Часть 1) // Искусственный интеллект и принятие решений. ISSN 2071–8594 2008 / 01. С. 80–97.
17. *Guarino N.* Formal Ontology and Information Systems. In: Guarino N. (ed.) Proc. 1st Int'l Conference on Formal Ontology in Information Systems, 3–15. IOS Press/Ohmsha, 1998.
18. *Клещев А. С.*, *Артемьева И. Л.* Математические модели онтологий предметных областей. Часть 3. Сравнение разных классов моделей онтологий // Научно-техническая информация. Сер. 2.
19. РБК обзор рынка. Обзор рынка дистанционного образования в России. http://www.rbc.ru/reviews/business-education-2008/chapter5-distance-1.shtml
20. *Кузнецов В.*, *Баринов А.* Web-технологии в образовании. Системы дистанционного обучения в Интернет. http://e-commerce.ru

Тарханов Иван Александрович. С. н. с. ИСА РАН. Окончил МФТИ в 2005 г. Количество печатных работ: 12. Область научных интересов: базы данных, базы знаний, документооборот, BPM. E-mail: ivant@cs.isa.ru