ОНЛАЙНОВАЯ ИНФРАСТРУКТУРА ОТКРЫТЫХ НАУЧНЫХ ДАННЫХ

Т.А. Полилова

С появлением интернета произошла заметная трансформация представлений о научных публикациях. Все чаще можно встретить научные издания, публикующие материалы исключительно в интернете. Для традиционных бумажных журналов стало нормой иметь интернет-проекцию. В обиход вошел новый термин — *онлайновое* издание — для обозначения издания в интернете. Этот термин, по-нашему мнению, подчеркивает более тесную связь автора со своей публикацией и возможность обновлять опубликованный материал. В западных странах, где сильны позиции сторонников идей открытой науки, стал широко использоваться термин — открытые научные данные. Сторонники открытой науки развивают две основные темы: свободный доступ к данным и разнообразие форм представления результатов научных исследований.

Перечисленные тенденции заставляют по новому взглянуть на пространство научных публикаций, в недрах которого формируются механизмы представления и агрегирования научных данных. Создание новой инфраструктуры требует, с одной стороны, консолидации научного сообщества, принимающего новые реалии, с другой стороны, движение в сторону открытых научных данных невозможно без поддержки государства. Нужны стратегические решения на уровне органов управления наукой, направленные на реализацию идей открытой науки.

Рассмотрим некоторые проблемы, связанные с развитием научного интернета.

Наука в интернете

Для научного сообщества интернет стал главенствующим способом получения информации. Ученые, преподаватели, студенты и аспиранты, используя разнообразные поисковые механизмы, ищут теперь необходимые материалы не в библиотеке, а в Сети. Интернет-публикация превратилась в основное средство донесения идей ученого до потребителя.

Интернет предлагает новые, несоизмеримо более продуктивные формы донесения информации до читателя, взаимодействия с читателем. Статья или книга в интернете становится живым, каждодневно улучшаемым проектом, аналогом динамичного интернет-портала в несколько уменьшенном масштабе. Публикация обычно атрибутируется электронным адресом автора, благодаря чему легко завязывается переписка с читателями. Другая форма коммуникации с посетителями сайта публикации — электронный форум, где каждый вправе разместить для общего сведения любые соображения, возникшие у него при прочтении работы. Еще одна важная форма взаимодействия — подписка. Посетитель сайта публикации может оставить там свой электронный адрес, в результате чего он начинает автоматически получать электронные извещения о всех важных событиях, происходящих в заинтересовавшем его направлении исследований.

В интернете автор может выбирать форму подачи своего материала, не будучи связанным ограничениями полиграфии. В его распоряжении весь спектр возможностей мультимедиа: цвет, звук, видео, анимация, интерактивная компьютерная графика и др. Применение гиперссылок формирует существенно более комфортную среду чтения. Например, по ссылке в тексте статьи читатель за доли секунды попадает на ее расшифровку в пристатейном библиографическом списке и далее на сам цитируемый источник (если он размещен в интернете). Прямо в тексте можно поместить интерактивные панели, позволяющие читателю тут же заказывать и получать результаты произвольных вычислений, извлекать заинтересовавшие его данные из массивов информации практически неограниченного объема и др.

Все большую популярность приобретают технологии семантического веба. Идея сематического веба заключается в создании системы описания знаний в достаточно формальном виде, доступном для обработки компьютером. Семантический веб представляет собой следующее поколение World Wide Web, где гипертекстовые документы дополняются описанием семантики этих документов. Формальная спецификация содержимого веб-документа дает возможность поисковой программе делать выводы о соответствии найденных документов поисковому запросу не только на основе лексических компонентов, составляющих текст этого документа, но и с привлечением семантики документа, описанной с помощью явно указанных метаданных. Построение научного информационного пространства с привлечением технологий семантического веба поможет структурировать научные ресурсы в согласованной системе понятий, улучшить качество поиска в научном интернете, создать основу для агрегации научных данных и содержательного анализа происходящих в науке процессов.

Поддержка государства

За рубежом научные интернет-публикации получают серьезную поддержку со стороны государства. Активно развивается движение «Открытых архивов», призывающее обеспечить свободный интернет-доступ к научным статьям. Все чаще выделение средств на науку обуславливается жестким требованием: если опубликованная на бумаге статься написана по результатам исследований, выполненных на государственные

средства, то не позднее, чем через полгода после выхода она должна быть размещена в интернете в свободном доступе (полгода задержки — уступка интересам издателей).

Вслед за ведущими европейскими странами соответствующий меморандум в 2013 г. выпустила и администрация президента США Б. Обамы. В меморандуме содержится призыв к федеральным агентствам сделать доступными и полезными для общественности, промышленности и научного сообщества непосредственные результаты научных исследований, финансируемых из федерального бюджета. В меморандуме отмечается, что научные исследования, поддерживаемые федеральным правительством, являются катализатором инновационных прорывов, приводящих в движение экономику.

С 2004 года Кибернетическая лаборатория Национального исследовательского совета Испании два раза в год публикует рейтинг Webometrics сайтов университетов и научно-исследовательских центров всего мира. Рейтинг вычисляется по несложной формуле, где основную роль играют число общедоступных размещенных на сайте научных работ и количество ссылок на них. Очередной рейтинг был опубликован по состоянию на февраль 2013 г. Среди вузов первые 13 мест заняли американские университеты. На первых местах, как обычно, оказались сайты Гарварда и Массачусетского технологического института, где несколько лет назад были приняты жесткие административные решения: размещать на своем сайте в свободном доступе все статьи, написанные сотрудниками вуза.

Аналогичные нормативно-правовые документы, обеспечивающие развитие национального научного интернет-пространства, должны появиться и в нашей стране. Необходимо разработать требования к нормативно-правовой базе открытых научных интернет-публикаций, подготовить и обсудить в научной аудитории содержание распорядительных документов и регламентов, которые будут стимулировать размещение научной продукции в интернете (на сайтах научных организаций, сайтах фондов, финансирующих научные исследования, тематических научных сайтах, сайтах инициативных научных проектов, электронных библиотек и т.д.).

Онлайновая инфраструктура научных публикаций

Основные результаты научной деятельности ученого представлены в научных публикациях. В силу сложившихся традиций большая доля научных результатов до сих пор публикуется в традиционных "бумажных" изданиях. Однако в последнее десятилетие ситуация заметно изменилась. Научные журналы все чаще выкладывают материалы выпусков в интернете. Появляются онлайновые научные издания. Функционирует электронная библиотека eLIBRARY.RU (http://elibrary.ru/) — информационный портал, содержащий электронные версии более 2500 российских научно-технических журналов. На базе eLIBRARY.RU реализуется проект РИНЦ (российский индекс научного цитирования). Создан общероссийский математический портал Math-Net.Ru (http://www.mathnet.ru/), агрегирующий публикации в области математики, предоставляющий профессиональному сообществу развитые средства поиска, представления на сайте проекта статей по математике. Реализуются проекты Открытых архивов. Один из таких состоявшихся проектов — проект Соционет (http://www.socionet.ru/), содержащий коллекции научных материалов по основным дисциплинам общественных наук.

Многие научные учреждения, в частности институты РАН, также являются издателями научной продукции. Институтское издание обычно имеет онлайновую проекцию и на регулярной основе представляется на сайте института В качестве примера можно привести онлайновую библиотеку препринтов сотрудников института, реализованную на сайте ИПМ им. М.В.Келдыша РАН (http://library.keldysh.ru/preprints/).

Сайты научных учреждений, научных интернет-проектов, открытые архивы представляется целесообразным интерпретировать как базовые элементы строящейся онлайновой инфраструктуры научных публикаций. В институтах РАН, в онлайновых проектах, связанных с публикацией научных трудов, сейчас используются разные подходы к представлению на сайтах научной продукции. При агрегировании таких компонентов возникает задача создания согласованной системы представления информации о научных изданиях (ресурсах).

Агрегирование научной продукции имеет смысл проводить в первую очередь на уровне метаданных. Метаданные – основные существенные сведения об издании, представленные в формальной структуре. Метаданные научной публикации указываются в момент размещения ее на сайте (например, на сайте научной организации или сайте научного проекта) в том объеме, в котором данный сайт взаимодействует с сайтами-агрегаторами, т.е. потребителями метаданных. Такими сайтами-агрегаторами являются сайты систем сбора, учета, анализа научной продукции.

Одной из важных и первоочередных задач является разработка механизмов агрегации результатов научной деятельности, которые должны учитывать динамичное развитие формируемой инфраструктуры, появление новых агрегаторов, расширение набора атрибутов научной публикации, изменение структуры метаданных.

Персональные страницы ученых

Необходимым компонентом онлайновой инфраструктуры открытых научных данных является персональная информация об авторах научных публикаций. Персональную информацию, как правило,

формирует (и поддерживает в актуальном состоянии) сам автор на своей персональной странице. Если персональная страница размещается на сайте организации, где работает автор, появляются дополнительные возможности контролировать состав и достоверность персональных данных. Так, например, на персональной странице сотрудника ИПМ им. М.В.Келдыша РАН (см., например, персональную страницу автора статьи http://keldysh.ru/persons/polilova.html) часть данных заполняется сотрудником отдела кадров на основании первичных документов: данных паспорта, наградных документов, свидетельств, дипломов доктора или кандидата наук, приказов о назначении на должность и т.д. На персональной странице сотрудника автоматически формируется ссылка на публикации автора, размещенные в электронной библиотеке.

Персональная страница – важный, первичный элемент инфраструктуры. В научном интернете адрес персональной страницы играет роль семантического связующего звена между статьей и ее автором. Этот адрес должен войти в состав метаданных, сообщаемых внешним библиографическим системам при размещении статьи на сайте. Именно от персональной страницы отталкиваются при поиске в интернете работ определенного автора, при подсчете числа цитирований и при выполнении других подобных операций.

Научному сообществу предстоит согласовать предложения по рациональному составу информации, размещаемой на персональной странице ученого. Вслед за этим должны быть разработаны регламенты, обслуживающие агрегирование персональных страниц ученых в единую онлайновую инфраструктуру открытых научных данных на базе известных технологий OpenID, ORCID, ResearcherID.

Объекты агрегирования

В онлайновой инфраструктуре открытых научных данных представляется целесообразным агрегировать следующие объекты.

- Публикации в научных периодических и сериальных изданиях.
- Отчеты по проектам НИР, НИОКР, ОКР.
- Авторефераты диссертаций на соискание ученой степени.
- Диссертации на соискание ученой степени.
- Монографии.
- Публикации в сборниках научных конференций, семинаров, симпозиумов и пр.
- Учебники, методические пособия.
- Статьи в энциклопедии.
- Научные авторские сайты.

Список объектов, вовлекаемых в строящуюся инфраструктуру, может расширяться. Например, в ближайшее время ожидается массовое признание такого важного интернет-объекта, как инициативная открытая рецензия члена экспертного сообщества на размещенную на общедоступном сайте статью. Разумеется, такого рода объекты займут в инфраструктуре почетное место.

Подходы к оценке качества и востребованности научных статей

В какой-то мере задачу агрегирования определенной части научной продукции в настоящее время решает онлайновая библиотека eLIBRARY.RU. Библиотека работает в основном с научными журналами. В эту библиотеку издательства могут передавать материалы своих научных изданий, заключив соответствующий договор. Передаются метаданные отдельных статей и, возможно, их полные тексты. Библиотека реализует проект РИНЦ (Российский индекс научного цитирования). Потребности РИНЦ во многом определяют и структуры метаданных: наряду с основными параметрами выпуска издания библиотека требует библиографические списки статей для организации подсчета числа цитирований.

Господствующий в настоящее время библиометрический метод оценки значимости журнальных статей ориентирован на бумажные издания. Метод основан на разборе списка используемой литературы и подсчете цитирований авторов. Достаточно очевидный и оригинальный в те годы способ оценить значимость статей и журналов в условиях интернета заметно уступает уверенно завоевывающим популярность методам оценки значимости интернет-ресурсов. В частности, представляет интерес подсчет числа закладок, сделанных посетителями интернет-ресурса, подсчет положительных рецензий, написанных членами экспертного сообщества и т.д. Создаваемая онлайновая инфраструктура позволит обеспечить новые подходы в оценке качества и востребованности научных статей.

Однако до тех пор, пока библиографические базы данных остаются востребованными, в онлайновой инфраструктуре имеет смысл предусмотреть возможность включения в состав метаданных научной публикации списка цитируемой литературы.

Живая публикация

Интернет позволяет автору непрерывно развивать, дополнять и совершенствовать свою размещенную на сайте работу, т.е. превратить ее в живую публикацию.

Технология живых публикаций требует применения соответствующих средств синхронного изменения информации на сайтах агрегаторов. Приведем пример. Публикация была дополнена автором, в результате

появилась ссылка на новый источник цитирования. Такое изменение приводит к изменению метаданных (расширился библиографический список цитируемых авторов). Следовательно, должна соответственно измениться информация в системах, выполняющих подсчет цитирований. Еще один пример. Публикация получила новую положительную рецензию. Соответственно должна поменяться информация на сайте агрегаторе, где учитывается число полученных положительных рецензий для определения рейтинга публикации.

Надежное хранение онлайнового научного контента

Заслуживает внимания проблема сохранения онлайнового научного контента. Еще недавно тут все было относительно благополучно. Научно-технический центр Информрегистр на своем сайте вел реестр зарегистрированных электронных научных изданий. Регистрация выпуска издания сопровождалась выдачей идентификационных номеров статьям, включенным в этот выпуск. Информрегистр выполнял также роль хранилища полных текстов онлайновых статей. В Информрегистр на электронных носителях представлялись копии выпусков онлайновых изданий. Несколько копий Информрегистр поставлял в ведущие библиотеки. Тем самым Информрегистр обеспечивал надежное долговременное хранение онлайнового компонента научного и культурного наследия, подобно Российской книжной палате, несколько столетий выполняющей ту же функцию для печатных изданий.

К сожалению, с 1 ноября 2012 г. Информрегистр прекратил регистрацию и хранение электронных научных изданий. В результате онлайновые издания стали чрезвычайно уязвимыми по сравнению с другими видами изданий: ссылки на материалы, опубликованные в онлайновых изданиях, вообще говоря, теряют доверие. Сайт онлайнового издания может исчезнуть (например, быть недоступен по техническим причинам), и ссылка на материалы, опубликованные в этом онлайновом издании, будет вызывать правомерный вопрос «А был ли мальчик?».

Но положение можно спасти. Утраченную в настоящее время функцию сохранения контента научного интернета могут взять на себя отдельные агрегаторы.

Таким образом, на данном этапе актуальным представляется решение следующих задач:

- Разработка концепции открытой инфраструктуры научных данных, в которой:
  - отражается современный уровень развития технологий представления научных данных и результатов научных исследований;
  - используются механизмы семантического веба – метаданные как основа структуризации научного контента и агрегирования научных данных;
  - описываются структура и функциональность базовых элементов строящейся инфраструктуры (в частности, сайтов научных организаций, персональных страниц ученых), а также необходимый набор метаданных, позволяющий агрегировать научные данные на специализированных сайтах-агрегаторах научной продукции;
  - предлагаются конкретные механизмы агрегирования результатов научных исследований, их систематического обновления и поддержания в актуальном состоянии;
  - предлагаются технологические решения для долговременного и надежного хранения научного контента.
- Разработка пакета нормативно-правовых документов и организационных регламентов, обеспечивающих создание и функционирование открытой инфраструктуры научных данных, для начала – в академическом секторе российской науки.

Задача создания механизмов поддержки развивающихся во времени научных ресурсов, их систематического обновления и поддержания в актуальном состоянии, динамического отслеживания возникающих между развивающимися ресурсами семантических связей не решена пока в требуемом объеме ни в России, ни в других странах с развитой информационно-коммуникационной инфраструктурой.

Решение перечисленных выше задач позволит заложить основу для поступательного развития онлайновой инфраструктуры открытых научных данных, сделать доступными для общественности, промышленности и научного сообщества актуальные результаты научных исследований.

Работа выполнена при поддержке гранта РФФИ № 13-01-00493 а.

ЛИТЕРАТУРА:

1. Т.А. Полилова. Инфраструктура научных публикаций // Препринты ИПМ им. М.В.Келдыша. 2009. № 15. 30 с. URL: http://library.keldysh.ru/preprint.asp?id=2009-15.
2. А.В. Ермаков, Т.А. Полилова. Статус научной публикации // Научный сервис в сети Интернет: труды Международной конференции (17-22 сентября 2012 г., г. Новороссийск): [электр. изд.]. 2012. С. 635-639. ISBN 978-5-211-06394-5.
3. М.М. Горбунов-Посадов. Живая публикация. — М.: ИПМ им.М.В.Келдыша, 2011. Обновлено 14.06.2013. — http://keldysh.ru/gorbunov/live.htm

4. Т.А. Полилова. Библиографическая ссылка и живая публикация // Вестник ЮУрГУ. Серия "Математическое моделирование и программирование". 2011. вып. 7, № 4 (211). С. 75-81.
5. М.М. Горбунов-Посадов. Интернет-активность как обязанность ученого // Информационные технологии и вычислительные системы. 2007. № 3. С. 88–93. URL: http://keldysh.ru/gorbunov/duty.htm
6. М.М. Горбунов-Посадов. Персональная веб-страница ученого. — М.: ИПМ им.М.В.Келдыша, 2011. Обновлено 04.03.2012. — http://keldysh.ru/gorbunov/person.htm
7. Т.А. Полилова. Персональные веб-страницы в научном сообществ // Научный сервис в сети Интернет: труды Международной конференции (19-2 сентября 2011 г., г. Новороссийск). Элект. издание, 2011. С. 476-479. ISBN 978-5-211-06229-0.