Все на борт! Вперед, к автоматизированной системе научного взаимодействия!

ГЕРБЕРТ ВАН ДЕ СОМПЕЛ (HERBERT VAN DE SOMPEL)
Лос-Аламосская национальная лаборатория (Los Alamos National Laboratory)

КАРЛ ЛАГОЗЕ (CARL LAGOZE)
Корнельский университет (Cornell University)

«Существующая система научного взаимодействия представляет собой не что иное, как отсканированную копию бумажной системы».

Это утверждение, которое мы произносили во время многочисленных презентаций на конференциях для того, чтобы произвести впечатление на аудиторию, было окончательно сформулировано в статье, выпущенной в 2004 г. [1]. Однако и по сей день оно в общем и целом справедливо. Хотя издатели научной литературы стали использовать новые технологии (такие как Интернет и документы PDF), в значительной степени упрощающие доступ к научным материалам, эти изменения не реализуют весь потенциал новых цифровых и сетевых возможностей. В частности, они не помогают преодолеть три недостатка широко распространенной системы научного взаимодействия:

- системные проблемы, в особенности неразрывная связь в системе публикации между подачей научной заявки на совместную авторскую разработку и процессом экспертного рецензирования;
- экономические проблемы, проявляющиеся в кризисе периодических изданий, что создает огромную нагрузку на библиотеки;

- технические проблемы, препятствующие реализации информационной инфраструктуры с поддержкой взаимодействия.

Наше беспокойство о состоянии научного взаимодействия разделяют многие специалисты по всему миру. Почти десять лет назад в сотрудничестве с представителями мирового научного сообщества мы создали группу Open Archives Initiative (OAI, Инициатива открытых архивов), которая оказала значительное влияние на направление и темпы развития движения Open Access (Открытый доступ). Протокол OAI-PMH и последующие попытки создания протокола OpenURL отражали нашу изначальную ориентированность на аспекты научного взаимодействия, связанные с процессами. Другие члены сообщества интересовались собственно научными информационными ресурсами. Например, Питер Мюррей-Раст (Peter Murray-Rust) занимался преобразованием структурированной, пригодной для машинной обработки информации (например, табличных данных и базовых координат, лежащих в основе графиков) в читаемый текст, который подходит только для чтения людьми [2].

Спустя десятилетие после начала нашей работы в этой области мы с радостью наблюдаем быстрые изменения, происходящие в различных форматах научного взаимодействия. Мы остановимся на трех областях, изменения в которых, на наш взгляд, достаточно значительны, чтобы свидетельствовать о фундаментальных переменах.

РАСШИРЕНИЕ ВОЗМОЖНОСТЕЙ АРХИВА НАУКИ С ПОМОЩЬЮ БАЗОВОГО СЛОЯ КОМПЬЮТЕРНОЙ ОБРАБОТКИ

Одним из движущих факторов для реализации компьютерной обработки научных материалов является шквальный поток литературы, который исключает для ученых возможность быть в курсе соответствующих научных знаний [3]. Для решения этой проблемы можно использовать агенты, которые *читают и фильтруют* научные материалы вместо ученых. Потребность в подобном механизме усиливается тем фактом, что ученым все чаще нужно осваивать и прорабатывать литературу по другим дисциплинам, связывая воедино и объединяя имеющиеся разрозненные результаты исследований, чтобы прийти к новым идеям. Это основная проблема для медико-биологических наук, характеризующихся большим числом взаимосвязанных дисциплин (таких как генетика, молекулярная биология, биохимия, фармакохимия и органическая химия). Например, отсутствие единообразно структурированных данных в различных отраслях биомедицины считается серьезным препятствием для межотраслевых исследований — передачи открытий в основных биологических и медицинских исследованиях для их применения в клиническом лечении пациентов [4].

В последнее время мы наблюдаем выраженное движение в сторону машинного представления знаний, содержащихся в медико-биологической литературе, что делает возможными логические рассуждения, преодолевающие межотраслевые барьеры. Для извлечения сущностей и отношений между ними из имеющейся литературы применяются передовые методики анализа текста, а для достижения универсального представления знаний были созданы общие онтологии. Этот подход уже привел к появлению новых открытий, основанных на информации, содержащейся в литературе, которую раньше мог прочитать только человек. Другие отрасли знаний последовали этому примеру. Некоторые инициативы позволяют ученым добавлять информацию о сущностях и их связях на этапе публикации статьи, чтобы отказаться от обработки статьи после ее выхода, что широко распространено в настоящее время [5].

Создание международной организации Concept Web Alliance, целью которой является предоставление глобальной межотраслевой платформы для *обсуждения, проектирования и, возможно, сертификации решений для обеспечения совместимости и пригодности к использованию обширных, разрозненных и сложных данных*, означает, что тенденция к машинной обработке данных серьезно воспринимается и научным сообществом, и отраслью научной информации. Создание машинного представления научных знаний поможет ученым и учащимся справиться с избытком информации. Оно позволит совершать новые открытия путем размышления над имеющимися знаниями, а также повысить скорость открытий, избавляя ученых от ненужных исследований и открывая перспективные пути для новых исследований.

ИНТЕГРАЦИЯ НАБОРОВ ДАННЫХ В АРХИВ НАУКИ

Хотя данные всегда были ключевой составляющей научных исследований, до сих пор к ним относились не как к первостепенным объектам в процессе научного взаимодействия — в отличие от научных статей с описанием открытий, сделанных на основе этих данных. Эта ситуация быстро и радикально меняется. Научное сообщество активно обсуждает и изучает возможности реализации всех основных функций научного взаимодействия — *регистрации, сертификации, информирования, архивирования и награждения* [1] — для наборов данных.

Например, пирамида данных [6] ясно демонстрирует, как обеспечение надежности (*сертификация*) и цифровой сохранности (*архивирование*) наборов данных становится насущной потребностью по мере того, как они выходят за рамки личного использования в сферу отраслевых научных сообществ и общества в целом. Международные инициативы, нацеленные на реализацию обмена научными данными [7], отражают необходимость в инфраструктуре, способствующей созданию общих наборов данных (*информирование*). А работы по формированию стандартного формата цитирования для наборов данных [8] подразумевают, что наборы данных являются основными научными артефактами. Эти инициативы отчасти мотивированы убеждением в том, что ученые должны заслужить хорошую репутацию (быть *награждены*) за наборы данных, которые они составили и предоставили в распоряжение других ученых. Примерно десять лет назад эти функции научного взаимодействия были главным образом применимы только к научной литературе.

ВЫЯВЛЕНИЕ ПРОЦЕССА НАУЧНОГО ВЗАИМОДЕЙСТВИЯ И ЕГО ИНТЕГРАЦИЯ В АРХИВ НАУКИ

Некоторые аспекты процесса научного взаимодействия были давно известны. Цитаты в публикациях отражают использование имевшихся знаний для создания новых идей. Таким образом, график научного цитирования помогает обнаружить аспекты научной динамики и, как результат, активно изучается для выявления связей между различными отраслями знаний, а также анализа и прогнозирования тенденций. Однако интерпретация графика научного цитирования часто ненадежна вследствие несовершенных методов ручного и автоматического поиска цитат и ссылок и сложных проблем с устранением авторской многозначности. Область охвата графика цитирования также ограничена (только ведущие журналы или только конкретные отрасли знания), и, к сожалению, самый репрезентативный график (Thomson Reuters) защищен правом собственности.

Проблема с графиком цитирования является отражением более широкой проблемы: отсутствие непротиворечивой, задокументированной и наглядной картины эволюции научной статьи в системе, а также отсутствие информации о природе этой эволюции. Проблема в том, что связи, известные на момент перехода научной статьи на новую ступень в цепочке ценностей, практически сразу после этого теряются, и часто навсегда. Реальная динамика научных знаний — взаимодействие и связи между научными статьями, авторами, читателями, оценкой качества статей, областями научного исследования и т.п. — чрезвычайно сложно поддаются восстановлению постфактум. Поэтому важно создать слой, лежащий в основе научного взаимодействия, — систему, которая будет фиксировать и обнаруживать такую динамику, связи и зависимости.

Решение этой проблемы возникает в рамках нескольких инновационных инициатив, обеспечивающих публикацию информации о научном процессе в пригодной для автоматической обработки форме в Интернете, предпочтительно в момент, когда происходят вышеописанные события, и следовательно, когда имеется вся необходимая информация.

В частности, что касается графика цитирования, метод веб-цитирования, разрабатываемый в проекте CLADDIER, представляет механизм кодирования точного, доступного для обхода графика цитирования в Интернете. Несколько инициатив ориентированы на внедрение авторских идентификаторов [9], которые позволят создавать менее противоречивые графики цитирования. График, снабженный семантикой цитирования, например, предложенной в проекте Citation Typing Ontology, также будет содержать информацию о причине цитирования артефакта — важном аспекте, который до сих пор оставался неясным [10].

Помимо обработки данных о цитировании предпринимались и другие усилия по разработке научного процесса, включая проекты по реализации обмена научными данными (процесс фокусировки внимания на научной информации), такие как COUNTER, MESUR и сервис научных рекомендаций bX. В совокупности эти проекты иллюстрируют широкие возможности применения подобной информации о процессе научного взаимодействия для создания коллекций, расчета новых показателей для оценки влияния научных артефактов [11], анализа текущих направлений исследований [12] и рекомендательных систем. В результате этой работы несколько проектов в Европе занимаются поиском технических решений для обмена подробными данными об использовании в Интернете.

Другой пример автоматизации процесса — успешная инициатива myExperiment, представляющая социальный портал для обмена описаниями вычислительных потоков работ. Аналогичные проекты в химии обеспечивают публикацию и обмен информацией из лабораторных журналов через Интернет [13].

Мы считаем эти инициативы особенно вдохновляющими, поскольку они позволяют нам представить следующий логический шаг — обмен информацией об источнике данных. Данные об источнике представляют журнал входных данных и этапов обработки при выполнении потоков работ и являются важным аспектом научной информации как для обеспечения уверенности в достоверности данных, так и для поддержки воспроизводимости результатов, требуемой от всех экспериментальных наук. Недавние работы в ИТ-сообществе [14] привели к созданию систем, способных хранить подробную информацию об источнике данных в единой среде. Мы полагаем, что информация об источнике данных, описывающая и связывающая потоки работ, наборы данных и процессы, представляет новый тип метаданных о процессе, который сыграет ключевую роль в науке, основанной на сетях и использующей большой объем данных. Ее значение можно сравнить с важностью описательных метаданных, данных о цитировании и об использовании данных в научной среде, основанных на публикациях. Следовательно, логично предположить, что в конечном итоге информация об источнике данных будет автоматизирована, чтобы она могла быть использована различными инструментами для поиска, анализа и оценки влияния некоторых основных продуктов новых научных знаний: потоков работ, наборов данных и процессов.

ВЗГЛЯД В БУДУЩЕЕ

Как говорилось выше, архив науки будет представлять собой результат объединения традиционных и новых научных артефактов, разработки автоматизированной основы научных знаний и автоматизации метаданных о научном процессе. Эти возможности достигнут своего максимального потенциала только при условии, что они будут реализованы в подходящей и обеспечивающей взаимодействие киберинфраструктуре, основанной на Интернете и соответствующих веб-стандартах. Применение Интернета будет не только способствовать целостности научного процесса, но и позволит гибко интегрировать научные дискуссии в более широкий контекст общественных дискуссий, ведущихся в Интернете.

За последнее время мы отмечаем растущую ориентированность на Интернет при разработке подходов к реализации научного взаимодействия. Это проявля-

ется в пробном или активном использовании идентификаторов URI, в частности HTTP URI для идентификации научных артефактов, понятий, исследователей и учреждений, а также применение форматов XML, RDF, RDFS, OWL, RSS и Atom для представления научной информации и знаний и обмена ими. Эти основные технологии все чаще дополняются совместимыми спецификациями, которые применяются и разрабатываются в отдельных научных сообществах. В общем, вырисовывается ситуация, когда все составляющие нового архива науки (автоматизированные и пригодные для чтения человеком) публикуются в Интернете в соответствии с общими веб-стандартами и отраслевыми спецификациями этих стандартов. После их публикации в Интернете они становятся доступными для просмотра, сбора и анализа как человеком, так и автоматизированными агентами.

Наша работа над спецификациями OAI Object Reuse & Exchange (OAI-ORE) [15], в которых определяется подход к идентификации и описанию ресурсов е-Науки, представляющих агрегации нескольких ресурсов, является иллюстрацией нового подхода, основанного на киберинфраструктуре, ориентированной на Интернет. Этот подход основывается на базовых веб-технологиях, а также следует основным принципам инициативы Linked Data, которая быстро развивается как наиболее масштабное проявление многих лет работы над проектом Semantic Web.

Описывая тенденции использования общих веб-технологий для научных целей, мы вспоминаем о Джиме Грее, который в ходе дискуссий, предварявших работы над спецификациями OAI-ORE, настаивал на том, что в любом решении должны применяться общие технологии потоковой передачи — RSS или Atom. Джим был прав, говоря о том, что для удовлетворения требований научного взаимодействия необходимо разработать множество специализированных компонентов киберинфраструктуры, а прочие компоненты будут доступны сразу как результат общих работ по веб-стандартизации.

Заглядывая в ближайшее будущее, мы вспоминаем одну из известных фраз Джима Грея: «Да будут все ваши проблемы техническими». Этим ироничным комментарием Джим хотел сказать, что за самыми сложными техническими проблемами стоит еще более фундаментальная проблема: интеграция киберинфраструктуры в потоки работ и практики ученых. Без этой интеграции даже самая лучшая инфраструктура не получит широкого распространения. К счастью, есть признаки того, что мы выучили этот урок опытным путем в результате многолетней работы над другими крупномасштабными инфраструктурными проектами, такими как Digital Libraries Initiative. Программа The Sustainable Digital Data Preservation and Access Network Partners (DataNet), финансируемая подразделением Office of Cyberinfrastructure Национального научного фонда США, недавно утвердила финансирование двух 10-летних проектов, изучающих киберинфраструктуру как социотехническую проблему, требующую знания технологий и понимания способов их интеграции в сообщества. Мы полагаем, что эта более широкая цель станет одним из самых важных факторов, которые будут способствовать изменению природы научного знания и способов его передачи в следующем десятилетии.

Мы уверены в том, что продолжающееся развитие Интернета, появление новых технологий, использующих его основные принципы, и понимание того, как люди используют технологии, в совокупности послужит основанием для фундаментально переосмысленной системы научного взаимодействия, удобной для человека и поддерживающей машинную обработку. С появлением этой системы мы с удовольствием откажемся от когда-то любимого нами сравнения существующей системы научного взаимодействия с отсканированной копией бумажной системы.

ССЫЛКИ

[1] H. Van de Sompel, S. Payette, J. Erickson, C. Lagoze, and S. Warner, «Rethinking Scholarly Communication: Building the System that Scholars Deserve», D-Lib Mag., vol. 10, no. 9, 2004, www.dlib.org/dlib/september04/vandesompel/09vandesompel.html.

[2] P. Murray-Rust and H. S. Rzepa, «The Next Big Thing: From Hypermedia to Datuments», J. Digit. Inf., vol. 5, no. 1, 2004.

[3] C. L. Palmer, M. H. Cragin, and T. P. Hogan, «Weak information work in scientific discovery», Inf. Process. Manage., vol. 43, no. 3., pp. 808—820, 2007, doi: 10.1016/j.ipm.2006.06.003.

[4] A. Ruttenberg, T. Clark, W. Bug, M. Samwald, O. Bodenreider, H. Chen, D. Doherty, K. Forsberg, Y. Gao, V. Kashyap, J. Kinoshita, J. Luciano, M. S. Marshall, C. Ogbuji, J. Rees, S. Stephens, G. T. Wong, E. Wu, D. Zaccagnini, T. Hongsermeier, E. Neumann, I. Herman, and K. H. Cheung, «Advancing translational research with the Semantic Web», BMC Bioinf., vol. 8, suppl. 3, p. S2, 2007, doi: 10.1186/1471-2105-8-S3-S2.

[5] D. Shotton, K. Portwin, G. Klyne, and A. Miles, «Adventures in Semantic Publishing: Exemplar Semantic Enhancements of a Research Article», PLoS Comput. Biol., vol. 5, no. 4, p. e1000361, 2009, doi: 10.1371/journal.pcbi.1000361.

[6] F. Berman, «Got data?: a guide to data preservation in the information age», Commun. ACM, vol. 51, no. 12, pp. 50—56, 2008, doi: 10.1145/1409360.1409376.

[7] R. Ruusalepp, «Infrastructure Planning and Data Curation: A Comparative Study of International Approaches to Enabling the Sharing of Research Data», JISC, Nov. 30, 2008, www.dcc.ac.uk/docs/publications/reports/Data_Sharing_Report.pdf.

[8] M. Altman and G. King, «A Proposed Standard for the Scholarly Citation of Quantitative Data», D-Lib Magazine, vol. 13, no. 3/4, 2007.

[9] M. Enserink, «Science Publishing: Are You Ready to Become a Number?» Science, vol. 323, no. 5922, 2009, doi: 10.1126/science.323.5922.1662.

[10] N. Kaplan, «The norm of citation behavior», Am. Documentation, vol. 16. pp. 179—184, 1965.

[11] J. Bollen, H. Van de Sompel, A. Hagberg, and R. Chute, «A Principal Component Analysis of 39 Scientific Impact Measures», PLoS ONE, vol. 4, no. 6, p. e6022, 2009, doi: 10.1371/journal. pone.0006022.

[12] J. Bollen, H. Van de Sompel, A. Hagberg, L. Bettencourt, R. Chute, and L. Balakireva, «Clickstream Data Yields High-Resolution Maps of Science», PLoS ONE, vol. 4, no. 3, p. e4803, 2009, doi: 10.1371/journal.pone.0004803.
[13] S. J. Coles, J. G. Frey, M. B. Hursthouse, M. E. Light, A. J. Milsted, L. A. Carr, D. De Roure, C. J. Gutteridge, H. R. Mills, K. E. Meacham, M. Surridge, E. Lyon, R. Heery, M. Duke, and M. Day, «An e-science environment for service crystallography from submission to dissemination», J. Chem. Inf. Model., vol. 46, no. 3, 2006, doi: 10.1021/ci050362w.
[14] R. Bose and J. Frew, «Lineage retrieval for scientific data processing: a survey», ACM Comput. Surv. (CSUR), vol. 37, no. 1, pp. 1—28, 2005, doi: 10.1145/1057977.1057978.
[15] H. Van de Sompel, C. Lagoze, C. E. Nelson, S. Warner, R. Sanderson, and P. Johnston, «Adding eScience Publications to the Data Web», Proc. Linked Data on the Web 2009, Madrid.