DOI: 10.36535/0203-6460-2020-01-3

Связывание научных ресурсов: построение жизненного цикла их связи*

Мэтью МАЙЕРНИК

(Matthew MAYERNIK)

Национальный центр атмосферных исследований, Университетская корпорация атмосферных исследований, США

Научные ресурсы, включая публикации, программное обеспечение, массивы данных и инструменты, создаются итеративно и взаимосвязано. Управление взаимосвязями, существующими между такими ресурсами и среди них, является основным требованием для информационных систем. Однако практически многие научные ресурсы существуют в режиме онлайн в виде дискретных связей, отделенных от других ресурсов, с которыми они тесно связаны. Надежная система для объединения научных ресурсов на широкой и устойчивой основе должна будет ориентироваться на комплекс сложных и взаимосвязанных потребностей. В статье представлены результаты и выводы трех различных проектов, которые направлены на поддержку более прочных связей между научными ресурсами. В ходе обсуждения подробно рассматриваются ключевые технические и институциональные проблемы, которые во времени могут рассматриваться как «жизненный цикл отношений»: идентификация, подтверждение, характеристика и сохранение отношений. Цель статьи – помочь в руководстве новыми исследовательскими инициативами и оперативными службами, нацеленными на интеграцию информации о взаимоотношениях в научные исследования.

ВВЕДЕНИЕ

Многие научные ресурсы сегодня доступны онлайн. Журнальные статьи публикуются через платформы в режиме онлайн, отчеты и другая серая литература предоставляются через институциональные и общецелевые хранилища, массивы данных все больше архивируются в доступных сетевых хранилищах данных, а пакеты программного обеспечения широко распространяются через GitHub и иные средства обмена кодами. Безусловно, не все ресурсы предоставляются онлайн. Многочисленные хорошо известные социальные, культурные и технические факторы сдерживают распространение научно-исследовательских продуктов, а ученые в некоторых случаях обеспечивают конкурентное преимущество за счет максимизации их уникального доступа к новым данным, средствам или знанию [1-3]. Конечно, это также относится к случаю, когда предоставление онлайн доступа к чему-либо автоматически не гарантирует его пользу или завоевание понимания со стороны широкой аудитории [4]. Тем не менее, тенденция ясна и предполагает, что научные ресурсы уже сейчас доступны онлайн или будут все больше становиться доступными онлайн. Политическое давление, технические возможности и социальные нормы – все это подталкивает к такой направленности [5-7]. В рамках научной работы наличие доступных онлайн ресурсов дает широкие возможности. Научные сообщества извлекают пользу из повышающегося доступа к огромному числу ресурсов, а отдельные ученые получают выгоду от появляющихся преимуществ библиографических ссылок и чтения [8, 9].

Если ресурсы доступны онлайн, то, естественно, возникает вопрос: можем ли мы связать ресурсы вместе? В конце концов, связи (ссылки) — это то, из чего состоит весь Интернет. Публикации, программное обеспече-

* Перевод Mayernik M. Scholarly resource linking: Building out a "relationship life cycle"//Proceedings of 81 st Annual Meeting ASIS&T. — http://www.asist.org/wp-content/uploads/Final-81st-Annual-Meeting-Proceedings-1. pdf

ние, массивы данных и другие научные ресурсы создаются итеративными и взаимосвязанными. Взаимосвязи между научными ресурсами можно охарактеризовать как образующие важную цепь, в которой они обеспечивают существенное значение для подходов к обнаружению, использованию, управлению и сохранению ресурса [10, 11]. Управление такими взаимосвязями в принципе является ключевой составляющей (основной компонентой) информационных систем [12]. В самом деле многие системы информации и данных управляют и усиливают различные взаимосвязи, в частности, взаимосвязи между словарными терминами и структурами содержания [13]. Однако на практике большинство научных ресурсов существуют онлайн как дискретные связи. Исследование реестра хранилищ научно-исследовательских данных (registry of research data repositories, https://www.re3data.org, исследование проводилось 28 марта 2018 г.) обнаружило только 48 хранилищ из 2040 идентифицированных как принимающих «код источника» и 26 хранилищ как поддерживающих «применение программного обеспечения», некоторые хранилища имеют и то, и другое. Эти цифры, скорее всего, не являются точными, но отражают, насколько услуги по хранению данных и сопровождению программного обеспечения в хранилищах разъединены. Наравне с этим, научные статьи принадлежат системам и хранилищам, которые, как правило, не собирают массивы данных или программного обеспечения.

Таким образом, организация связи научных ресурсов требует подхода, который осуществляет навигацию между многими научными учреждениями и техническими системами. Установление и управление отношениями между информационными ресурсами стало общей темой в информатике и изучении технологий. Данная статья не стремится проанализировать всю релевантную литературу и инициативы в этой сфере. Полезные обзоры можно найти в других источниках [14-17]. Наоборот, статья представляет результаты и выводы трех различных проектов, направленных на организацию связи между данными и литературой. В ней подчеркиваются основные проблемы идентификации, подтверждения, характеристики и сохранения отношений между научными ресурсами и обсуждается, как эти проблемы различаются при взгляде во времени в будущее или прошлое. Целью статьи является помощь в руководстве новым исследованием и оперативными службами, нацеленными на интеграцию информации об отношениях в более полном объеме в научные исследования.

ИССЛЕДОВАНИЯ В СФЕРЕ СВЯЗЫВАНИЯ

В этом разделе представлены три отдельных проекта, реализованных в последние четыре года. Каждый проект отличался своей целью, масштабом и партнерами, но их общей темой было исследование подходов к перекрестному связыванию схожих научных ресурсов. Как отмечают авторы работы [16] в своей недавней ретроспективе, относящейся к серии научных ресурсов и проектов по интероперабельности инфраструктуры, из-за разнообразия проблем и вовлеченных участников «едва ли можно точно знать, как следует начать работу в направлении растущей интероперабельности». Указанные проекты следует рассматривать, как попытку дать старт множеству параллельных начинаний в поисках понимания и рассмотрения вероятного исследования проблем перекрестного связывания научных ресурсов.

Ниже отдельно обсуждается каждый проект. Каждый подраздел описывает соответствующие цели, подходы и релевантные действия проекта. Также определяются некоторые важные уроки, усвоенные через работу в каждом проекте. Следуя описаниям этих проектов, приводится междисциплинарное обсуждение, связывающее воедино точки зрения по каждому проекту в целях разработки широко применяемых выводов, сделанных на основе усилий по перекрестному связыванию научных ресурсов.

Перекрестное связывание от хранилища к хранилищу

Первый проект (январь 2015 г. – март 2016 г.) содержал экспериментальную попытку провести обмен связанной информацией между двумя хранилищами – одним хранилищем данных и одним хранилищем литературы. Цель состояла в том, чтобы позволить двум хранилищам напрямую взаимодействовать для обмена информацией о ссылках на ресурсы, которые имели известные отношения, но располагались и управлялись по отдельности, такие как, например, когда массивы данных, расположенные в хранилище А, использовались для производства публикаций, расположенных в хранилище В. Задумкой данного проекта была разработка процесса взаимообмена между двумя хранилищами, позволяющая исследователям, депонирующим ресурсы в одной системе, инициировать депонирование близких источников в другом хранилище. В идеале связи между близкими ресурсами, содержащимися в двух отдельных системах, могут быть обменены и видимы в соответствующих хранилищах. С точки зрения системы, целями были идентификация, охват, проверка и использование особенностей хранилища, позволяющих создавать, обменивать и поддерживать во времени связи между близкими ресурсами при небольших технических препятствиях и минимальных усилиях со стороны куратора хранилища. Полное от начала до конца применение этой идеи не было завершено, как обсуждается далее. Но процесс работы над проектом позволил нам исследовать основные потребности в том, как может быть достигнуто (получено) перекрестное связывание хранилищ. У нас также был семинар в конце проекта, на котором произошла более широкая дискуссия по его темам [17].

Подход к проекту состоял в: 1) установлении случаев использования и участников; 2) разработке технических потребностей для этих случаев использования и участников и 3) разработке системных функциональностей, которые могли бы удовлетворить потребности случаев использования и участников. Рис. 1 отражает четыре сценария организации связи между ресурсами, которые относятся к случаям использования и участникам, для связи документа с соответствующими ему данными. В сценариях #1-3 рис. 1 создатель контента (например, автор публикации и/или данных) может быть заинтересованным в депонировании двух новых ресурсов или депонировании нового ресурса, относящегося к депонированному ранее ресурсу. В сценарии # 4 кураторы хранилища могут быть заинтересованы в идентификации связей между ресурсами, уже являющимися частью существующих у них собраний.

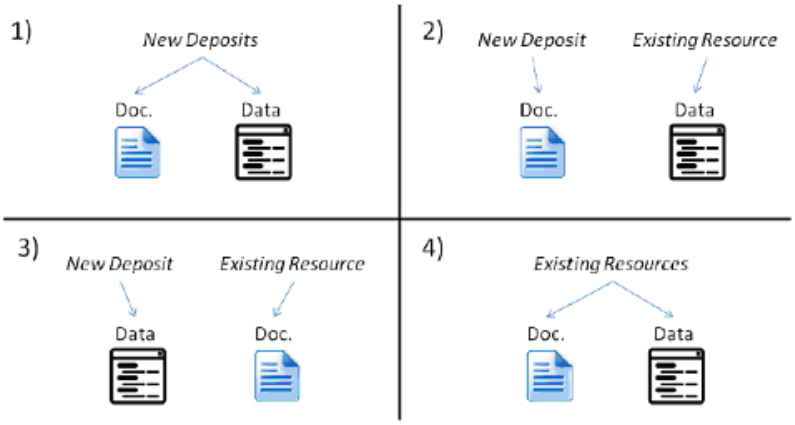

Рис. 1. Сценарии рабочих процессов организации связи между хранилищами

Эти сценарии представили ряд поддерживающих задач, включая: а) провайдеры массива данных могут загружать публикации в одно хранилище и депонировать массивы данных в другое (или наоборот); б) кураторы контента хранилища будут оповещены при попытке инициировать новые представления из внешней системы; в) авторы публикаций или данных будут иметь возможность обмениваться депонированными ресурсами с коллегами и соответствующим образом ссылаться на них; г) потребители результатов научных исследований (люди или машины) могут пользоваться хранилищами для поиска ресурсов и их взаимосвязей. При подразделении этих задач на подзадачи и другие потоки данных были сформированы следующие технические потребности в качестве основы для установления и поддержания надежных связей между хранилищами связанных ресурсов:

• Механизм оповещения для отправки информации о связанных ресурсах между хранилищами;

• Модель данных для обмена метаданными, созданная в целях:

выявления наследия метаданных (например, авторов или их мест работы, являющихся общими для разных ресурсов);

выявления типов взаимосвязей (например, массив данных соответствует публикации);

выявления во времени информации, необходимой для поддержки взаимосвязей между ресурсами в качестве появляющихся обновлений (например, создание нового массива данных или создание новой публикации на основе существующего массива данных).

• Эффективные интерфейсы депонирования ресурса, допускающие простую декларацию перекрестных связей между данными и публикациями в хранилище;

• Эффективный показ интерфейса, делающий перекрестные связи в хранилище прозрачными для пользователя.

Изучался сервис SHARE Notify (https://share.osf.io) как средство оповещения, позволяющее хранилищам взаимодействовать и обмениваться между собой информацией о взаимосвязях. К моменту нашего исследования SHARE Notify разрабатывался Центром открытой науки как еще один сервис третьей стороны по созданию оповещений о «событиях» вокруг научных публикаций. Наш интерес в использовании SHARE Notify заключался в отправке оповещений о «событиях» взаимосвязи между нашими хранилищами, при этом данный сервис выступал в качестве посредника между двумя хранилищами. Обращение к стороннему промежуточному сервису, такому как SHARE Notify, имело место по причине того, что он исключал потребность каждого хранилища напрямую связываться с другим. Хотя наш случай использования ставил своей целью связать два хранилища, для нас представляло интерес средство распространения информации о вазаимосвязях нескольким потенциальным партнерам. Таким образом, нам нужен был подход, который, вероятно, мог определить масштаб поддержки обмена информацией о взаимосвязях между несколькими хранилищами. Нам хотелось найти третью сторону, которая могла бы обеспечить основной реестр выпуска научных «событий» через обратную связь метаданных. У сервиса SHARE Notify имелся потенциал удовлетворить наши желания, хотя наш случай использования не был для него приоритетным.

Проводились пробные эксперименты по использованию сервиса SHARE Notify, но в конечном счете по ряду причин не были полностью выполнены рабочие процессы во время реализации нашего проекта. Во-первых, сервис SHARE Notify сам был на стадии разработки и итераций. Поэтому он представлял подвижную цель наших потребностей. Модель данных сервиса SHARE Notify и интерфейсы прикладного программирования не были завершены на момент старта нашего проекта. Это основной урок, полученный из данного проекта. Стабильность стороннего сервиса оповещений является внешним фактором, который может воздействовать на надежность локальных рабочих процессов.

Другим, усвоенным из проекта и не связанным с сервисом SHARE Notify, уроком было то, что подобная нашей попытка должна быть агностической относительно идентификации схем, используемых для установления ресурсов. Некоторые из изучаемых нами технических рабочих процессов требовали определения массивов через DOI, в особенности, чтобы использовать преимущество хранилищ метаданных, принятое агентствами регистрации DOI (например, CrossRef и

DataCite). Однако многие ресурсы в двух задействованных в этом проекте хранилищах не были (и предположительно не должны были быть) приписаны к DOI.

Отслеживание связей ресурсов через библиографические ссылки

Описываемый здесь второй проект (с 2014 г. по настоящее время) фокусировался на разработке и средствах оценки автоматического отслеживания научных инфраструктур в научной литературе с помощью постоянных идентификаторов (PID) цитирования. Относительно связи интерес состоял в отслеживании связей между статьями и соответствующими ресурсами, такими как массивы данных, которые использовались для производства статей. Стимулом данного проекта был рост интереса к «цитированию данных», т. е. присвоение данных постоянных идентификаторов, таких как DOI, данным с целью стимуляции их цитируемости и возможности отслеживания наравне с традиционной научной литературой [18, 19]. Однако для нас интерес заключался в более широком изучении пользы постоянной идентификации научных ресурсов, включая научное оборудование (или набор инструментов), пакеты научного программного обеспечения, вычислительные системы и коммуникационные сети. Воспользуемся многозначным термином «научные инфраструктуры», чтобы выделить эту широкую группу. Поскольку присвоение и использование PID научно-исследовательским инфраструктурам является относительно новой разработкой, было проведено очень мало оценок, которые системно изучали бы воздействие таких идентификаторов, присвоенных подобным инфраструктурам. Таким образом, цели нашего проекта состояли из: 1) развития понимания того, как методологически и последовательно анализировать научное влияние исследовательских инфраструктур, и 2) разработки автоматизированных технологий, дающих возможность отслеживать научные инфраструктуры с целью более эффективной работы. Отдельные вопросы исследования касались, например, того, как со временем изменялись ссылки на научно-исследовательские инфраструктуры относительно присвоения PID данным, программному обеспечению и иным компонентам научных инфраструктур.

Первой задачей данного проекта было проведение оценки на основе ситуационного анализа того, насколько часто исследовательские инфраструктуры идентифицируются и отражаются в научной литературе через постоянные идентификаторы (PID - Persistent IDentifiers) цитирования. В исследовании [20] цитирования и библиографические ссылки на четыре ресурса – два массива данных, пакет программного обеспечения и чрезвычайная вычислительная легкость – собирались вручную с помощью поиска в сервисе Google Scholar и анализировались, чтобы оценить способы, которыми ссылаются на ресурсы пользующиеся ими ученые, через приведение характеристики того, как часто на данные ресурсы делались ссылки в статьях с помощью полнотекстового описания, упомянутого в благодарности или эксплицитно цитируемого в библиографическом списке. Результаты этого исследования показали, как на самом деле использовались постоянные, присвоенные четырем анализируемым ресурсам, идентификаторы в ссылках в опубликованных статьях на основе роста. Но здесь не было постоянной последовательной модели во всех четырех ситуационных анализах относительно того, как этот рост управлялся. Наравне с этим анализ обнаружил, что практики библиографических ссылок со временем менялись, а степень этих изменений варьируется в значительной мере от ресурса к ресурсу. Основным выводом из этих результатов является то, что изменение установленных практик библиографических ссылок и признание массива данных будут вероятно более трудными, чем создание новых практик библиографических ссылок для иного рода продуктов, таких как возможности программного обеспечения или вычислений.

Вышеупомянутое исследование проводилось в основном вручную. Второй задачей в рамках данного проекта была разработка вычислительных алгоритмов/методов с целью облегчения оценки подобного рода проекта. Разработка алгоритмов и методов отслеживания придерживалась типового подхода машинного обучения, сконцентрированного на трех направлениях: 1) сбор публикаций-кандидатов для классификации, 2) разработка экспериментальной методологии по классификации и 3) автоматизированный подход к проведению классификации и анализа документа. Данный подход признается многообещающим благодаря первоначальной проверке классификаторов документов на способность правильно определять, действительно ли документ из тестового массива использует интересующую нас вычислительную способность, основанную на характеристиках метаданных и полного текста документа. Ограничение этого исследования, таким образом, намного больше и состоит в том, что основано на относительно небольшом массиве документов. Пробный массив (исследуемый вручную и присваивающий обозначения) включает около 300 документов, а тестовый массив (присваивающий обозначения, но еще не изученный или обработанный) содержит около 120 документов. Эти цифры далеки от идеальных в исследованиях по машинному обучению, которые часто используют тысячи или даже миллионы документов.

Основным выводом этого проекта является то, что трудно предсказать масштаб человеческого опыта, но он (опыт) крайне важен для идентификации и подтверждения взаимосвязей в тех случаях, где нет возможности использовать подходы на основе вычислений. Машинная автоматизация легко масштабируется, но гарантия и точность измерения автоматизированными средствами сбора связей очень непросты.

Оба исследования, предпринятые в ходе этого проекта, столкнулись с одинаковой проблемой, главным образом той, что сбор большого числа документов, на которые опирается анализ любых тенденций на основе литературы, весьма труден вне узких областей, таких как биомедицина (PubMed), физика (arXiv.org) и астрономия (Astronomical Data Service), где масса литературы хранится в публично доступных и машиночитаемых системах. В других областях, включая сферы интереса данного проекта, литература распределена по многим издательским платформам, не допускающим любую всеохватывающую возможность доступности через ЭВМ. Google Scholar, хотя и имеет достаточно большой охват областей, не позволяет какое-либо значительное автоматическое извлечение литературы. Данные методологические трудности служат общей предпосылкой для многих исследований, нацеленных на выполнение и анализ метрик влияния научно-исследовательских инфраструктур [18]. Эти ограничивающие масштаб про-

блемы авторского права, наравне с редакционной политикой журналов, издательскими платформами и разницей в форматах статей, представляют иные неконтролируемые факторы для автоматизации такого вида исследований.

EarthCollab – связывание по семантической сети

Третий проект, представляющий интерес для данной статьи и называемый EarthCollab (2014 – 2018 гг.), фокусировался на использовании семантической сети и технологий связанных данных, чтобы облегчить координацию и организацию сложных научных проектов и их продуктов. С технологической точки зрения цель проекта заключалась в разработке информационных систем, демонстрирующих, как геонауки могут усилить связанные данные для производства более когерентных методов обнаружения информации и данных в крупных междисциплинарных проектах и виртуальных организациях. Мотивацией к этому проекту служило улучшение обнаружения и обмена информацией в целях оказания содействия исследовательскому и научному сотрудничеству, стимуляции ученых к более простому поиску людей, организаций и научно-исследовательских ресурсов, отвечающих их занятости. Пакет программного обеспечения семантической сети VIVO (http://vivoweb.org/) был выбран из-за его опоры на модель данных, сконцентрированную вокруг сети и сфокусированную на представлении взаимосвязей между ее объектами [21]. Рис. 2 отражает, как взаимосвязи «многие-ко-многим» (many-to-many) существуют среди (и внутри) научных ресурсов, проектов, людей и организаций.

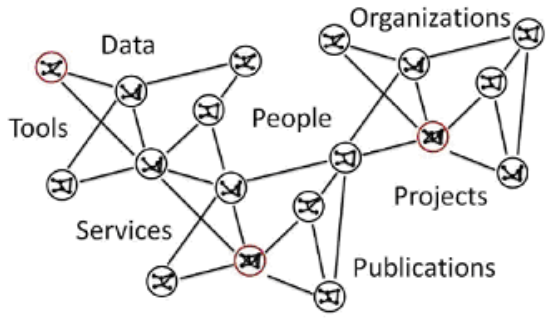

Рис. 2. Сетевая наука

Проект использовал VIVO для обеспечения ученых сетевыми интерфейсами в рамках двух целевых научно-исследовательских областей, чтобы они изучали людей, публикации, платформы и массивы данных внутри их соответствующих сообществ. Эти системы, названные «Connect UNAVCO» (https://connect.unavco.org) и «Arctic Data Connects» (http://vivo.eol.ucar.edu), более года были открыты для публичного пользования. В иллюстрации того, какого рода информация находилась в этих системах: например, по состоянию на 1 апреля 2018 г. система «Connect UNAVCO» содержала записи почти 6000 научных документов, 4307 массивов данных, 3841 научно-исследовательский сайт, 228 грантов, сведения относительно 803 человек и 381 организации. Система «Arctic Data Connects» с более пристальным вниманием на ситуационный анализ по выбору арктических научно-исследовательских проектов содержала записи 354 массивов данных, 26 грантов, сведения о 146 ученых и 53 организациях. Информация в этих системах появлялась из сочетания существующих баз данных метаданных и недавно созданных метаданных. Ресурсы представлены через модель данных семантической сети с помощью использования многочисленных онтологий [22]. В семантической сети онтологии применяются для определения типов объектов (классов) и отношений между ними (свойств). Все, что угодно, может быть представлено в качестве объекта первого порядка, как декларируется в соответствующих онтологиях. В результате получается сетевой набор информации, в котором каждые массивы данных, публикации, научные инструменты, исследовательские проекты и сайты представлены в модели данных отдельными объектами, имеющими определенный тип отношений с другим объектом. В частности, программное обеспечение VIVO показывает сетевую страницу каждого объекта, предоставляющую информацию об объекте вместе со связями с другими объектами, выраженными через эксплицитно заявленные отношения. Например, сетевая страница массива данных внутри одной из систем покажет все известные ссылки на соответствующие публикации, организации, гранты, создателей и инструменты.

Другой составляющей этого проекта было осуществление разработки программного обеспечения с целью добавить возможности пакету программного обеспечения VIVO. В течение первых двух лет ведения проекта был разработан прототип подхода «перекрестного связывания» для обмена информацией на примерах VIVO. Мотивацией для данной работы являлось то, что участники многих научно-исследовательских проектов находятся в разных организациях. Некоторые из этих организаций уже пользуются VIVO, чтобы управлять информационными профилями преподавателей и персонала. Новая особенность перекрестного связывания была развита в целях поощрения обмена информацией об отдельных людях или объектах, общих для разных примеров VIVO. Задача новой особенности – сократить дупликации информации между системами и стимулировать распределения авторитетной информации об отдельном объекте, такой как в случае наличия у данного лица VIVO профилей в разных организациях. Возможность перекрестного связывания была развернута в рамках системы «Connect UNAVCO» и передана базе кода открытого источника ядра VIVO.

Основным уроком, усвоенным из проекта EarthCollab, было представление взаимосвязей об использовании и переработке онтологий между центрами научных ресурсов. Для упрощения интеграции и обмена данными в геонаучном сообществе нашей целью в рамках данного проекта была, насколько это возможно, переработка существующих онтологий в ходе реализации проекта онтологий и сетевых применений. Переработка существующих онтологий оказалась не такой легкой, как предполагалось первоначально. Трудность в работе по использованию онтологий была в концептуальном моделировании, в котором определены и отображены представляющие интерес ключевые объекты и отношения. Только после этого можно делать следующий шаг — главным образом поиск релевантных существующих онтологий, представляющих вашу концептуальную мо-

дель, что не является тривиальной задачей. В нашем проекте ни одна онтология не поддерживала потребности нашего проекта. Таким образом, сочетались компоненты пары разных онтологий и создавались расширения онтологии потребителей в качестве необходимости устранения пробелов [22].

ОБСУЖДЕНИЕ

Все, описанные выше, три проекта фокусируются на проблемах, касающихся связывания вместе научных ресурсов в сетевой среде. Проекты велись параллельно с некоторым перекрытием участников, но без прямого повтора в полученных результатах. Принципы работы проектов, однако, перекрывались. Вместе они обеспечивают более широкий взгляд на компоненты и участников попыток перекрестного связывания научных ресурсов, чем каждый в отдельности. Такого рода подход «параллельных усилий» оказался весьма полезен в ситуациях, когда оптимальный результат (или путь к результату) не ясен с самого начала [23].

В этом разделе представлено всестороннее обсуждение потребностей перекрестного связывания научных ресурсов, основанное на знании и опыте, полученном при одновременной работе над многими релевантными проектами. Сначала категоризируем по типу наиболее заметные потребности и связанные в пространстве этой проблемы задачи, затем опишем, как по-разному выглядят эти потребности и задачи при взгляде в прошлое и будущее во времени.

Потребности перекрестного связывания научных ресурсов

Надежная система связывания научных ресурсов в широкой и устойчивой манере должна будет преодолеть ряд сложных и взаимосвязанных требований. Настолько, насколько научные данные и создание, управление и сохранение программного обеспечения могут быть характеризованы через модель жизненного цикла [24, 25], можно характеризовать потребности инфраструктуры связывания научных ресурсов через «жизненный цикл их связи», включающий потребность в идентификации, подтверждении, характеристике и сохранении информационных отношений и связей. Каждая из этих областей обсуждается ниже.

Идентификация отношений. Отношения между научными ресурсами, столь полезные для обнаружения, понимания и применения ресурса, должны быть декларированы кем-то или каким-то объектом. Откуда могут (или должны) происходить декларации этих отношений? Отношения, не декларируемые явно, иногда могут определяться вычислительными процессами, иногда полагаться на статистические измерения идентификации скрытых отношений между отдельными объектами или словарными терминами, включенными в текст [26]. Но эти вычислительные техники, как правило, более успешны в особых случаях применения и трудны для обобщения. Другой очевидный источник информации об отношениях находится в создателе связанных ресурсов. Банальностью поколения метаданных для научных ресурсов является то, что ученые наилучшим образом позиционированы для предоставления метаданных о своих собственных ресурсах с учетом их непосредственного знания о том, как эти ресурсы были созданы и как использовались. Но попытки собрать информацию о взаимосвязях между данными, программным обеспечением, публикациями и иными научными ресурсами напрямую от ученых также наталкиваются на известные проблемы создания метаданных, главным образом те, что ученые проявляют мало инициативы по отношению к четкому описанию этих взаимосвязей и что знание отдельных отношений может быстро снижаться и распределяться между командами участников [27, 28]. Сбор информации об отношениях непосредственно от ученых, таким образом, получается трудной задачей, которая плохо поддается масштабированию из-за распределения в сети участников научного сообщества. Другим потенциальным источником информации об отношениях является опубликованная литература. Как обнаружилось в ходе нашего проекта *Tracing identifiers,* цитирования данных и программного обеспечения в самом деле формально все больше приводятся в ссылках, тем самым делая вероятно возможным автоматизированный анализ цитирований. Но наш проект также обнаружил, что временной интервал для укоренения этих новых цитирований будет очень долгим. Таким образом, идентификация отношений является важным ограничителем роста любой попытки по развитию подходов на основе связывания научных ресурсов. Эта задача также имеет бинаправленную составляющую, главным образом такую, что даже если идентификация отношений становится более прямолинейной, то как пропагандируются эти отношения? Иными словами, если я знаю о связи между моим и вашим ресурсами, как вы узнаете об этой связи? Как декларация моей связи распространится на вас? Это условие подогревало наш интерес в обслуживании уведомления о взаимосвязи в рамках нашего проекта перекрестного связывания хранилищ. Сегодня нет всестороннего агрегатора информации об отношениях, который мог бы служить в качестве подобной общей службы уведомления. В отдельном случае связывания данных с литературой, подход Scholix и близкая служба взаимосвязи данных с литературой (DLI), способны обслуживать основные роли централизованного стороннего агрегатора деклараций взаимосвязей [29, 30].

Подтверждение отношений. Подтверждение представляет следующую задачу и условие. Какой бы метод не использовался в идентификации отношений между научными ресурсами, декларации отношений нуждаются в подтверждении через некий процесс, который опять может основываться на человеческом или машинном труде. Основной проблемой процесса подтверждения является простое подтверждение относительно того, какие объекты соотносятся. Цифровые объекты рассеивают границы и могут изменяться со временем. Многие подходы по идентификации взаимосвязи полагаются на использование постоянных идентификаторов, таких как DOI, чтобы гарантировать постоянство и фиксированность. Тем не менее, технически PID действуют как локаторы ресурсов, к которым они приписаны, а не как идентификаторы этих ресурсов [31, 32]. Постоянные идентификаторы (PID) также приписываются на различных уровнях градации разного рода ресурсов, и нет повсеместно принятых правил управления решениями относительно того, как PID должны быть приписаны сложным цифровым объектам [33]. На практике использование PID не всестороннее, как обнаружил наш про-

ект *Tracing identifiers*. Многие библиографические ссылки на данные, программное обеспечение и иные источники все еще происходят через неформальные благодарности или внутритекстовые ссылки. Подтверждение деклараций отношений также зависит от понятия авторитетного источника информации об отношениях. Является ли информация об отношениях, полученная от создателя ресурса или из опубликованных статей, более авторитетной, чем информация из других источников? Служба DLI, отмеченная в предыдущем абзаце, предоставляет информацию об организациях, публикующих декларации об отношениях. Но должны ли все провайдеры рассматриваться авторитетными источниками информации о взаимосвязях? Механизмы обеспечения доверия в декларациях об отношениях являются основой любого подхода подтверждения отношений.

Характеристика отношений. Для многих целей основной акт декларации о существовании взаимосвязи не является очень полезным. Больше информации о типе отношений также можно собрать. Иными словами, использование взаимосвязи часто требует дополнительного описания самой взаимосвязи. Подход семантической сети основан на этой возможности, а именно на том, что отношения должны быть обозначены как особые типы наименований. Многие онтологии и словари создавались, чтобы определять особые типы отношений, которые могут встречаться между научными ресурсами [17], но эти типологии отношений очень разнообразны и непостоянны. В нашем проекте EarthCollab внимание сосредотачивалось на переработке существующих онтологий таким образом, чтобы представлять объекты и их отношения интероперабельно. Это кажется более сложным, чем предполагалось заранее, в силу того факта, что многие онтологии моделировали одинаковые типы ресурсов (например, массивы данных) и отношения (например, отношения ссылок) по-разному [22]. Дополнительной трудностью характеристики отношений является то, что разные применения информации об отношениях могут нуждаться в различных уровнях описания объектов при любом типе отношений. Подход Scholix, например, определяет очень простую модель данных для представления отношений [30]. Что касается других применений, включая наш проект EarthCollab самая простая модель данных не поддерживала цели обнаружения и понимания данных. Общий вопрос, касающийся характеристики отношений, состоит в том, что модели данных и схемы метаданных не всегда включают способы представления информации об отношениях или они имеют разные требования к описанию отношений. Модель данных SHARE Notify, например, не имела явного способа представления отношений в то время, когда изучался потенциал ее использования как службы уведомления об отношениях в рамках нашего проекта перекрестного связывания хранилищ.

Сохранение отношений. «Связывание таблиц объектов» является самой очевидной проблемой сохранения отношений между научными ресурсами на основе сети. Сетевые сайты изменяют URL или неожиданно исчезают, вызывая каскад ошибок любых ссылок, указывающих на сайт, который перестал существовать. Эта проблема представляет важный вопрос для проектов, которые стремятся использовать научную литературу как источник информации об отношениях. Постоянные идентификаторы (PID) снова являются решением этого вопроса. Все системы PID работают через перенаправляющие серверы. Поддержка перенаправлений представляет институциональную проблему, поскольку является технической. Организации, которые регистрируют DOI, требуются, например, для поддержки резолюции их идентификаторов и связанных языковых страниц. Кроме этого, проблемы сохранения затрагивают уровень описания, необходимый для понимания отношений с течением времени и /или по мере изменения во времени сообществ пользователей ресурсов и их отношений. Процессы, связанные с написанием и задокументированием отчета Американской национальной комиссии по изменению климата, например, межорганизационного отчета–соглашения по борьбе с изменением климата, значительно изменились в последнее время, преследуя цель обеспечить высоко структурированную информацию об отношениях для иллюстрации особых связей между научными потребностями и соответствующими данными и научно-исследовательскими статьями [34]. Авторы работы [35] приводят другой наглядный пример того, как потребности пользователя и потребности документации могут со временем значительно смещать акцент даже в одном и том же научном ресурсе. Следует отметить, что в двух приведенных примерах создание и курирование этой информации входило в ответственность соответствующего персонала. Сохранение понимания отношений является, таким образом, продолжающимся процессом, который может требовать серьезного опыта. Как часть потребности этого сохранения, подходы, подобные Scholix, и системы, такие как служба DLI, должны знать о своих собственных моделях поддержки, чтобы порождать доверие потенциальных пользователей.

Взгляд во времени на прошлое и будущее

Вопросы, связанные с идентификацией отношений, подтверждением, характеристикой и сохранением, представляют разные вызовы и потребности при взгляде во времени в прошлое и будущее. Рабочие процессы проекта перекрестного связывания хранилищ, изображенные на рис. 1, дают об этом четкое представление. Смотря в будущее во времени, вызовы концентрируются на том, как выстроить отношение декларация/идентификация в качестве установившейся и надежной части научного издательства, архивирования данных и программного обеспечения и технологий хранилища. Как отмечают авторы [36], «эффективный охват данных, включая источник и метаданные, легче всего осуществляется на начальном этапе рабочего процесса» [36, с. 17]. В сценарии, в котором ресурсы созданы заново и депонированы в системах хранилищ, представляется возможным сбор информации об отношениях, по мере их производства.

Смотря назад в прошлое, связывание научных ресурсов включает поиск доступной литературы (и другой документации) вместе с потенциальным непосредственным запросом создателей научных ресурсов относительно отношений, существующих между имеющимися ресурсами. Поиск литературы может быть ручной или автоматизированной, как в нашем проекте *Tracing identifiers*, с предполагаемыми достоинствами и недостатками любого подхода. Ограничения авторского права и лицензий явно уменьшают современные усилия по поиску литературы, за исключением академических специальностей, где хорошо организованы модели открытого доступа.

Таблица

Техническая и институциональная работа, необходимая для поддержки связывания научных ресурсов

	Взгляд во времени в прошлое	**Взгляд во времени в будущее**
Идентификация взаимосвязей	Техническая работа – подходы к извлечению данных для выборки ссылок на основе PID из опубликованной литературы Институциональная работа – взаимодействие с издателями и финансирующими организациями по открытой научной литературе для получения всестороннего анализа данных	Техническая работа – разработка агрегаторов взаимосвязей и связанные открытые сетевые службы Институциональная работа – 1. продвижение соответствующего использования PID. 2. разработка и адаптация подходов сообщества к распределению взаимосвязей (например, Scholix)
Подтверждение взаимосвязей	Техническая работа – подходы к извлечению данных для выборки неформальных ссылок на основе PID из опубликованной литературы с помощью оценки доверия Институциональная работа – взаимодействие с издателями и финансирующими организациями	Техническая работа –разработка/обновление моделей данных для поддержки взаимосвязи источника и достоверных утверждений Институциональная работа – разработка организационных сетей доверия для декларации отношений
Характеристика взаимосвязей	Техническая работа – соответствие нестандартным декларациям отношений в сообществе онтологий и схемах метаданных Институциональная работа – сочетание описаний объекта и взаимосвязей в потребности целевых сообществ и/или применений	Техническая работа – разработка/обновление моделей данных и схем метаданных для постоянно представленных взаимосвязей и объектов связи Институциональная работа –координация соглашений по семантике взаимосвязей внутри определенных сообществ или отдельных применений
Сохранение взаимосвязей	Техническая работа – сочетание сканирования ссылок и средств сетевого архивирования Институциональная работа – курирование информации о взаимосвязях во времени, итеративно обновленные взаимосвязи как необходимость поддержать потребности пользователя	Техническая работа – разработка/адаптация пакетных средств, гарантирующих сохранность связей между ресурсами во времени Институциональная работа – 1. разработка подходов по поддержке для агрегаторов взаимосвязей. 2. курирование информации о взаимосвязях во времени

Прямой контакт и работа с создателями ресурса по установлению информации об отношениях в существующих ресурсах также имеет очевидные ограничения, и, вероятно, будет скорее всего происходить только в контексте особых проектов или применений при необходимости в данной информации. Например, в рамках нашего проекта EarthCollab изучались протоколы запросов ученых непосредственно по сбору информации о связях между данными и научными статьями, поскольку сами статьи, как правило, не содержат достаточной для нас информации, чтобы уверенно подтвердить сами отношения.

Таблица отражает различного рода работу, необходимую для поддержки широкого связывания научных ресурсов, фокусируясь на прошлом и будущем. Каждая ее ячейка далее разбивает потребности работы на технические и институциональные. Как обсуждалось ранее [37], курирование научных ресурсов содержит нечто большее, чем просто техническая работа. Институциональные факторы, такие как нормы и ожидания сообщества, доступность посредников в поддержке кураторской работы, стандарты разработок и адаптации и установленный порядок,– всё играет важную роль в определении успеха кураторских усилий. Таким образом, в табл. подчеркивается, что и технические, и институциональные разработки важны для получения надежных инфраструктур связывания научных ресурсов. Таблица, кажется иллюстративной, но не исчерпывающей. Отличные от авторской точки зрения работы, цитируемые в статье, представляют дополнительные возможности акцентировать внимание на имеющейся и будущей работе в указанной сфере.

ЗАКЛЮЧЕНИЕ

Попытки связать научные ресурсы надежными и устойчивыми способами сталкиваются с многими сложными проблемами, распространяющимися на технические и организационные факторы. В статье обсуждаются три проекта, каждый из которых сфокусирован на одном или нескольких аспектах этих проблем. В ней синтезируются основные результаты и уроки, усвоенные в ходе этих проектов, и представлены принципы работы

четырех основных сфер потребностей, главным образом, вопросов, касающихся потребности в идентификации, подтверждении, характеристике и сохранении информации о взаимосвязях между научными ресурсами. Также подчеркивается, как попытки построить связывающие инфраструктуры будущего сталкиваются с разными проблемами в отличие от инициатив по сбору и характеристике связей между уже существующими данными, программным обеспечением, публикациями и иными научными ресурсами.

Одновременный запуск разных проектов позволил нам использовать подход «параллельных усилий», чтобы понять, какого рода инициативы больше подходят для решения конкретных задач, относящихся к связыванию научных ресурсов. Удовлетворение локальных потребностей, например, желание связать две конкретные локальные системы, может быть найдено через ограниченную техническую трудность. Разработка перекрестных организационных и перекрестных институциональных решений по поддержке связей ресурсов, тем не менее, требует стандартов и координационной работы и на это потребуется длительное время. Ни одна техническая система или инфраструктура скорее всего не сможет обеспечить общее решение описанных в статье проблем из-за вовлеченности широкого спектра заинтересованных сообществ. Итерация и институциональная работа будут ключом к решению.

Информационные науки имеют опыт и научно-исследовательскую возможность стать крупными вкладчиками в предстоящие попытки связать научные ресурсы. Проблемы, затронутые в этой статье, касаются научной коммуникации, подходов к работе с метаданными, библиометрии, цифрового хранения, сетевой архитектуры,– которые все являются и исторически, и в настоящее время областями исследования в рамках дисциплин библиотековедения и информатики. Инфраструктуры связывания научных ресурсов должны будут найти способы усилить результаты и средства, относящиеся ко всем этим областям, чтобы успешно продвигаться вперед.

Благодарность. Хочу выразить признательность многочисленным сотрудникам Национального центра атмосферных исследований, Корнельского университета и UNAVCO, внесшим вклад в рассмотренные три проекта. Описанная в статье работа проводилась при поддержке грантов Национального фонда научных исследований США (гранты №№ 1449668, 1448480, 1440213 и 1440181). Национальный центр атмосферных исследований финансируется Национальным фондом научных исследований.

ЛИТЕРАТУРА

1. *Mitroff I.I.* Norms and counter-norms in a select group of the Apollo moon scientists: A case study of the ambivalence of scientists// American Sociological Review. — 1974. — Vol. 39, No. 4.— P. 579-595. — https://doi.org/10.2307/2094423.

2. *Fienberg S.E., Martin M.E., Straf M.L.* (Eds). Part I: Report of the Committee on National Statistics. In Sharing Research Data. — Washington, D.C.: National Academy Press, 1985. — http://www.nap.edu/catalog/2033/sharing-research-data

3. *Borgman C.L.* The conundrum of sharing research data// Journal of the American Society for Information Science and Technology.— 2012.— Vol. 63, No. 6. — P. 1059-1078. — https://doi.org/10.1002/asi.22634

4. *Mayernik M.S.* Open data: Accountability and transparency// Big Data & Society. —2017. — Vol. 4, No. 2.— https://doi.org/10.1177/2053951717718853

5. *Willinsky J.* The unacknowledged convergence of open source, open access, and open science//First Monday. — 2005. — Vol. 10, No. 8. — http://ojphi.org/ojs/index.php/fm/article/view/1265/1185

6. *Woelfle M., Olliaro P., Todd M.H.* Open science is a research accelerator// Nature Chemistry. — 2011.— Vol. 3, No. 10.— P. 745-748. — http://doi.org/10.1038/ nchem.114

7. *Kriesberg A., Huller K., Punzalan R., Parr C.* An analysis of federal policy on public access to scientific research data// Data Science Journal. — 2017.— Vol. 16: paper 27. — https://doi.org/10.5334/dsj-2017-027

8. *Hitchcock S.* The effect of open access and downloads ('hits') on citation impact: A bibliography of studies.— University of Southampton, 2013. — https://eprints.soton.ac.uk/354006/

9. *Piwowar H.A., Vision T.J.* Data reuse and the open data citation advantage// PeerJ. — 2013. — Vol. 1: e175. — https://doi.org/10.7717/peerj.175

10. *Van de Sompel H., Payette S., Erickson J., Lagoze C., Warner S.* Rethinking scholarly communication// D-Lib Magazine. —2004.— Vol. 10, No. 9. — https://doi.org/10.1045/september2004-vandesompel

11. *Pepe A., Mayernik M., Borgman C.L., Van de Sompel H.* From artifacts to aggregations: Modeling scientific life cycles on the semantic web// Journal of the American Society for Information Science and Technology.— 2010. — Vol.. 63, No. 3.— P. 567-582. —https://doi.org/10.1002/asi.21263

12. *Kent W.* Data and Reality: Basic Assumptions in Data Processing Reconsidered. — New York: North-Holland, 2018.

13. *Bean C.A., Green R.* (Eds). Relationships in the Organization of Knowledge. — Boston, MA: Kluwer, 2001.

14. *Borgman C.L.* Scholarship in the Digital Age: Information, Infrastructure, and the Internet. Cambridge, MA: MIT Press, 2007.

15. *Lagoze C.J.* Lost Identity: The Assimilation of Digital Libraries into the Web. Ph.D. diss. — Cornell University, 2010.

16. *Van de Sompel H., Nelson M. L.* Reminiscing about 15 years of interoperability efforts// D-Lib Magazine. — 2015. — Vol. 21, No. (11/12). — http://doi.org/10.1045/november2015-vandesompel

17. *Mayernik M.S., Phillips J., Nienhouse E.* Linking publications and data: Challenges, trends, and opportunities// D-Lib Magazine. — 2016. — Vol. 22, No. (5/6). — https://doi.org/10.1045/may2016-mayernik

18. *Mayernik M.S., Hart D.L., Maull K.E., Weber N.M.* Assessing and tracing the outcomes and impact of research infrastructures// Journal of the Association for Information Science and Technology. — 2017. — Vol. 68, No. 6. — P . 1341-1359. —https://doi.org/10.1002/asi.23721

19. *Silvello G.* Theory and practice of data citation// Journal of the Association for Information Science and Technology. — 2017. — Vol. 69, No. 1. — P. 6-20. — https://doi.org/10.1002/asi.23917

20. *Mayernik M.S., Maull K.E.* Assessing the uptake of persistent identifiers by research infrastructure users// PLoS ONE. — 2017. — Vol. 12, Nol. 4, e0175418. — https://doi.org/10.1371/journal.pone.0175418

21. *Borner K., Conlon M., Corson-Rikert J., Ding Y.* (Eds.). VIVO: A Semantic Approach to Scholarly Networking and Discovery. — San Rafael, CA: Morgan & Claypool, 2012.

22. *Mayernik M.S., Gross M.B., Corson-Rikert J., Daniels M.D., Johns E.M., Khan, H., Stott D.* Building geoscience Semantic Web applications using established ontologies// Data Science Journal.— 2016.—Vol. 15, article 11.— P. 1-10. — https://doi.org/10.5334/dsj-2016-011

23. *Lenfle S., Loch C.* Lost roots: How project management came to emphasize control over flexibility and novelty// California Management Review. — 2013.— Vol. 53, No. 1.— P. 32-55. — https://doi.org/10.1525/cmr.2010.53.1.32

24. *Carlson J.* The use of life cycle models in developing and supporting data services// J.M. Ray (Ed.), Research Data Management: Practical Strategies for Information Professionals (pp. 63-86). — West Lafayette, IN: Purdue Univ. Press, 2014.

25. *Lenhardt W.C., Ahalt S., Blanton B., Christopherson L., Idaszak R.* Data management lifecycle and software lifecycle management in the context of conducting science// Journal of Open Research Software.— 2014.— Vol. 2, No.1, e15. — https://doi.org/10.5334/jors.ax

26. *Sheth A., Ramakrishnan C., Thomas C.* Semantics for the Semantic Web: The implicit, the formal and the powerful// International Journal on Semantic Web and Information Systems. — 2005. — Vol. 1, No. 1. — P. 1-18. — https://doi.org/10.4018/jswis.2005010101

27. *Michener W.K., Brunt J.W., Helly J.J., Kirchner T.B., Stafford S.G.* Nongeospatial metadata for the ecological sciences// Ecological Applications — 1997. — Vol. 7, No. 1.— P. 330-342.

28. *Edwards P.N., Mayernik M.S., Batcheller A., Borgman C.L., Bowker G.C.* Science friction: Data, metadata, and collaboration in the interdisciplinary sciences// Social Studies of Science. — 2011. — Vol. 41, No. 5.— P. 667-690. — https://doi.org/10.1177/0306312711413314

29. *Burton A., et al.* The data-literature interlinking service: Towards a common infrastructure for sharing data-article links// Program. — 2017.— Vol. 51, No. 1.— P. 75-100. — https://doi.org/10.1108/PROG-06-2016-0048

30. Burton A., et al. The Scholix Framework for interoperability in data-literature information exchange// D-Lib Magazine. — 2017. — Vol. 23, No. (1/2). — https://doi.org/10.1045/ january2017-burton

31. *Thompson H.S.* What is a URI and why does it matter? // Ariadne. — 2010. — Vol. 65. — http://www.ariadne.ac.uk/issue65/thompson-hs/

32. *Duerr R., Downs R., Tilmes C., Barkstrom B., Lenhardt W.C., Glassy J., Bermudez L., et al.* On the utility of identification schemes for digital earth science data: An assessment and recommendations// Earth Science Informatics. — 2011, Vol. 4, No. 3. — P. 1-22. — https://doi.org/10.1007/s12145-011-0083-6

33. *Mayernik M. S.* Bridging data lifecycles: Tracking data use via data citation workshop report. NCAR Technical Note, NCAR/TN-494+PROC. — Boulder, CO: National Center for Atmospheric Research (NCAR), 2013.— https://doi.org/10.5065/D6PZ56TX

34. *Tilmes C., Fox P., Ma X., McGuinness D.L., Privette A.P., Smith A., … Zheng J.G.* Provenance representation for the National Climate Assessment in the Global Change Information System// IEEE Transactions on Geoscience and Remote Sensing. — 2011. — Vol. 51, No. 11. — P. 5160-5168. — https://doi.org/10.1109/tgrs.2013.2262179

35. *Baker K.S., Duerr R.E., Parsons M.A.* Scientific knowledge mobilization: Co-evolution of data products and designated communities// International Journal of Digital Curation. — 2015. — Vol. 10, No. 2.— P. 110-135. — http://doi.org/10.2218/ijdc.v10i2.346

36. *Marchionini G., Lee C.A., Bowden H., Lesk M.* (Eds.). Curating for Quality: Ensuring Data Quality to Enable New Science. Final Report: Invitational Workshop Sponsored by the National Science Foundation, Sept. 10-11, 2012, Arlington, VA. — 2012. — http://openscholar.mit.edu/sites/default/files/dept/files/altman2012-mitigating_threats_to_data_quality_ throughout_ the_curation_lifecycle.pdf

37. *Mayernik M.S.* Research data and metadata curation as institutional issues// Journal of the Association for Information Science and Technology. — 2016. — Vol. 67, No. 4. — P. 973-993. — https://doi.org/10.1002/asi.23425