Методические рекомендации по разработке репозиториев

ФОНД ПРЕЗИДЕНТСКИХ ГРАНТОВ

NEICON
ЭЛЕКТРОННАЯ ИНФОРМАЦИЯ

НАЦИОНАЛЬНЫЙ АГРЕГАТОР ОТКРЫТЫХ РЕПОЗИТОРИЕВ РОССИЙСКИХ УНИВЕРСИТЕТОВ

Москва
2018

Методические рекомендации по разработке репозиториев

Москва
2018

УДК 025.135:004.9
ББК 73я7
М 54

Перевод материалов Confederation of Open Access Repositories (COAR) выполнен Н. Г. Поповой, Я. Ю. Моисеенко и А. Л. Поповой

Редактор М. Е. Шварцман

Технический консультант А. Н. Борбунов

М 54 Методические рекомендации по разработке репозиториев / под ред. М. Е. Шварцмана. — М.: Ваше цифровое издательство, 2018. — 34 с.
ISBN 978-5-6040408-2-9

Настоящие рекомендации разработаны на основе отчета рабочей группы по развитию нового поколения репозиториев Конфедерации репозиториев открытого доступа (COAR), опубликованного 28 ноября 2017 г. по адресу https://www.coar-repositories.org/files/NGR-Final-Formatted-Report-cc.pdf. Предназначены для разработчиков репозиториев, аналитиков, сотрудников научных библиотек.

Выполнено с использованием гранта Президента Российской Федерации на развитие гражданского общества, предоставленного Фондом президентских грантов. Грант № 17-2-003857

УДК 025.135:004.9
ББК 73я7

ISBN 978-5-6040408-2-9

Содержание

Введение

Широкое распространение систем репозиториев в высших учебных заведениях и исследовательских институтах создает основу для распределенной, глобальной сетевой инфраструктуры, поддерживающей систему научных коммуникаций. В настоящее время существует довольно большое количество литературы, посвященной выбору программного обеспечения для репозитория, установки и настройки его. Из последних работ можно рекомендовать для изучения «Электронная библиотека: инструкция по установке. Рекомендации для библиотек по организации собственных репозиториев открытого доступа», составленную И. В. Бегтиным и А. С. Горбуновой и опубликованную в «Научном корреспонденте» (http://nauchkor.ru/pubs/elektronnaya-biblioteka-instruktsiya-po-ustanovke-5a37c2627966e11ea210792b). Тем не менее платформы репозиториев по-прежнему используют технологии и протоколы, разработанные почти двадцать лет назад, еще до бума Интернета, социальных сетей, семантической паутины и вездесущих мобильных устройств. По большей части именно это является причиной того, почему репозитории не полностью реализовали свой потенциал и функционируют в основном как пассивные получатели опубликованных результатов исследований своих пользователей. Для повышения ценности сети репозиториев нам необходимо расширить их функциональные возможности. В настоящих рекомендациях описаны подходы, которые могут быть применены проектировании современного репозитория, рассчитанного на активное взаимодействие с окружающей научно-информационной инфраструктурой. Возможно, далеко не все из описанных технологий могут быть использованы в настоящее время в силу ряда причин. Однако авторы не сомневаются, что всем проектировщикам репозиториев нужно знать о современных тенденциях и перспективах развития репозиториев.

Настоящие рекомендации разработаны на основе отчета рабочей группы по развитию нового поколения репозиториев Конфедерации репозиториев открытого доступа (COAR), опубликованно-

го 28 ноября 2017 г. по адресу https://www.coar-repositories.org/files/NGR-Final-Formatted-Report-cc.pdf

Современная система распространения исследовательских работ, в которой доминируют коммерческие издатели, далека от идеала. В экономическом смысле как цены подписок, так и размеры платы за подготовку статьи к публикации завышены и, вероятно, продолжат расти недопустимыми темпами. Кроме того, в международной издательской системе существует значительное неравенство с точки зрения как доступа, так и участия. Встроенные в систему стимулы, принуждающие исследователей публиковаться в традиционных издательских центрах, только укрепляют эти проблемы.

Мы считаем, что глобально распределенная сеть репозиториев может быть использована для создания более устойчивой системы обмена результатами научных исследований. Репозитории могут обеспечить доступ к результатам исследований всего мира, а также дать возможность каждому ученому и институту участвовать в глобальной сети научных исследований. Создание дополнительных сервисов, предназначенных для оценки показателей использования, рецензирования и общения исследователей между собой, повысит значимость и популярность репозиториев в научной среде.

Репозитории должны стать основой глобальной сетевой инфраструктуры, поддерживающей функционирование системы научных коммуникаций. В рамках данной инфраструктуры, управляемой самим научным сообществом, может быть разработан целый спектр дополнительных услуг для дальнейшего стимулирования исследований и внедрения инноваций.

При проектировании нового Репозитория следует учесть следующее:

- в его задачи может входить обеспечение доступа к разнообразным ресурсам, включая опубликованные статьи, пре-

принты, наборы данных, рабочие документы, изображения, программное обеспечение и т. д.;

- репозиторий должен быть ресурсоориентирован, то есть ресурсы должны быть основой всех сервисов и инфраструктуры;

- репозиторий является частью сети, и при проектировании должны быть запланированы связи между репозиториями для обмена данными, метаданными и данными пользовательской активности.

В данных рекомендациях описываются основные функции, на которые следует обратить особенное внимание при проектировании репозитория, а также технологии, стандарты и протоколы, которые будут способствовать разработке новых сервисов.

1. Основные принципы организации репозитория

Распределенный контроль

Распределенный контроль или управление научными ресурсами (препринтами, постпринтами, исследовательскими данными, вспомогательным программным обеспечением и т. д.) и научной инфраструктурой — важный принцип, на который нужно опираться при создании репозитория. В противном случае небольшой круг участников системы научных коммуникаций получит слишком большой контроль и может установить монополию. Распределенные сети более устойчивы и менее подвержены такому риску.

Инклюзивность

Различные учреждения и регионы имеют уникальные и специфические потребности и условия (например, язык, политика и приоритеты). При создании репозитория нужно учитывать различные нужды и условия разных регионов и отраслей науки.

Интеллектуальная открытость и доступность

Научные ресурсы по возможности должны быть представлены в удобном формате и быть доступны всем, что повысит их ценность и позволит широко применять в интересах как научного сообщества, так и всего социума.

Устойчивость

Учреждения и исследовательские организации должны стать основными участниками глобальной сети, способствуя долгосрочной устойчивости ресурсов.

Функциональная совместимость

Репозитории должны следовать общим направлениям развития, функциональным возможностям и стандартам, обеспечивающим взаимодействие между институтами и позволяющим им совместно использовать внешних поставщиков услуг.

2. Основные принципы проектирования репозитория

Фокус на ресурсах, а не на связанных с ними метаданных

Исторически сложилось, что технические решения были сосредоточены на метаданных, которые описывают научные ресурсы, а не на ресурсах как таковых. Подход, при котором научные ресурсы и их метаданные рассматриваются как равноправные веб-ресурсы со своими URI-адресами, позволяет выстраивать между ними эффективные взаимосвязи.

Прагматизм

При наличии выбора следует выбирать более простой подход. По возможности мы выбираем технологии, решения и примеры, которые уже широко используются. На практике это означает, что мы предпочитаем использовать стандартные веб-технологии, где это возможно.

Эволюция вместо революции

Предпочтительнее разрабатывать решения, корректируя существующее программное обеспечение и системы, которые уже широко используются во всем мире, чтобы эффективнее использовать существующую инфраструктуру.

Следование международным стандартам

Следует стремиться максимально использовать признанные конвенции и стандарты там, где это возможно, вместо использования более богатых, сложных и разнообразных своих собственных решений. Новые стандарты следует вводить только тогда, когда возникают конкретные и специфические потребности с целью свести к минимуму возможные сложности взаимодействия с внешними системами.

Взаимодействие с пользователями, где бы они ни находились

Вместо того чтобы постоянно просить пользователей покидать среду, в которой они работают, и работать с одной из наших систем, мы хотим интегрировать наши инструменты в среды и системы, в которых пользователи уже находятся.

3. Основные характеристики репозиториев

Репозиторий должен обеспечить доступ к разнообразным ресурсам, включая опубликованные статьи, препринты, наборы данных, рабочие документы, изображения, программное обеспечение и т. д. Созданный репозиторий должен иметь следующие характеристики.

Ресурсоориентированность

Репозиторий должен быть ресурсоориентирован, поскольку его ресурсы становятся центром сервисов и инфраструктуры. В глобальной сети репозиториев распределенные и разнообразные ресурсы находятся в открытом доступе и в настоящее время могут однозначно идентифицироваться с помощью уни-

фицированных (единообразных) идентификаторов ресурса Uniform Resource Identifier (URI), а не посредством неточных описательных метаданных. Каждый отдельный ресурс может быть использован отдельно или связан с другими ресурсами. Вся эта совокупность ресурсов составляет уровень контента, который служит основой для разработки дополнительных сервисов, таких как рецензирование, использование в социальных сетях, создание рекомендаций, учет использования и т. д. Таким образом, репозитории становятся важной составляющей в глобальной сети научных ресурсов.

Взаимосвязанность

Репозиторий является сетевым инструментом, изначально спроектированным для сетевого взаимодействия. Соединения между репозиториями устанавливаются путем внедрения двунаправленных ссылок в результате действия соответствующих сервисов, анализирующих метаданные и данные пользовательской активности, предоставленные репозиториями. Связи между ресурсами позволят создать распределенные репозитории, многоуровневые научные сети и станут катализатором создания эффективного союза научных коммуникаций и исследовательских инфраструктур, устраняя разобщение между местами, где мы занимаемся наукой, и местами, где мы публикуем результаты научных исследований.

Удобство для машинной обработки

Репозиторий должен быть ориентирован не только на пользователя-человека, но и на пользователя-робота, созданного для автоматизации процесса обмена информацией между репозиториями. Такой подход позволит разрабатывать широкий спектр глобальных сервисов использования репозиториев при уменьшении затрат на их создание и развитие. При этом не следует ограничиваться только предоставлением возможности поиска метаданных для внешних программ, нужно стараться предоставлять доступ к полному разнообразию своих ресурсов (полному тексту, методам навигации, информации об использовании и т. п.).

Активность

Для успешной работы администратор репозитория должен обеспечить постоянное внедрение новых версий, обновление данных и связь между ресурсами. Содержимое репозитория не должно быть статично, оно меняется со временем. Администратор репозитория не может находиться в пассивном ожидании получения информации, он должен сам её выявлять, добавлять и активно уведомлять связанные системы об изменениях в репозитории.

4. Функционал репозиториев

При проектировании репозитория следует предусмотреть возможность использования в нем следующих технологий:

1. применение унифицированных идентификаторов ресурса;
2. размещение информации об используемой лицензии;
3. навигация для облегчения процедуры поиска;
4. взаимодействие с пользователями (аннотирование, комментирование, рецензирование);
5. передача контента;
6. сервисы дискавери;
7. статистические метрики посещаемости ресурса;
8. идентификация пользователей;
9. аутентификация пользователей;
10. стандартизованные метрики использования ресурса;
11. архивирование.

4.1 Применение унифицированных идентификаторов ресурса

При проектировании репозитория нужно стремиться максимально использовать унифицированные идентификаторы всех

ресурсов и их частей для исключения неоднозначности при поиске и обеспечении постоянного доступа к ним. Так, для идентификации статей можно использовать DOI, для идентификации авторов — ORCID, а для ссылок на полный текст использовать не URL, который может меняться, а URI и соответствующий разрешитель (резолвер) ссылок.

При публикации ресурса в виде HTML страницы возникает проблема неоднозначности в определении того, чем является объект, на который ведет ссылка. Например, это может быть ссылка на постоянный идентификатор URI. Эта проблема может быть решена посредством использования таких типизированных HTTP-ссылок, структура которых включала бы соответствующий тип ссылки.

Технологические решения, стандарты, протоколы, поддерживающие такой функционал

- Можно рекомендовать использовать технологию **Signposting** (http://signposting.org). Это подход, который делает научный веб более дружественным для компьютеров. Он использует «типовые связи» (Link Relation Types, https://www.iana.org/assignments/link-relations/link-relations.xhtml) как средство обозначения наиболее часто встречающихся отношений между ресурсами и/или их частями.

- Использование указателей (Signposting) позволяет информировать программу, анализирующую HTML код страницы о характере ресурсов, к которым ведет ссылка. Использование указателей (signposting) позволяет автоматически обнаруживать разнообразные ресурсы, относящиеся к научному объекту, включая библиографические описания, унифицированные идентификаторы, лицензии, информацию об авторах или различные ресурсах, являющихся частью научного объекта.

4.2. Информация об используемой лицензии

В идеальных условиях доступ к научным объектам должен предоставляться без каких-либо ограничений на их дальнейшее применение. Однако в реальности всё несколько иначе и в большинстве случаев правообладатели регулируют условия использования их продукта. Такие правила должны быть ясно сформулированы для любого веб-ресурса, что является частью научного объекта, причем они должны легко обнаруживаться как человеком, так и машиной. Для пользователей это может быть достигнуто посредством закрепления за каждой лицензией легкоузнаваемого символа (логотипа). Для машин проблема может быть решена с помощью корректно структурированных гиперссылок, перенаправляющих пользователей на URI конкретной лицензии, регулирующей условия пользования научным продуктом. Если информация о лицензии будет встроена в сам ресурс, инструменты наподобие библиографических менеджеров смогут передавать ее своим пользователям для сохранения в их базе данных. Поисковые роботы, выполняющие задачу архивирования электронных ресурсов или интеллектуального анализа данных, будут действовать согласно ограничениям, наложенным лицензией, и самостоятельно принимать решение о возможности сбора данных и их дальнейшей обработки. Использование универсальных лицензий типа Creative Commons позволяет пользователям (как людям, так и машинам) быстро и легко определять, какие ограничения вменяются той или иной лицензией.

Технологические решения, стандарты, протоколы, поддерживающие такой функционал

- Простым и стандартизированным способом предоставления разрешения от индивидуальных и корпоративных правообладателей на использование их творческого продукта другими лицами являются открытые лицензии компании **Creative Commons**. В настоящее время уже существует обширное и постоянно растущее сообщество авторов, распространяющих свои работы по этой лицензии, и большое

количество ресурсов, которые можно копировать, распространять, редактировать, воспроизводить и дополнять без письменного разрешения от правообладателя, осуществляя все операции в рамках закона об авторском праве https://creativecommons.org/licenses/.

4.3. Навигация для облегчения процедуры поиска

Научный ресурс может быть представлен в сети целым набором объектов, каждый из которых имеет свой URL. Например, это может быть HTML-страница, файл научной статьи в PDF- или HTML-формате, один или несколько сопутствующих блоков данных, библиографическое описание ресурса в одном или нескольких форматах и т. д. В то время как человек может легко переключаться между этими ресурсами, распознавая их принадлежность к одному и тому же научному ресурсу, для машины это достаточно сложная задача. В некоторых случаях в репозиториях применяется перенаправление к библиографической информации, описывающей научный объект, с помощью гиперссылок с тегами, которые указывают формат цитируемого объекта: bibtex, RIS, DC и т. д. Библиографические менеджеры или поисковые роботы, находящиеся в процессе поиска или архивирования информации, не всегда способны легко обнаружить доступ к таким метаданным. Пытаясь решить поставленную задачу, этим инструментам приходится прибегать к эвристическим алгоритмам, специфичным для каждого репозитория. Кроме того, когда такие программы попадают не на основную страницу, а на сопутствующие ей ресурсы — будь это PDF-документ или пакет данных, — они не способны самостоятельно перейти от них к другим ресурсам, являющимся частью ресурса. Для того чтобы увеличить вероятность обнаружения таких ресурсов, их структура должна стать очевидной для программ-роботов. Эта проблема может быть решена использованием корректно структурированных гиперссылок (с правильно выстроенным соотношением между ними) и индикаторов формата. В этом случае появляется возможность связывания между

собой всех веб-ресурсов, что в совокупности составляют научный объект.

Технологические решения, стандарты, протоколы, поддерживающие такой функционал

- Использование технологии **Signposting** — см. п. 1 «Унифицированные идентификаторы ресурса».

4.4. Интерактивные возможности (аннотирование, комментирование и рецензирование)

Ценность репозитория будет больше, если в нем предоставлена возможность пользователям добавлять свою информацию к уже имеющимся ресурсам. Это может быть такая деятельность, как аннотирование, комментирование и экспертное рецензирование. Для любого пользователя необходимо иметь возможность комментировать и рецензировать научные работы коллег, причем комментарии и рецензии к ним должны быть доступны всем — таким образом становится возможна более объективная оценка качества работы. Для исследователя крайне важно иметь возможность объединять информацию из разных репозиториев, создавая общую картину проведенных исследований на основе сочетания разрозненных цифровых объектов. Для организации, финансирующей исследование, необходимо получать доступ к экспертным заключениям и наукометрическим показателям научного продукта, созданного отдельными авторами.

Функционал, с помощью которого эти виды деятельности могут осуществляться, не обязательно должен быть обеспечен самим репозиторием. Его разработка может быть доверена и стороннему производителю, который специализируется на создании сопутствующего контента. Поддерживая развитие сопутствующего контента таким образом, репозитории имеют хороший шанс стать центром научной коммуникации, а следовательно, способ-

ствовать научной дискуссии и коллективной работе. Для обеспечения этой возможности нужно добиться полной совместимости между репозиторием и такими сервисами сторонних производителей. При этом нужно обратить внимание на то, как репозиторий должен быть извещен о том, что такой контент был создан. Эти методы позволят репозиторию обнаруживать сопутствующий контент и встраивать его в собственную структуру. Чтобы безошибочно привязать сопутствующий контент к его создателю, чрезвычайно важен процесс и аутентификации пользователей (см. пп. 4.8 и 4.9 «Идентификация пользователей» и «Аутентификация пользователей»).

Технологические решения, стандарты, протоколы, поддерживающие такой функционал

- **Activity Stream 2.0** (https://www.w3.org/TR/activitystreams-core/) является удобным инструментом для описания взаимодействий между ресурсами, включая комментирование, выражение одобрения («лайки»), обмен информацией и др. Такого рода взаимодействия описываются JSON-LD и используют словарь Activity Stream 2.0 (https://www.w3.org/TR/activitystreams-vocabulary/). Несмотря на то что ресурс словаря нацелен на взаимодействие пользователей обычных социальных сетей, можно создавать его расширения для научных социальных сетей.

- **Web Annotation Model** и **Web Annotation Protocol** (https://www.w3.org/TR/annotation-model/ и https://www.w3.org/TR/annotation-protocol/) — это механизмы представления аннотаций (включая рецензии, комментарии и др.) и соответствующие протоколы по созданию и управлению ими. Аннотации представляются с использованием словаря RDF и могут быть преобразованы в JSON-LD. Протокол основан на HTTP и REST.

- **International Image Interoperability Framework** (IIIF, http://iiif.io/) представляет собой семью пользовательских интер-

фейсов (API), которые позволяют воспроизводить, делиться и обрабатывать изображения с целью аннотирования, описания, предоставления авторизованного доступа и др. Несмотря на то что данная технология предназначена для работы со специфичным контентом репозитория, ее можно считать хорошим примером того, как проявляется будущее поколение репозиториев.

Что касается технологий, которые позволяют проинформировать репозиторий о том, что был создан сопутствующий контент, а также технологических решений по встраиванию этой информации, см. п. 7 «Статистические метрики посещаемости ресурса».

4.5. Передача контента

Как уже писалось выше, репозиторий должен проектироваться на ресурсоцентрированной парадигме. Согласно этой парадигме научная информация не копируется в разных системах, но располагается в одном месте, а доступ к ней осуществляется с помощью гиперссылок. Несомненно, существуют ситуации, когда требуется копирование информации. В основном это связано с необходимостью повысить скорость передачи данных. При решении таких задач, как индексирование, интеллектуальный анализ информации и т. п., нужно обработать очень большое количество информации, и, если обрабатываемая информация находится в разных репозиториях, может не хватить производительности сети для обеспечения процессора данными. Хотя нужно отметить, что технологии распределенной обработки тоже существуют и вполне могут быть применены в ряде случаев.

Таким образом, репозитории должны предоставлять возможность переносить контент «по требованию» разных сервисов, для обеспечения бесперебойной работы программ обработки данных, архивирования и т. п. Передача контента «по требованию» позво-

ляет всем участникам процесса научной коммуникации получать доступ и передавать актуальный контент научных объектов. Когда действия по поиску и передаче информации производятся во внешние по отношению к репозиторию инфраструктуры, их менеджеры должны иметь возможность своевременно получать/передавать контент из репозитория, что подразумевает периодическую синхронизацию их ресурсов с постоянно пополняющимися ресурсами репозитория (вновь создаваемыми, обновляемыми или удаляемыми).

Технологические решения, стандарты, протоколы, поддерживающие такой функционал

- Из проверенных временем решений не стоит забывать про **OAI-PMH** (http://www.openarchives.org/pmh/), а из более современных технологий можно обратить внимание на следующие.

- **IPFS** (https://ipfs.io/) — многообещающий peer-to-peer протокол, разработчики которого стремятся сделать Всемирную сеть более безопасным, быстрым и открытым пространством. Система IPFS удобна в тех случаях, когда большому количеству пользователей, каждый из которых является узлом IPFS, необходимо поделиться большим количеством данных

- **ResourceSync** (http://www.openarchives.org/rs/toc) — это спецификация, основанная на SiteMaps, может быть использована менеджерами репозиториев для предоставления информации, которая позволяла бы сторонним пользователям синхронизировать их ресурсы с изменяющимися (создаваемыми, обновляемыми и удаляемыми) ресурсами репозитория на постоянной основе. Хотя файлы SiteMaps позволяют поисковикам находить метаданные и другие материалы репозитория, расширение ResourceSync предлагает дополнительные возможности, например возможность отслеживания только лишь внесенных изменений. Среди другого

удобного функционала — предоставление метаданных, связанных с процедурами синхронизации, и типизированные гиперссылки для дальнейшего поиска. ResourceSync может быть применен для синхронизации как контента, так и метаданных. Система использует формат SiteMaps XML.

- **SWORD** (Simple Web-Service Offering Repository Deposit) (http://swordapp.org/about/) — удобный протокол для массового перенесения контента из одного места в другое.

4.6. Участие в дискавери-сервисах

Возможность одновременного поиска научных ресурсов сразу в большом количестве репозиториев имеет огромное значение для становления статуса репозиториев как важных составляющих в научной коммуникации. Часто пользователю нужно находить интересующие материалы репозиториев через агрегаторы или другие поисковые сервисы, такие как BASE, CORE, OpenAIRE и другие.

Дискавери-поиск достаточно эффективен, причем и в тех случаях, когда необходим перенос контента. Чем в большем количестве дискавери-сервисов участвует репозиторий, тем выше вероятность того, что пользователь (или машина) эти ресурсы обнаружат. Таким образом, поддержка дискавери-поиска дает возможность ресурсам избежать ошибки, при которой тот или иной ресурс как будто бы «перестает существовать», если он не попадает в топ поисковых запросов.

Технологические решения, стандарты, протоколы, поддерживающие такой функционал

- **ResourceSync** (см. п. 5 «Передача контента»)
- Использование указателей (**Signposting**) (см. п. 1 «Унифицированные идентификаторы ресурса»)

- Карты сайтов (SiteMaps) широко используются веб-разработчиками для информирования поисковых программ о том, какие страницы на сайтах доступны для обнаружения поисковыми роботами. В простейшем виде SiteMap представляет собой XML-файл, который фиксирует URI для каждого доступного ресурса в совокупности со всеми доступными, имеющими к нему отношение метаданными в целях оптимизации поискового процесса (например, дата последнего изменения, частота изменений). Менеджеры репозиториев могут использовать SiteMaps как самый прямолинейный способ сообщить о ресурсах репозитория поисковым механизмам (https://www.sitemaps.org/).

4.7. Статистические метрики посещаемости ресурса

Репозитории должны в режиме реального времени отслеживать статистику всех событий (информацию об изменениях, дополнениях, комментировании, аннотировании, рецензировании, доступе к ресурсу, количестве скачиваний и т. д.), имеющих отношение к каждому объекту, хранимому в репозитории.

Авторы и читатели должны иметь возможность получать информацию об интересующих их ресурсах не только ретроспективно и посредством специальных запросов, но и в режиме реального времени. Для этого нужно ввести в эксплуатацию механизмы уведомлений. В зависимости от конкретного случая это может быть «точечное» уведомление, например отправка извещения автору о цитируемости его работы по электронной почте. В другом случае это могут быть уведомления, распространяемые по подписке. Например, пользователь, заинтересованный в получении информации о поступлении новых документов по определенной тематике, подписывается на соответствующий канал и регулярно получает списки новых поступлений по его тематике. Для этого дополнительные сервисы должны собирать информацию о пользовательской активности, анализировать ее и рассы-

лать на ее основе уведомления. Возможно использовать автоматизированные системы рекомендаций, которые, основываясь на прошлых (в том числе анонимных) запросах пользователей, сами предлагают пользователям подходящие им объекты, находящихся в репозиториях. Для того чтобы достичь такого уровня функциональности, в научной среде необходима уникальная идентификация всех научных объектов и акторов научной коммуникации (авторов, рецензентов, научных институтов и т. п.).

Технологические решения, стандарты, протоколы, поддерживающие такой функционал

- **Activity Streams 2.0** (см. п. 4 «Интерактивные возможности (аннотирование, комментирование, рецензирование»)

- **Linked Data Notifications** (https://www.w3.org/TR/ldn/) — это протокол уведомлений, в соответствии с которым любой ресурс может формировать и поддерживать список изменений, куда будут направляться уведомления об всех изменениях данного ресурса. Например, приложения, выполняющие задачи аннотирования, комментирования или рецензирования, могут отправить уведомление о выполненной операции. Уведомление передается с помощью JSON-LD и использует словарь Stream Streams 2.0. Это позволит размещать информацию об изменениях в списке изменений либо представлять их в агрегированном виде для дальнейшего использования машинами с помощью WebSub (см. далее).

- **ResourceSync Change Notifications** (http://www.openarchives.org/rs/notification) является протоколом публикации/подписки, созданным на основе WebSub и предназначенным для рассылки уведомлений о любых обновлениях ресурсов репозитория (создание/обновление/удаление) подписчикам. ResourceSync Change Notifications могут использоваться для поиска и синхронизации как контента, так и метаданных. Протокол основан на формате SiteMaps XML.

- **Signposting** (см. п. 1 «Унифицированные идентификаторы ресурса»)

- **WebMention** (https://www.w3.org/TR/webmention/) — это простой способ уведомлять любой источник о том, что с ним взаимодействовал другой ресурс. Он позволяет, например, устанавливать двунаправленные ссылки.

- **WebSub** (https://www.w3.org/TR/websub) — протокол публикации/подписки, согласно которому средство публикации посылает информацию об обновлении ресурса в определенный канал центра обмена, впоследствии передающий эти обновления абонентам канала. Репозиторий может публиковать уведомления о произошедшем обновлении одному или нескольким каналам, например по одному каналу для каждого типа действий (цитирование, рецензирование, аннотирование и т. д.). Это может быть достигнуто способом, аналогичным ResourceSync Change Notifications. Агрегирующие приложения могут подписываться на эти каналы репозитория.

Существуют и другие протоколы обмена сообщениями, например:

- **AMQP** (Advanced Message Queueing Protocol) — широко поддерживаемый открытый протокол для передачи сообщений между компонентами системы с низкой задержкой и на высокой скорости. При этом семантика обмена сообщениями настраивается под нужды конкретного проекта;

- **Apache Kafka** (http://kafka.apache.org/) — распределенный программный брокер сообщений, разработанный в рамках Apache Software Foundation. Написан на языке программирования Scala.Эти протоколы обеспечивают общий механизм связи между издателями любых видов веб-контента и их подписчиков.

4.8. Идентификация пользователей

Репозитории должны поддерживать создание сопутствующего контента, содержащего сведения об аннотировании, комментировании, рецензировании, равно как и о других имевших место формах взаимодействия пользователей с их объектами, хранящимися в репозитории. Если пользователи будут применять унифицированные международные идентификаторы при аутентификации в репозитории, это может способствовать ведению конструктивных дискуссий и укреплению социальных связей. Идентификация пользователей может поддерживать работу персонализированных сервисов, таких как рассылка целевых уведомлений и рекомендаций, что поможет пользователям легче и эффективнее ориентироваться в крупномасштабных распределенных наборах данных. В идеале хорошо бы уметь идентифицировать пользователей и распознавать, что те или иные конкретные действия, предпринимаемые пользователем в пределах любого из репозиториев в сети, принадлежат именно этому пользователю (вне зависимости от того, аутентифицирован этот пользователь или нет). Такая способность позволит репозиториям сотрудничать на международном уровне при анализе пользовательской активности. Для лучшего понимания того, насколько широко используется контент из глобальной сети репозиториев по всему миру, следует также фиксировать и активность анонимных пользователей.

Технологические решения, стандарты, протоколы, поддерживающие такой функционал

- **ORCID** (Open Researcher and Contributor ID, https://orcid.org) — некоммерческий проект, в рамках которого каждому автору научной работы или другого научного вклада присваивается уникальный код. Профиль каждого автора представлен в форме, удобной для восприятия как человеком, так и машиной. Машиночитаемый профиль основан на модели RDF и использует схему FOAF. Организация ORCID предоставляет разные услуги по аутентификации, см. «Аутентификация пользователей».

- **Social Network Identities,** идентификаторы пользователя в социальных сетях, создаются многими социальными сетями. В большинстве случаев эти платформы предоставляют возможность комплексной аутентификации, основанной на идентификаторах пользователя, которые используются в социальных сетях (подробнее см. п. 4.9 «Аутентификация пользователей»).

- **WebID** (https://www.w3.org/2005/Incubator/webid/spec/identity/) представляет собой унифицированный идентификатор ресурса URI, который назначен отдельному пользователю (человеку, организации, группе и т. д.) и формируется в домене, как правило, принадлежащем данному пользователю. По адресу WebID находится машиночитаемый профиль, описывающий пользователя. Данный профиль основан на модели RDF, полностью контролируется пользователем и использует схему FOAF. WebID обычно используется в сочетании с аутентификацией при помощи WebID/TLS (подробнее см. п. 9 «Аутентификация пользователей») и способом авторизации через списки управления доступом из веб (Web Access Control Lists, Web ACL).

4.9. Аутентификация пользователей

Применение пользователями унифицированных идентификаторов URI при взаимодействии с объектами репозитория (аннотировании, комментировании или рецензировании) может способствовать ведению конструктивных научных дискуссий и укреплению социальных связей. Однако только идентификации пользователя в момент взаимодействия с контентом недостаточно. Заявленный пользователем идентификатор должен быть в первую очередь подтвержден провайдером, присвоившим ему этот идентификатор. Таким образом, репозитории обязаны поддерживать технологии, позволяющие осуществлять проверку подлинности пользователей, идентификаторы которых созданы

как кодом ORCID, так и социальными сетями (например, «Твиттером», «Гуглом», «Фейсбуком» и т. п.).

Для любого пользователя важно, чтобы репозиторий распознавал его самого, равно как и других пользователей, поскольку это позволит ему оставаться на связи с теми пользователями, которых он знает, оставлять комментарии и получать информацию о контенте, представляющем для него интерес. Для менеджера репозитория необходимо пресекать случаи ненадлежащего обращения пользователей с контентом. Требуя от пользователей идентифицировать самих себя и верифицировать свою идентичность, провайдер снижает риск такого ненадлежащего обращения.

Технологические решения, стандарты, протоколы, поддерживающие такой функционал

- **HTTP Signatures** (https://github.com/asonix/http-signatures) представляют собой способ аутентификации, концептуально похожий на WEBID/TLS. Однако это более общий подход, поскольку он не связан исключительно с глобальным идентификатором WebID. Кроме того, в дополнение к аутентификации этот подход позволяет проверить, была ли связь между клиентом и сервером фальсифицирована или нет. В настоящее время этот подход находится в процессе стандартизации в IETF, за ходом которой имеет смысл внимательно следить.

- **OpenID Connect** (http://openid.net/connect/) — это третье поколение OpenID-технологии, которое представляет собой аутентификационную надстройку над протоколом авторизации OAuth 2.0. OpenID Connect позволяет интернет-ресурсам проверить личность пользователя на основе аутентификации, выполненной авторизационным сервером. Для работы используется RESTful API, описанное в спецификации. Также в OpenID Connect определены дополнительные механизмы для надежного шифрования и цифровой подписи. Стандарт позволяет использовать дополнительные возможности,

такие как управление сессиями и обнаружение OpenID-провайдеров. Основные поставщики идентификаторов социальной сети уже поддерживают OpenID Connect. В настоящее время приложение ORCID находится в стадии бета-тестирования (http://openid.net/connect/).

- **WebID/TLS** (https://www.w3.org/2005/Incubator/webid/spec/tls/) — это протокол, обеспечивающий безопасную аутентификацию пользователей на основе Transport Security Layer protocol (TSL), X.509 Certificates и WebID с соответствующим профилем. Согласно данному протоколу аутентификация осуществляется простым способом через выбор сертификата из предлагаемых браузером. Сертификат используется для составления ответа на полученный от вызываемой стороны (сервера) запрос с помощью закрытого ключа пользователя, а также для передачи идентификатора пользователя. Этот идентификатор направляет сервер к профилю пользователя, в котором содержится закрытый ключ, позволяя серверу тем самым проверить, был ли дан корректный ответ на запрос вызываемого сервера. Хотя этот подход к аутентификации является одновременно простым, эффективным и полностью распределенным, его введение в оборот повсеместно до сих пор не состоялось в силу ряда причин, в том числе проблем с генерацией сертификатов и проблем с пользовательским интерфейсом.

4.10. Стандартные метрики использования ресурсов

Репозитории должны уметь накапливать и вести обмен данными о действиях пользователей, чтобы обеспечить оценку сервисов, которые предоставляют репозитории. Сбор показателей важен для оптимизации и управления деятельностью репозитория, а также чтобы продемонстрировать значимость репозитория для авторов и других пользователей.

Для каждого автора важно знать, как часто используются его статьи, база данных или другие ресурсы, чтобы у него была возможность сравнить результаты со статьями других авторов и в итоге получить объективный и стандартизированный способ оценки собственного вклада в науку. Для любой финансирующей организации важно использовать метрики использования ресурсов в качестве единого параметра, который поможет оценить научный вес исследования, ею финансируемого. Для отделов, управляющих научной деятельностью, важно использовать более широкий спектр параметров для оценки научного веса исследования, включая метрики репозитория, чтобы затем включить их в отчеты, оценивающие научный вклад отдельного исследования.

Методологии для измерения использования ресурсов должны быть стандартизированы. Показатели также должны быть надежными и внушать доверие, чтобы их можно было сравнивать аналогичными, полученными в различных репозиториях. В то же время, если репозитории содержат копии одной и той же статьи, они должны уметь вести обмен данными и суммировать свои метрики использования ресурса, что, в свою очередь, позволит автору (и другим пользователям) видеть полную статистическую картину. Важно обратить внимание на то, что создание надежной и независимой системы стандартных метрик использования ресурсов в Интернете сможет создать противовес монополизации рынка коммерческими издательствами. Тем не менее, учитывая присущие количественным параметрам ограничения в плане оценки качества и актуальности исследования, именно оценка качества, реализуемая посредством аннотирования, рецензирования и комментирования, имеет наивысшее значение.

Представление метрик использования ресурса может быть выполнено в любом из двух режимов: в режиме периодических запросов (например, используя SUSHI) или в режиме постоянного отслеживания появления информации на сервере поставщика услуг, которые в настоящее время специфичны для каждого поставщика информации (например, Google-analytics, IRUS-UK,

OpenAIRE) с использованием таких инструментов, как Piwik (бесплатная система веб-аналитики с открытым исходным кодом (https://matomo.org/)), RAMP (Repository Analytics and Metrics Portal (http://ramp.montana.edu/)) и т. п. Однако одна из основных проблем, связанных с представлением статистических метрик использования ресурсов, заключается в обеспечении прозрачности и сопоставимости статистических показателей разных репозиториев. Эта проблема не может быть решена только технологически, необходимо принятие общих стандартов.

Технологические решения, стандарты, протоколы, поддерживающие такой функционал

- Технологии, необходимые для передачи ресурсов на платформы архивирования — см. п. 5 «Передача контента».

- **COUNTER** (https://www.projectcounter.org/) — это стандарт, который определяет перечень и форму предоставления статистических данных по использованию электронных ресурсов. Стандарт, известный также как «Свод правил» (Code of Practice), обеспечивает поставщиков и издателей возможностью предоставить библиотекам и провайдерам данных сопоставимые сведения об использовании ресурсов.

- **SUSHI** (Standardized Usage Statistics Harvesting Initiative, http://www.niso.org/standards-committees/sushi) — эта инициатива нацелена на определение протокола передачи статистических данных для автоматического сбора отчетов по использованию онлайновых ресурсов. SUSHI стал стандартом NISO в 2007 году и был обновлен в 2014-м (получил номер ANSI/NISO Z39.93–2014) путем добавления протокола, который описывает автоматическое отправление запроса и получение ответа для статистических отчетов в формате COUNTER.

- **Etag**, или **entity tag** (https://ru.wikipedia.org/wiki/HTTP_ETag) — часть HTTP, протокола World Wide Web. Это один из нескольких механизмов, с помощью которых HTTP обеспечи-

вает веб-проверку кэша и который позволяет клиенту делать условный запрос. Это позволяет кэшу быть более эффективным и экономит пропускную способность, так как веб-серверу не нужно отправлять полный ответ, если содержимое не изменилось. ETag также может быть использован для оптимального управления многопоточностью как способ, чтобы помочь предотвратить одновременное обновление и перезапись ресурса.

- **ETag** — это закрытый идентификатор, присвоенный веб-сервером на определенную версию ресурса, найденного на URL. Если содержание ресурса для этого адреса меняется на новое, назначается и новый ETag. Использование в таком ключе ETags аналогично использованию отпечатков пальцев, можно быстро сравнить и определить, являются ли две версии ресурса одинаковыми или нет. Сравнение ETag имеет смысл только с Etag с одного и того же URL, идентификаторы, полученные из разных URL-адресов, могут быть, а могут не быть равны, вне зависимости от ресурсов, так что их сравнение не имеет какого-либо смысла.

- Большой опыт сбора статистики использования ресурсов накоплен в **IRUS-UK** (Institutional Repository Usage Statistics UK, http://irus.mimas.ac.uk/), который является национальным сервисом сбора статистики использования онлайновых ресурсов в Великобритании. На сайте этого сервиса регулярно публикуются полезные материалы о современных технологиях сбора статистики.

4.11. Архивирование

Открытый доступ гарантирует доступ к данным не только в текущий момент, но и в будущем. Необходимо предусмотреть такие способы архивирования контента, которые обеспечат стабильное функционирование всей системы взаимосвязанных ре-

позиториев. Для пользователя репозитория, будь то независимый ученый или научная организация, важно, чтобы их научные результаты хранились продолжительное время. Кроме того, важно быть уверенным, что статью или другой материал можно будет восстановить в случае неполадок в работе системы репозитория. Совершенно не обязательно, чтобы каждый репозиторий создавал свои собственные системы архивирования — скорее нужно разработать общие стандарты и протоколы, для обеспечения интероперабельности. Кроме того, при архивировании необходимо сохранить сложную взаимосвязь ресурсов различных репозиториев на различных уровнях (данных, метаданных и их взаимосвязях). Плюс необходимо сохранять постоянно изменяющийся контент репозиториев в режиме реального времени. Этого можно достичь только посредством внедрения новых технологий в процессы создания и передачи информации. Также нужно учесть, что для обеспечения долговременной сохранности более надежными представляются открытые форматы, Несмотря на повсеместное распространения формата PDF, для хранение полных текстов статей репозитории должны работать со всеми распространенными открытыми форматами (например, LaTex и TEI), которые при загрузке в репозиторий должны быть проверены соответствующими валидаторами, например разработанными в рамках проекта Data Documentation Initiative (DDI, http://www.ddialliance.org/).

Технологические решения, стандарты, протоколы, поддерживающие такой функционал

- Электронное архивирование представляет собой хранение электронного контента в неизменном виде на протяжении времени для обеспечения к нему постоянного доступа. Архивирование — это сложнейший процесс, базирующийся на целом комплексе принципов, стандартов, практик и технологий. Огромную работу по выявлению лучших практик и специализированных технологий ведут сообщества профессионалов в области электронного архивирования. Поэтому авторы настоящих рекомендаций не ставили перед собой цели сделать детальный обзор всех существующих техноло-

гий, а скорее концентрировалась на способности репозиториев переносить их полнотекстовый контент на платформы архивирования. Основные технологические решения, обеспечивающие такую задачу, представлены в п. 5 «Передача ресурсов».

Благодарности

Составление настоящих методических рекомендаций проводилось в рамках проекта «Национальный агрегатор открытых репозиториев российских университетов», поддержанного грантом Президента Российской Федерации на развитие гражданского общества № 17-2-003857.