Единая идентификация библиографических метаданных: проблемы и решения

Мазов Н.А.
Институт нефтегазовой геологии и геофизики им. академика А.А. Трофимука СО РАН, Новосибирск, Россия
Гуреев В.Н.
Государственный научный центр вирусологии и биотехнологии «Вектор», Новосибирская область, р.п. Кольцово
Жижимов О.Л.
Институт вычислительных технологий СО РАН, Новосибирск, Россия

В реферативных библиографических базах данных для упрощения обмена информацией используются уникальные идентификаторы для каждого информационного источника, позволяющие легко их отыскивать. В настоящее время нет единого стандартизованного принятого способа идентификации журнальных статей, авторов и др., несмотря на то, что в последние годы введено в действие немалое число различных идентификаторов. Проблема идентификации становится особенно актуальной при использовании наукометрических баз данных, в связи с появлением различных веб-сервисов, позволяющих проводить комплексную обработку данных с дальнейшей интеграцией полученных данных. В настоящей работе описывается решение для идентификации библиографических метаданных научных публикаций на основе использования идентификатора SICI.

Введение

В последние годы в области, касающейся научно-технической иинформации (НТИ), произошел глобальный переворот. Издатели и агрегаторы НТИ предоставляют на своих порталах преимущественно библиотечные функции, а с другой стороны, библиотеки также принимают на себя издательские функции, когда берутся за создание репозитариев электронных документов или открытых архивов. Издательское и библиотечное дело вошли в ту фазу, когда происходит объединение того, что прежде было в компетенции и квалификации каждого из них, тогда как цели их расходятся.

Традиционно рассматриваемая как техническая задача, идентификация стала важным элементом в организации доступа к электронным публикациям и прочим интеллектуальным артефактам. Идентификация сейчас является основополагающим компонентом так называемой «политэкономии информации» [1], согласно которой контроль над техническими средствами доступа к ресурсам настолько же важен (а иногда и более важен!), как и контроль над самими ресурсами. В этой статье мы рассмотрим современные тенденции развития систем идентификации и достижения в этой области.

Обзор систем идентификации

«Идентификаторы – это имена, или нити, привязанные к определенным условным знакам, которые при соответствующем использовании гарантируют уникальность» [2]. Строго говоря, системы идентификации попросту связаны с обозначением уникального опубликованного или медийного продукта. Как бы то ни было, необходимость в доступе к электронным публикациям, рассеянным по коммуникационным сетям через метаданные, сделала идентификацию лишь одним компонентом (хотя и очень важным) более широкой задачи хранения и передачи коммуникационных пакетов для библиографического учета, обнаружения ресурсов, электронной доставки документов. Ниже приведен небольшой беглый обзор существующих в мире идентификаторов, используемых для опубликованных объектов и которые обычно используются библиотеками.

ISBN. Международный стандартный книжный номер, созданный в 1970 г. (ISO 2108). ISBN является «разумным» идентификатором для опубликованных монографий, поскольку каждый из его элементов несет определенное значение [3]. Существует три уровня управления ISBN: международный, национальный и индивидуальный. На международном уровне управление осуществляется в Международном агентстве ISBN, расположенном в Государственной библиотеке (Берлин. На национальном уровне управление ISBN осуществляется национальными агентствами, состоящими из независимых издателей, ассоциаций издателей и (или) книгопродавцев, а также некоторых специализированных отделов национальных библиотек. На индивидуальном уровне управление ISBN осуществляется самими издателями, которые присваивают ISBN каждой из опубликованных книг. Современная организация сети ISBN имеет два основных недостатка:

- нет централизованной базы данных;
- нет применения на практике, т.е. возможности узнать название ресурса путем клика по ISBN.

Кроме того, ISBN достигает пределов своих возможностей, и ряд агентств ISBN скоро израсходует список доступных номеров. В настоящее время ведутся работы по пересмотру ISBN.

ISSN. Международный стандартный серийный номер [4] (ISO 3297) создан в 1975 г. ISSN – это «немой» идентификатор. Т.е., строка, не содержащая информации относительно содержания или происхождения периодического издания. ISSN и имеет двухуровневое управление. На международном уровне база данных, содержащая более миллиона записей ISSN (реестр ISSN), поддерживается Международным центром по ISSN. На национальном уровне управление возложено на национальные агентства по ISSN, которыми могут быть как

специализированные отделы национальных библиотек, так и организации по работе с документами внутри национальных исследовательских центров.

ISRN. Международный стандартный номер отчета (ISO 10444), который устанавливает единообразный формат для создания уникальных, но совместимых номеров, используемых при идентификации, организации и размещении технических отчетов.

ISMN. Международный стандартный музыкальный номер [5] (ISO 10957) принят в 1993 г. Аналогично коду ISBN, ISMN применяется в отношении нотных изданий. ISMN разработан для рационализации процесса обработки записей о нотных изданиях и их библиографических.

ISWC. Международный стандартный код музыкальный произведений (ISO 15707) [6]. Принят ISO в 2001 г. ISWC – уникальный постоянный признанный международно ссылочный номер для идентификации музыкальных произведений.

ISAN. Международный стандартный аудиовизуальный номер, который определяет аудиовизуальные произведения и их выражения (ISO 15706) [7]. Принят в 2002 г..

DOI. Идентификатор цифрового объекта, который обеспечивает способы постоянной идентификации фрагмента интеллектуальной собственности (произведения) в цифровой сети [8]. «Немой» идентификатор DOI состоит из четырех компонентов:

- *нумерация*: присвоение буквенно-цифровой строки (числа или имени) записи интеллектуальной собственности, которую определяет DOI;
- *описание*: связь метаданных с записью, которая определяется присвоенным DOI;
- *разрешение*: обеспечение разрешающих сервисов с использованием интернет-технологий, которые делают идентификатор «действенным» в цифровой сети (эти сервисы основаны на Handle Systemm [9] и включают открытое множество протоколов, область имен и базовую реализацию протоколов);
- *принципы*: правила управления операциями системы в социальной инфраструктуре.

Общую политику DOI диктует Международная некоммерческая организация DOI, основанная в 1998 г. Между тем присвоение DOI – это бизнес для ряда регистрирующих агентств, обеспечивающих регистрирующие организации такими сервисами, как присвоение префикса, регистрация DOI и полезная инфраструктура для поддержания данных.

Среди агентств DOI на сегодняшний день наиболее успешным является CrossRef [10], который обеспечивает всестороннее обслуживание применительно к идентификации объекта (в частности, идентификации статей) в периодических изданиях и полнотекстовом поиске этих объектов, т.е. связи между ссылками. CrossRef выбрал такую структурную бизнес-модель, в которой с издателя взимается плата, изменяемая в зависимости от объема выпуска и количества отыскиваемых объектов. Кроме того, для каждого DOI требуется ежегодный

депозитный взнос. Таким образом, финансовое бремя ложится на издателей, а не на потребителей.

SICI. Идентификатор объектов периодических изданий и статей. Стандарт SICI представляет собой расширяемый механизм для уникальной идентификации как выпуска периодического издания, так и объекта внутри него (например, статьи) [11]. SICI был принят Организацией по национальными информационным стандартам США (NISO) в 1996 г., но не был принят ISO, и сейчас используется совместно с ISSN в библиотеках Северной Америки. Такой ведущий агрегатор электронных ресурсов как Ingenta, также принял SICI для обеспечения связи с третьими сторонами. Его использование бесплатно. В настоящее время SICI не имеет сети агентств на международном или национальном уровнях.

Беглый обзор идентификаторов в предыдущем параграфе показал, как системы идентификации могут быть частными в своих целях, но должны быть универсальными в своем охвате. Тем не менее, системы идентификации обязаны обслуживать как можно большее число сообществ, чтобы получить повсеместное применение.

Возможность связи между различными системами идентификации с метаданными и родственными службами изменяет модели институциональных идентификационных систем. Полезность, гибкость и реализуемость идентификатора больше не относятся исключительно к его использованию в качестве указателя на документы. Идентификаторы в настоящее время отсылают к сервисам. Рынок систем идентификации определенно сдвинулся от подхода, ориентированного на продукт, в сторону подхода, ориентированного на сервис.

Старательно разработанные инструменты, такие как функциональные требования к библиографическим записям от IFLA-UBCIM, могут оказаться непригодными для применения к идентификаторам [12]. Согласно этим требованиям, библиографические записи состоят из работы, выражения, воплощения и объекта. Работа – это «отдельное интеллектуальное творение или художественное творчество»; выражение – это «интеллектуальная или художественная реализация работы»; воплощение – это «физическое овеществление выражения работы», а объект – это «единичный экземпляр воплощения».

Каким образом классифицировать DOI в терминах вышеперечисленных требований? DOI не идентифицирует ни работу, ни выражение, ни воплощение. Иначе, он может идентифицировать их все, поскольку между категориями, перечисленными в требованиях, не проводится различий. «Мета-идентификатор» DOI буквально интерпретирует совпадения в коммуникационных сетях и применяет их к любому цифровому объекту, выставляя значимым лишь то, что подпадает под защиту авторских прав. DOI присваивают всем объектам, к которым его можно «прилепить». Например, значение геологического журнала заключается в его рисунках, без которых текст и соответствующие подписи принесут мало

пользы. В статье по статистике диаграммы и таблицы – наиболее значимые части, и текст часто лишь дополняет их.

В ином аспекте проблема идентификации поставлена Инициативой открытых архивов (OAI), возможно, наиболее успешной инициативой библиотечного и научного сообщества в пользу бесплатного доступа к информации [13]. Протокол открытых архивов по сбору метаданных (OAI-PMH) работает над доступностью электронных репозитариев и возможностью их взаимодействия. Идентификатор, описанный протоколом OAI, не присваивается (как это обычно бывает у идентификаторов) объекту или ресурсу, находящемуся в архиве, а скорее устанавливает связь между записью метаданных и содержащимся в записи идентификатором (URL, URN, DOI и др.). Формат метаданных Дублинского ядра (ДЯ) является обязательным, ДЯ предоставляет элемент идентификатора, который можно использовать для этой цели. Тем не менее поставщики данных могут выбрать ту систему идентификации, которая будет уникальной в мировом масштабе в пространстве имен OAI [14].

Это свойство обязано действительно «открытой» природе электронных архивов, куда включаются документы, разбросанные по различным коммуникационным каналам (книги, научная периодика и пр.). Идентификатор OAI сам по себе «открытый». Любопытно отметить, что при адресации к постоянному объекту институциональные цифровые репозитарии используют идентификационную систему CNRI для сохранения идентификационных номеров цифровых ресурсов и последующего их разрешения. Так происходит, например, в репозитариях DSpace, E-prints и др., каждый из которых совместим с протоколом OAI [15].

Если рассматривать общие тенденции в системах идентификации сами по себе, то невозможно объяснить сегодняшние институциональные модели и порой напряженные взаимоотношения между этими системами. Также они не помогут понять и того, почему первенство от библиотек переходит к сообществу издателей. Объяснение этой тенденции нужно искать во внутренней динамике области идентификации и в позициях, которых придерживаются игроки обоих сообществ.

Что ожидает традиционные идентификаторы в конкурирующей среде? После бурных 1990-х гг. рынок электронных публикаций чрезвычайно расширился – почти все издатели НТИ к настоящему времени трансформировали свои публикации в электронный формат. Тем не менее, DOI можно считать победителем в сегодняшней сфере идентификации. Его успех подтверждается количеством присвоенных DOI. Из идентификатора цифровых объектов DOI превратился в цифровой идентификатор объектов – он способствует управлению цифровыми записями, решает существующие проблемы и обладает высоким уровнем совместимости.

После успеха DOI работники сферы идентификации были вынуждены изменить свои позиции и пересмотреть свои стратегии и методы работы. Системы идентификации используют технологии как средство расширения их активности и как инструмент для ответа на новые требования.

Заключение

В последнее десятилетие не только увеличилось количество систем идентификации, но и повысилась их значимость. Системы идентификации теперь являются важным элементом любой информационной политики, поскольку они представляют собой инструмент библиографического контроля, поиска и получения информации, исследования ресурса, управления информацией и авторскими правами, а также облегчают все эти процессы. Видимость творческого и научного контента в сети зависит от надежной идентификации. Так что она представляет собой еще и возможность для пользователя выбирать материал бесплатно и без искажения, которое несет с собой коммерческое влияние.

Список литературы

1. Ghislaine Chartron, Jean-Michel Salaün. "La reconstruction de l´économie politique des publications scientifiques." *Bulletin des Bibliothèques de France*, 45, 2000, №. 2, p. 32-42.
2. Amy Brand, Frank Daly, Barbara Meyers. *Metadata demystified*. http://www.niso.org/standards/resources/Metadata_Demystified.pdf.
3. ISBN, http://www.isbn.org/standards/home/index.asp.
4. ISSN, http://www.issn.org:8080/pub.
5. ISMN, http://www.ismn-international.org/.
6. ISWC, http://www.nlc-bnc.ca/iso/tc46sc9/iswc.htm.
7. ISAN, http://www.nlc-bnc.ca/iso/tc46sc9/isan.htm.
8. DOI, http://www.doi.org/
9. Handle System, http://www.handle.net/>
10. CrossRef, http://www.crossref.org/.
11. SICI, http://sunsite.berkeley.edu/SICI/.
12. IFLA Study Group on the Functional Requirements for Bibliographic Records. http://www.ifla.org/VII/s13/frbr/frbr.pdf.
13. OAI, http://www.openarchives.org/.
14. OAI PMH, http://www.openarchives.org/OAI/openarchivesprotocol.html>
15. OSI, http://www.soros.org/openaccess/pdf/OSI_Guide_to_Institutional_Repository_Software_v1.pdf.