Идентификация в информационных библиографических системах: проблемы и решения

Identification in information bibliographic systems: Problems and solutions

В. Н. Гуреев, Н. А. Мазов
Институт нефтегазовой геологии и геофизики
им. академика А. А. Трофимука СО РАН,
Новосибирск, Россия

Vadim Gureyev and Nikolay Mazov
A. A. Trofimuk Institute of Petroleum Geology and Geophysics,
Siberian Branch of Russian Academy of Sciences,
Novosibirsk, Russia

В настоящее время нет единого принятого и стандартизованного способа идентификации журнальных статей, авторов и других элементов библиографических метаданных, несмотря на то, что в последние годы введено в действие немалое число различных идентификаторов. Проблема идентификации становится особенно актуальной при использовании одной и той же информации в различных наукометрических и библиографических базах данных, когда необходимо проводить комплексную обработку данных с дальнейшей интеграцией полученных данных. Необходимость единых идентификаторов за пределами одной системы является глобальным требованием. Различные инициативы построения идентификаторов и систем регулярно появляются в научной печати. Однако они пока не достигли необходимой степени интероперабельности. В настоящей работе представлены общие проблемы, связанные с идентификаторами в широкой научной области, проанализирован ряд имеющиеся систем и технологий, приведены современные инициативы для устранения пробела в области идентификации.

Today, there is no uniform accepted and standardized way of identification of journal articles, authors and other bibliographic metadata elements despite many identifiers developed in recent years. The problem is especially acute when using the same information in various scientometric and bibliographic databases for complex data processing with further integration of data obtained. The globally compatible uniform identifiers are needed, and many proposals are made in science periodicals. However, the necessary degree of interoperability has not been achieved yet. The authors discuss general problems of identifiers in science and research, available systems and technologies are analyzed, modern initiatives are reviewed.

Введение

Для упрощения обмена информацией в реферативных библиографических базах данных (БД) и информационных системах принято использовать уникальные идентификаторы для различных информационных источников внутри системы, позволяющие легко их отыскивать. Внутрисистемные идентификаторы хорошо работают в рамках системы для определения и накопления информации об объектах. Поскольку научная информация становится более открытой и доступной, релевантные информационные объекты повторяются во множестве систем, поддерживающих совместно используемые описания. Так, например, в каждой из этих систем один и тот же человек, наиболее вероятно, будет иметь свой внутрисистемный идентификатор, и в каждой системе описания имени могут иметь орфографические варианты записей в другом написании. Необходимы механизмы управления идентичностью, чтобы обеспечить управление различными идентификаторами. Средства для описания и связи внутрисистемных и внешних сущностей являются, таким образом, необходимыми как в глобальном, так и во внутрисистемном масштабе, охватывающих множество информационных систем.

Несмотря на очевидную актуальность проблемы, в настоящее время в мире нет единого стандартизованного принятого способа идентификации журнальных статей, авторов, их мест работы и др., несмотря на то, что в последние годы введено в действие немалое число различных глобальных идентификаторов [1–3]. Тем не менее, авторам практически неизвестны работы, за редким

исключением [4–5], где бы обсуждались проблемы идентификации авторов, профилей авторов, организаций, публикаций.

Проблемы идентификации в информационных системах

Далее будут рассмотрены проблемы уникальной идентификации и пути решения на примере идентификации персон, публикаций (метаописаний) и организаций.

1. Идентификация персоны

В рамках одной научной организации или библиографической БД (например, БД Трудов сотрудников) [6–8] научные сотрудники обычно идентифицируются при помощи уникального идентификатора, или номера. Однако в основном такой идентификатор является уникальным только в определенной информационной системе или службе, и часто каждая организационная единица, такая как отдел кадров, информационно-библиотечный центр или бухгалтерия организации, создает и поддерживает свои собственные уникальные идентификаторы для каждого работника с его собственными идентификационными признаками. Для единого представления, например, продукции научного работника, он должен единообразно и однозначно идентифицироваться во всех системах, имеющих его публикации. Для этого необходимо решение следующих задач: во-первых, должен быть создан идентификатор, который будет уникальным не только в контексте одной системы, но и во всей организации или даже за пределами организации и, во-вторых, различные идентификаторы одного и того же сотрудника должны быть связаны друг с другом.

Общая потребность в уникальной идентификации персон состоит в их роли как авторов. Отыскивание всех публикаций определенного ученого на основе его заданных имени и фамилии не приводит к удовлетворительному и полному перечню всех его публикаций, поскольку они не являются уникальными для всех авторов в мире. Кроме того, фамилия или имя конкретной персоны может измениться.

Система идентификации персоны является одной из наиболее важных проблем. Уникальный идентификатор должен пройти через сравнение записи на наличие ошибок, основанное на библиографических записях, и публикации должны быть снова интегрированы с идентификаторами авторов. Это позволит избегать создания новых идентификаторов при объединении записей с новыми библиографическими массивами.

Известной технологией, которая в настоящее время поддерживает использование уникальных идентификаторов, является система единственной записи (SSO – Single Sing-On). Она позволяет устанавливать персональную подлинность (аутентифицировать) однажды, и впоследствии использовать внешне- или внутриорганизационные услуги без повторной регистрации. Системой открытого доступа для поддержки аутентификации, основанной на решениях SSO, является, например, система Shibboleth [9].

2. Идентификация публикации

Уникальная идентификация публикации в научной области является важной. Когда ученый публикует статью, он хочет делать эту публикацию доступной для цитирования, поскольку число ссылок считается общей мерой признания его работы. Фактически, вся научная деятельность строится на опубликованных результатах, и поэтому предшествующая работа и, следовательно, публикация должны сделаться доступными посредством однозначной ссылки. В настоящее время считается обычным и нормальным сделать публикации доступными через всемирную сеть Интернет с дополнительным требованием долговременного хранения. Следовательно, от системы требуется хранить связи с публикациями постоянно, обеспечивая как уникальный указатель ресурсов URL, так и идентификатор, который является единым для издателя и независимой системы (т.е. постоянным идентификатором). Таким образом, это также необходимо потому, что URL, который используется для связи с определенной публикацией, должен функционировать, даже если местонахождение сервера, где хранится публикация, было изменено. Уникальные идентификаторы публикаций сопровождаются не только ссылками и местом хранения, но, кроме того, запросами или обменом, а также системными перемещениями.

Обычно различные публикационные порталы, такие как E-Library, российская научная электронная библиотека (http://elibrary.ru), PubMed, бесплатная база данных Национальной медицинской библиотеки США (http://www.ncbi.nlm.gov/pubmed/), база данных Web of Science корпорации Thomson Reuters или база данных Scopus издательства Elsevier и др., хранят информацию о публикациях и обеспечивают интерфейсы для выгрузки метаданных. На основе этих метаданных ученые могут импортировать информацию о своих публикациях, а также ссылках. Чтобы однозначно идентифицировать одну и ту же публикацию в различных электронных порталах, создается единый идентификатор, поскольку иначе идентификация описаний зависит от сравнений названий публикаций или соответствующих авторских имен, которые подвержены ошибкам, бывают неоднозначными, и поэтому просто не обнаруживаются. К сожалению, во всех перечисленных системах используются внутренние идентификаторы публикаций, а единственным связующим звеном является идентификатор DOI (Digital Object Identifer) [10].

3. Идентификация организаций

Приписывание уникальных и постоянных идентификаторов требуется не только публикациям или ученым, но также полезно для описания научных организаций. Такой идентификатор должен значительно улучшить качество данных, позволит существенно улучшить анализ научных областей. А финансирующие организации, научные проекты определенно получат выгоду от наличия глобального уникального идентификатора организации. Авторам известно, что европейская комиссия ввела так называемые коды идентификации стран-членов Европейского союза PIC – Participant Identification Codes (http://cordis/europa/eu/fp7/pp-pic_en.html) для седьмой рамочной программы. Коды PIC позволяют однозначно идентифицировать организации и их подразделения, тогда как у представляющих документы единиц нет необходимости повторять связанную с организацией информацию при каждом представлении. Код PIC не является публичным.

Представленные случаи использования идентификации являются обычными в научной среде, а выбранные сущности – персона, публикация и организация, т.е. деятели и продукция, несомненно являются самыми важными в смысле научных параметров, для которых однозначная идентификация – это вклад в качество. Здесь важно также отметить, что глобальные идентификационные системы требуют не просто технологических решений, но и необходимого управления.

Глобальные идентификаторы – существующие решения

В научном мире ликвидация глобального пробела в идентификаторах признается решающим моментом для улучшения качества информационных систем, и в то же время для того, чтобы дать возможность крупномасштабного распространения информации или многократного ее использования. В этом отношении были предприняты многочисленные международные инициативы [11].

Следует отметить, что если бы научные сущности, такие как персоны, публикации и организации, имели бы уникальные идентификационные номера, то это бы существенно облегчило жизнь всему научному сообществу. Действия, направленные на глобальную идентификацию ученых, развиваются сегодня стремительно. Первым появился идентификатор ученого ResearcherID [12]. Более позднее решение представляет ORCID – (Open Researcher & Contributor ID) [13]. Однако если идентификаторы персон (авторов, сотрудников или ученых) стали наиболее актуальными сейчас, то инициативы идентификации публикаций начались больше десятилетия назад с работы CrossRef – агентства по регистрации официального идентификатора цифрового объекта DOI. Примерно в то же время в библиотечной области был инициирован виртуальный международный авторитетный файл VIAF (Virtual International Authority File). За пределами научной среды существуют всеобщие уникальные идентификаторы UUID (Universally Unique Identifiers), унифицированные указатели ресурсов URL (Uniform Resource Locator) или OpenID. Перечень рассмотренных авторами здесь инициатив и систем не претендует на полноту. Нами сделана попытка обеспечить обзор самых общеизвестных и популярных инициатив и систем для рассмотрения возможного использования их в научной области.

Идентификатор ученого ResearcherID компании Thomson Reuters

Идентификатор ученого – ResearcherID компании Thomson Reuters – официально был представлен на сайте (http://www.rtseacherid.com/) в январе 2008 г. Он разработан в дополнение к БД Web of Science и на данный момент представляет собой отдельную систему, связанную с БД WoS лишь косвенно. Регистрация проходит в реальном времени, а введенные сведения не подвергаются модерации. Его цель – приписывание уникального идентификатора каждому зарегистрировавшемуся автору. Идентификатор был одобрен для использования, поскольку был первой глобальной схемой, готовой и доступной для идентификации ученого. В настоящее время ResearcherID – это бесплатная услуга для всего научного сообщества. Идентификатор ResearcherID открывает доступ к измерениям числа ссылок, поиску и связыванию с соавторами и выгрузке публикаций из базы данных Web of Science или EndNote Web. Профили этого идентификатора содержат многочисленные варианты имен. Учитывая, что вводимые автором сведения не отражаются в наукометрической БД WoS, авторам предоставлена возможность вносить в списки работ также отсутствующие в WoS публикации. При этом при выводе публикаций и при поиске по автору в WoS будет указан его идентификатор ResearcherID, по ссылке на который можно будет перейти к более полному списку работ автора и получить более полные библиометрические показатели по его публикациям. Также возможен поиск по самому идентификатору ResearcherID в Web of Science, что облегчает отбор публикаций автора из авторских множеств. Постоянный сетевой адрес ссылки имеет вид http://www.researcherid.com/rid/B-1327-2012, где B-1327-2012 – возможный идентификатор автора в ResearcherID (последние четыре цифры – год регистрации автора). Данные о публикациях в профиле полностью синхронизируются с данными в системе ORCID (см. ниже). Таким образом, можно вносить изменения лишь в одной из систем, а затем синхронизировать информацию.

Открытый идентификатор ученого и автора ORCID

Наиболее прогрессивной и более общей инициативой является, несомненно, ORCID (http://www.orcid.org) – открытый идентификатор ученого и автора (научного сотрудника). Представлена в конце 2009 г. «для решения проблемы неоднозначности имени автора в научной коммуникации». В августе 2010 г. была создана ведущая некоммерческая организация, членство в которой быстро растет. Система ORCID запущена во второй половине 2012 г. и представляет собой совместную разработку нескольких издательств, университетов и научных сообществ. Основной целью системы является создание стандартов идентификации авторов научных работ. Записи в ORCID хорошо синхронизируются с записями с БД Scopus. С одной стороны, есть возможность перенести все публикации конкретного автора из Scopus в ORCID, указав номер авторского профиля в Scopus (Scopus AuthorID). С другой стороны, если работы автора в Scopus распределены по нескольким профилям, то при включении этих работ в ORCID служба технической поддержки Scopus также отредактирует профиль автора в самой БД Scopus. Таким образом, автор может опосредованно редактировать свой профиль в Scopus, и в этом преимущество ORCID перед ResearcherID. В системе ORCID также предусмотрена полезная опция ручного добавления публикаций, отсутствующих в Scopus. Однако, в отличие от ResearcherID, в ORCID отсутствует опция импорта файлов в формате RIS. Как уже было отмечено выше, система полностью синхронизирует данные с ResearcherID.

Система CrossRef и идентификатор DOI

Система CrossRef (http://www.crossref.org) начала действовать в 2000 г. благодаря некоммерческой независимой организации, созданной ведущими научными издателями мира и именуемой ассоциацией PILA (Publishers International Linking Association).

Инициатива была больше услугой библиографической связи с использованием идентификатора цифрового объекта DOI. Первоначальная миссия библиографической связи, или DOI, была позже расширена «для возможности простой идентификации и использования надежного электронного контента путем содействия совместной разработке и применению поддерживающей инфраструктуры» [14].

Идентификатор цифрового объекта DOI – это уникальная буквенно-цифровая строка, которая обеспечивает способы постоянной идентификации объекта интеллектуальной собственности

в цифровой сети. Система CrossRef связывает с каждым DOI множество основных метаданных, а URL указывает на полный текст в сети. В сфере научных публикаций DOI может присваиваться всем видам журнальных публикаций, а также главам монографий. Основные функции DOI применительно к научным публикациям – постоянство ссылки на цифровой объект вне зависимости от его местоположения в сети, когда производится перенаправление на действующий URL (см. ниже), возможность цитировать статьи, уже прошедшие рецензирование и выставленные онлайн, но ещё не сформированные в номер, возможность поиска публикации по DOI в библиографических БД.

Этот идентификатор присутствует у большинства публикаций международных периодических изданий и практически отсутствует в российском журнальном сегменте, что значительно снижает международную видимость российских публикаций. Это связано с коммерческим характером идентификатора DOI для издателей, которые оплачивают присвоение идентификаторов и дальнейшее ежегодное обслуживание регистрирующим организациям (наиболее крупная из которых – CrossRef). Данные организации, в свою очередь, гарантируют постоянный и точный доступ по DOI к цифровому объекту (публикации), который будет сохраняться даже при изменении местоположения объекта в сети Интернет и смене сетевого адреса. Например, идентификатор DOI 10.6017/ital.v32i4.3421 через систему CrossRef (http://dx.doi.org/) разрешает доступ к URL, позволяющему доступ к полному тексту публикации: http://dx.doi.org/10.6017/ital.v32i4.3421.

Идентификатор DOI состоит из двух частей, где первая часть (приставка) обозначает издательство, а вторая (суффикс) идентифицирует публикацию или ее часть. Так, в вышеприведённой ссылке цифры 10.6017 обозначают издательство American Library Association и будут одними и теми же для всех журналов общества, а символы ital.v32i4.3421 обозначают конкретную статью журнала Information Technology and Libraries.

В современных статьях идентификатор DOI обычно указывается на первой странице публикации. Ряд издательств инициативно присваивает DOI всему архиву своих публикаций, который может насчитывать несколько столетий. В этом случае идентификаторы DOI старых публикаций доступны через библиографические БД или на сайте журнала.

Виртуальный международный авторитетный файл VIAF

Виртуальный международный авторитетный файл VIAF (Virtual International Authority File) (http://www.oclc.org/viaf/) управляется международной службой библиотек OCLC (Online Computer Library Center, Inc.), объединенных для обеспечения доступа к крупным мировым авторитетным файлам. Файл VIAF был инициирован Библиотекой конгресса США, Немецкой национальной библиотекой, Национальной библиотекой Франции и OCLC. Сам файл выглядит как строительный блок семантической сети. Наиболее крупные библиотеки поддерживают перечни имен людей, организаций, конференций и географических мест и других сущностей – они называются авторитетными (нормативными) файлами, которые VIAF предназначает для включения и многократного использования.

Унифицированные идентификаторы ресурсов URI

Обозримость семантической сети по отношению к документам за пределами сети осуществляется через связь данных посредством применения основного ряда принципов, а именно использования унифицированных идентификаторов ресурсов URI (Uniform Resource Identifiers) как названий вещей. Связанные данные отсылают к URI как глобальным идентификаторам [15]. Технически URI это структурированная строка знаков, применяемая для опознания имени или ресурса. Синтаксис URI включает схему URI (например, http, ftp, file) с последующим знаком двоеточия и затем специальной частью схемы. Идентификатор URI может быть относительным или абсолютным, например: resource.txt или http://example.org/resource.txt . Фактически, каждый раз создается новый идентификатор URI в связи с каждым новым фрагментом информации, который мы храним, давая ему имя.

Универсально уникальныq идентификатор UUID

Организация OSF (Open Software Foundation) рекомендовала использование так называемых универсально уникальных идентификаторов UUID (Universally Unique Identifiers) в виде программного обеспечения. «Любой человек может создать UUID и использовать его для идентификации чего-либо с полной уверенностью, что один и тот же идентификатор никогда не будет создан кем-либо не умышленно для идентификации чего-либо еще» [Википедия]. Организация OSF некоммерческая организация, основанной в 1988 г. в США для создания открытого стандарта по внедрению операционной системы UNIX. В 1994 г. OSF заявила новую организационную модель, знаменуя окончание своей разработки программного обеспечения. Тем не менее UUID все еще широко используется. По своей сути UUID это 16-байтовый (128-битовый) номер. В своей канонической форме он представлен 31-й шестнадцатиричной цифрой, разделенной на пять групп знаком тире, всего из 36 знаков в форме 8-4-4-4-12, например: 550e8400-e29b-41d4-a716-446655440000 [Википедия].

Открытый идентификатор OpenID

OIDF (OpenID Foundation) [16] это международная некоммерческая организация, основанная в июне 2007 г. и отвечающая за предоставление, продвижение и защиту технологий OpenID. Эта организация представляет собой открытое сообщество разработчиков, поставщиков и пользователей и содействует сообществу обеспечением инфраструктуры и помощи в продвижении и поддерживании расширенного применения OpenID. Идентификатор OpenID это децентрализованный протокол аутентификации (установления подлинности личности, авторизации), который делает его легким для людей при записи и учете доступа к сети. Среди спонсирующих членов можно выделить такие компании как Google, Microsoft, PayPal. Идентификатор OpenID не является специально адресованным научной области, а действует в общем информационном пространстве.

Заключение

Представленные в настоящей работе системы и идентификаторы имеют схожие цели – стремление к глобально уникальной системе идентификаторов и, следовательно, к сетевой научной информационной инфраструктуре или научному информационному пространству. Так, система CrossRef больше концентрируется на научном результате, хотя стремится к более общему охвату, тогда как ORCID и ResearcherID имеют в центре внимания ученого, связанного с его научным результатом. Инициатива VIAF исходит из библиотечной области, обеспечивающая строительные блоки, т.е. авторитетные файлы. Системы ORCID, ReseaecherID и CrossRef с DOI четко направлены на идентификацию своих имен и сосредоточены на самих научных сущностях. При этом они существенно отличаются от URL, UUID и OpenID, которые являются больше технологическими.

Из описанных здесь случаев использования очевидна необходимость глобально уникальной и постоянной идентификации научных сущностей. Мы рассмотрели системы идентификаторов и инициативы в широкой научной области, чтобы лучше понять ключевые проблемы, связанные с идентификацией.

Список литературы:

1. Paskin N. Information identifiers // Learn. Publ. – 1997. – Vol. 10. – № 2. – P. 135.
2. Yao L., Tang J., Li J. A unified approach to researcher profiling // Proceedings of the ieee/wic/acm international conference on web intelligence. – 2007. – P. 359–365.
3. Бездушный А.Н., Кулагин М.В., Серебряков В.А. и др. Предложения по наборам метаданных для научных информационных ресурсов // Вычислительные технологии – 2005. – Т. 10. – С. 29–48.
4. Jцrg B., Hцllrigl T., Sicilia M.-A. Entities and Identities in Research Information Systems. E-Infrastructures for Research and Innovation // Linking Information Systems to Improve Scientific Knowledge Production: Proceedings of the 11th International Conference on Current Research Information Systems (June 6–9, 2012, Prague, Czech Republic). – P. 185–194.
5. Jцrg B., Ruiz-Rube I., Sicilia M-A., Dvorak J., Jeffery K., Hцllrigl T., Rasmussen H.S., Engffer A. Vesterdam T., Garcia Barriocanal E. Connecting closed world research information systems through the linked open data web // International Journal of Software Engineering and Knowledge Engineering (IJSEKE). – 2012. – Vol. 22.

6. Мазов Н.А., Гуреев В.Н. Проблемы идентификации метаданных в наукометрических базах данных Web of Knowledge, Scopus и РИНЦ на примере профилей авторов // Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса: 19-я междунар. конф. «Крым 2012» (2–10 июня 2012 г., г. Судак): Труды конф. – М.: Изд-во ГПНТБ России, 2012. – С. 1–4. – http://www.gpntb.ru/win/inter-events/crimea2012/disk/124.pdf
7. Мазов Н.А., Гуреев В.Н. Идентификация библиографических метаданных научных публикаций в различных базах данных: проблемы и решения // Материалы 8-й Междунар. конф., посвящ. 60-летию ВИНИТИ РАН «Актуальные проблемы информационного обеспечения науки, аналитической и инновационной деятельности», «НТИ – 2012». (28–30 ноября 2012 г., ВИНИТИ РАН, г. Москва). – Москва, 2012. – С. 123–124. – http://www.viniti.ru/download/russian/prog8.pdf
8. Мазов Н.А., Гуреев В.Н. Новые методы формирования публикационного профиля научной организации в сети науки // Научные и технические библиотеки. – 2013. – № 12. – С. 42–48
9. http://shibboleth.internet2.edu [Дата обращения: 18.04.2014]
10. http://dx.doi.org [Дата обращения: 18.04.2014]
11. Enserink M. Are you ready to become a number? // Science. – 2009. – P.1662–1664.
12. http://www.rtseacherid.com [Дата обращения: 18.04.2014]
13. http://www.orcid.org [Дата обращения: 18.04.2014]
14. CrossRef: A short history. – 2009. – http://www.crossref.org/08downloads/CrossRef10Years.pdf
15. Bizer C., Heath T., Berners-Lee T. Linked data – the story so far // International Journal on Semantic Web and Information Systems. – 2009. – Vol. 5, Iss. 3. – P. 1–22.
16. http://openid.net/foundation/ [Дата обращения: 18.04.2014]