

Включение обсерваторских данных в систему цитирования DOI

Р. Ю. Лукьянова¹

Получено 11 декабря 2013 г.; опубликовано 20 декабря 2013 г.

В настоящее время возрастает роль наукометрических показателей для измерения эффективности научных учреждений. В качестве индикаторов используются различные количественные параметры, среди которых важное место занимает публикационная активность и цитируемость. Система DOI (Digital Object Identifier, цифровой идентификатор объекта) уже довольно давно применяется для научных публикаций. Обсуждаются подходы к присвоению DOI данным геомагнитных обсерваторий и глобальным системам данных, собираемым в Мировых центрах данных (МЦД). Включение обсерваторских данных в различные индексные системы и системы цитирования важно для широкого распространения, сохранности данных, а также защиты авторских прав. **КЛЮЧЕВЫЕ СЛОВА:** Геомагнитные обсерватории; наукометрические показатели; Digital Object Identifier (DOI).

Ссылка: Лукьянова, Р. Ю. (2013), Включение обсерваторских данных в систему цитирования DOI, *Вестник ОНЗ РАН*, 5, NZ9001, doi:10.2205/2013NZ000120.

1. Введение

Основой развития любой науки являются наблюдения. В настоящее время происходит интенсивное развитие новых технологических платформ для проведения экспериментов и мониторинга состояния окружающей среды. Быстро совершенствуются методы сбора, первичной обработки и хранения огромных массивов получаемой информации. Эти работы, в отличие от оставшихся в прошлом простых методов слежения и фиксации небольшого количества параметров, становятся все более сложными и дорогостоящими, требуют специальных знаний и оборудования. Это особенно хорошо заметно в области науки о Земле, где мониторинг ведется во всех геосферах, начиная от глубоких слоев земной коры и до дальних областей околоземного космоса. В международном геофизическом сообществе растет понимание того, что получение новых данных не может более рассматриваться как некая вспомогательная деятельность, а становится полноценной научной задачей.

С другой стороны, в настоящее время как во всем мире, так и в России наблюдается возрастающий спрос со стороны финансовых учреждений на наукометрические показатели для измерения производительности труда на-

учных институтов, лабораторий и групп. В качестве индикаторов успешности работы коллективов и отдельных ученых используются различные количественные параметры, среди которых важное место занимает публикационная активность и востребованность результатов. Каждое научное учреждение заинтересовано в увеличении таких показателей, как число научных публикаций, число цитирований, индекс Хирша, импакт-фактор и др. Это актуально и для обсерваторий, занимающихся геофизическими наблюдениями и сбором данных.

В ответ на вызовы времени пришло осознание необходимости выработки новой парадигмы взаимодействия наблюдательной и аналитической составляющих научного процесса и внедрения новой культуры обращения к данным наблюдений.

Начиная с 2007 года, на уровне различных ассоциаций, таких как Международный совет по науке (International Council for Science, ICSU), Комитет по данным для науки и техники (Committee on Data for Science and Technology, CODATA), Американский геофизический союз (American Geophysical Union, AGU), Международная ассоциация геомагнетизма и аэронавтики (International Association of Geomagnetism and Aeronomy, IAGA) и др. и входящих в них междисциплинарных и специализированных рабочих групп, началось обсуждение принципов цитирования данных, основанное на понимании того, что такое цитирование, наряду с привычным цитированием литературных источников, должно быть неотъемлемой частью любой научной публикации.

Данная статья содержит краткий обзор системы DOI (Digital Object Identifier, в переводе цифровой идентифи-

¹ФГБУ Геофизический центр Российской академии наук, Москва, Россия

катор объекта) и обсуждение вопросов ее применения к геомагнитным данным.

2. Декларация о принципах цитирования данных

В течение последних пяти лет достаточно большое количество заинтересованных сообществ изучало состояние вопроса о том, какие обратные связи существуют между пользователями и производителями данных. Изучалась практика цитирования источников данных, которая позволяла бы применять наукометрические технологии и измерять эффективность и востребованность работы наблюдательных сетей.

На основе широкого обсуждения был разработан проект Декларации о принципах цитирования данных (Declaration of Data Citation Principles). Этот документ создан рабочей группой Data Citation Synthesis Group из Force11, объединившей опыт заинтересованных специалистов из разных областей знаний. Предварительная рабочая версия документа размещена на их официальном сайте [<http://www.force11.org/datacitation>].

В Декларации говорится о том, что данные следует считать полноценными результатами исследований, которые, будучи использованы при создании научного продукта, подлежат формальному цитированию. Должна быть выработана единообразная, но достаточно гибкая форма цитирования, позволяющая получить метаданные, определить источник основных данных в сети и получить информацию о различных версиях наборов данных. Кроме того, цитирование должно повышать наукометрические показатели производителей данных.

3. Цитирование геомагнитных данных

Проблемы цитирования данных весьма актуальны для геомагнитных исследований, которые в большой степени основываются на информации, поступающей из обсерваторий. В связи с этим на заседании рабочей группы “Данные и индексы” (Дивизион V “Геомагнетизм” IAGA), проходившем во время XXII Ассамблеи IAGA (Мерида, Мексика, 23–29 августа 2013 г.) был поднят вопрос о цитировании геомагнитных данных.

Возникла дискуссия о необходимости найти решение, которое позволило бы производителям и хранителям данных (обсерваториям, Мировым центрам данных (МЦД) и другим системам сбора) получить должную меру своей востребованности через международно признанные схемы. Цифровой идентификатор объекта (Digital Object Identifier, или DOI), который уже довольно давно применяется в научных публикациях, кажется подходящим решением. Чтобы исследовать эту возможность и выработать рекомендации для сообщества IAGA о едином способе использования DOI для данных магнитных обсерваторий и геомагнитных индексов, была создана проблемно-

ориентированная рабочая группа “DOI для данных магнитных обсерваторий и индексов”. На первом этапе в группу вошли представители Великобритании, Германии, Дании, Индии, России, США, Франции, Чехии, Швеции, Японии. Группа ставит вопрос о создании культуры цитирования данных пользователями. Задача может быть решена, если геомагнитное сообщество найдет общую схему присвоения DOI продукции данных, и если эта политика будет поддержана МЦД и INTERMAGNET.

4. Принципы построения DOI

Система DOI предназначена для идентификации интеллектуальной собственности в виртуальной среде (“The DOI system”, электронный ресурс, [<http://www.doi.org/hb.html>]). Международная федерация DOI назначает Регистрационные агентства (РА), которые выделяют DOI-префикс, регистрируют DOI-имя и обеспечивают инфраструктуру для поддержки метаданных. Основными сферами применения DOI в науке являются:

- Цитирование научной литературы через систему CrossRef, объединяющую более 3000 издательств;
- Цитирование наборов научных данных через систему DataCite, в которую входят научные библиотеки, центры технической информации, научные центры данных.

DOI состоит из уникальной последовательности букв и цифр, которая разделена знаком “/” на префикс и суффикс. Префикс (в котором первой всегда стоит цифра 10) назначается РА. Суффикс назначается издательским агентством. Пример DOI, который назначен данной статье:

10.2205/2013NZ000120

Здесь префикс 10.2205 состоит из кода 10 (назначенного для всех DOI) и кода регистранта 2205 (который РА CrossRef назначило для Геофизического центра РАН). После знака “/” следует суффикс, в котором содержатся метаданные об объекте: год издания, двухбуквенный идентификатор журнала и цифровой идентификатор статьи (номер статьи в шестизначном формате). Текстовая ссылка на данный объект записывается в форме: doi:10.2205/2013NZ000120, а гиперссылка – в форме: <http://dx.doi.org/10.2205/2013NZ000120>.

5. DOI для геомагнитных данных

Уже есть некоторый опыт по получению DOI для геомагнитных данных. Так, японские МЦД и Университет Киото находятся в процессе получения DOI для японских обсерваторий. Этим вопросом занимаются в Геофизическом центре Гельмгольца в Потсдаме (GFZ), в Британской геологической службе (BGS), Центре данных в

Боулдере (NGDC). Между тем, в настоящее время еще нет общего решения, которое удовлетворило бы потребности всех институтов и обсерваторий. Первым шагом в обсуждении должно быть выявление потребностей каждого учреждения и возможностей системы DOI. Возникают следующие вопросы:

1. В мире есть несколько РА, и они ориентированы, в общем, на различную тематику. Получение и сохранение DOI обычно связано с финансовыми расходами, которые составляют до нескольких тысяч долларов в год (как, например, в РА CrossRef и DataCite). Однако существуют и другие РА, а также, при наличии ресурсов, возможна организация дочерних РА, ассоциированных с основными. Нужно выяснить стоимость и критерии выбора регистрирующего органа.
2. Каждый DOI связан с “целевой страницей” (landing page) в Интернете. Значит ли это, что страницы должны располагаться на сервере института (обсерватории), либо они могут быть расположены, например, в МЦД, который предоставляет данные и хранит метаданные, или возможны другие варианты?
3. Нужно ли иметь один DOI для определенного продукта (секундные, минутные, часовые и т.п. значения), или лучше иметь один DOI на одну обсерваторию для временного ряда с любой дискретностью? Что даст большее количество цитирований?
4. Должен ли префикс идентифицировать владельца данных, или префикс должен быть связан с IAGA, таким образом определяя все обсерватории, которые зарегистрированы с официальным кодом IAGA?

Представляется разумным включить IAGA-код обсерватории в суффикс, и также включить обозначение, которое определяло бы принадлежность к классу “магнитные обсерватории” в отличие от другого класса, например, “сейсмические обсерватории”, которые могут иметь такой же код. Поскольку IAGA имеет дело только с данными геомагнитных обсерваторий, мы считаем, что имеет смысл использовать в суффиксе 3-буквенный обсерваторский код, который не меняется, и определенное временное разрешение, например: КАК.1sec (обсерватория Какиока, 1-секундные значения магнитной индукции). Для пользователей лучше и понятней, если структура DOI несет значимую информацию, а также если запись DOI для геомагнитных данных имеет однотипный вид.

Например, суффикс DOI может содержать набор знаков “GM_OBS_HR”, где

- “GM” означает геомагнетизм;
- “OBS” – IAGA-код обсерватории;
- “HR” – аббревиатура типа данных.

Желательно, чтобы был один DOI для глобального продукта (компиляции данных), который позволял бы проследить DOI для всех используемых индивидуальных

данных. Это означало бы, что когда цитируется DOI продукта (компиляции), каждый отдельный поставщик данных получает свою единицу цитирования. Этого можно достичь, если МЦД или INTERMAGNET будут иметь свои DOI. Пользователи, которые работают с данными одной или нескольких обсерваторий, могут привести отдельные DOI, а пользователи глобальных массивов должны иметь возможность привести только один DOI, который дает единицу цитирования каждой обсерватории.

Геомагнитному сообществу надо уяснить, какова цель получения DOI. Здесь следует учесть три аспекта:

1. DOI должен гарантировать воспроизводимость научного результата, т.е. он должен однозначно определять набор данных и его источник.
2. Нам нужно показать, насколько часто используются наши данные, т.е. нам нужны прямые ссылки и цитирование наших данных, аналогичное тому, как это делается для ссылок на названия публикаций в научных журналах. DOI будет инструментом, обеспечивающим автоматический счет цитирований (например, через Thomson Reuters, Google Scholar, SCOPUS) в случае, если данные использованы в публикации.
3. Важно, чтобы введение DOI не потребовало бы много времени (и денег) и не меняло бы нашу обычную практику обращения и распространения данных.

В мировой практике есть различные подходы к получению DOI. Процедура может быть как весьма дорогостоящей, так и бесплатной. Это зависит от РА и издательского агентства. Так, например, Центр Гельмгольца получил права издательского агентства (т.е. может предоставлять DOI) в пределах своего “Узла данных”, где этот DOI бесплатный. Было бы хорошо, если те, кто тесно связан с МЦД или INTERMAGNET, и те, кто уже знает, что для них будет требоваться DOI для данных, выяснили возможности своих институтов в плане предоставления бесплатных DOI. Например, Центр Гельмгольца мог бы предоставлять бесплатно DOI и для других институтов. И конечно было бы удобно, если бы институты, на базе которых находятся МЦД или центры INTERMAGNET, сами предлагали бы функцию предоставления DOI для обсерваторий.

В настоящее время инструменты поиска постоянно развиваются. Ожидается, что каскадные DOI и перекрестные ссылки будут хорошо работать в этой системе. Имеет смысл, чтобы INTERMAGNET и МЦД присваивали бы определенный DOI для собираемого ими глобального массива данных, который цитировал бы все наборы данных, входящих в этот массив. Такая же схема применяется и для геомагнитных индексов, которые рассчитываются по данным нескольких обсерваторий. Каждый раз, когда цитируется глобальный DOI, каждый из связанных DOI тоже получает единицу цитирования, вне зависимости от того, имеют ли все обсерватории свои DOI. Эти связанные DOI могли бы быть частью метаданных и могли бы

обновляться каждый раз, когда входящая в систему, но не имеющая до этого времени своего DOI обсерватория, получала бы его. Таким образом, не нужно требовать от всех обсерваторий, чтобы они получали DOI – нужно постараться построить такую систему, в которой легко могли бы участвовать все обсерватории, желающие с помощью DOI подтверждать частоту использования своих данных.

Следующий практический вопрос – каким количественным параметрам присваивать DOI: всем данным обсерватории или каждому продукту (секундным, минутным, часовым значениям) по отдельности. В целях воспроизводимости научных результатов лучше было бы иметь индивидуальный DOI для каждого стандартного продукта, который может быть скачан по отдельности. Можно сделать кросс-ссылки между продуктами через средние значения, начиная с секундных. Здесь важно, чтобы геомагнитное сообщество пришло к общему для всех обсерваторий соглашению.

Нет необходимости иметь DOI для каждой обсерватории, но нужно иметь DOI для каждого стандартного типа данных, доступных в МЦД и INTERMAGNET, т.е. минутные, среднечасовые, среднегодовые величины. При этом однако возникает вопрос о версиях минутных данных (предварительные, окончательные и т.п.) – должны ли они иметь каждая свой DOI, или один DOI для всех версий с упоминанием в ссылке о том, какая версия использовалась. Скорее всего, второе предпочтительнее, чтобы уменьшить количество DOI.

6. Первые практические шаги

Институт Американской геологической службы в Голдене, на базе которого работает ряд обсерваторий, МЦД и INTERMAGNET GIN, делится своим опытом по организации системы DOI и дает некоторые практические советы:

1. При составлении суффикса DOI используется случайный набор знаков.
2. Каждая геомагнитная обсерватория регистрируется с одним DOI-именем для всего ряда данных определенной частоты (например, часовые значения обс. ESK).
3. Для многих институтов, имеющих обсерватории, обязательно или желательно использовать РА, связанное с данным институтом.
4. РА сильно отличаются друг от друга по способу ведения дел, стоимости, требованиям к метаданным и др. Например, японское РА “Japan Link Center” (JaLC) разрешает регистрантам самим определять набор метаданных и размещать у себя целевую страницу. Целевые страницы DOI будут размещаться в одном из японских МЦД в одном и том же формате, что позволит уменьшить общую стоимость. РА “DataCite” не работает непосредственно

с регистрантами, а делегирует сбор метаданных и целевых страниц аффилированным центрам данных (аналогия такого рода иерархии – работа с метаданными BGS через инфраструктуру NGDC). РА “figshare” (<http://figshare.com>) разрешает запись свободного текста, но не разрешает регистрантам определять свои собственные метаданные или держать собственные целевые страницы.

5. Для многих РА обозначаемый объект (набор данных) должен быть неизменным. Любое обновление должно пройти через регистрационный процесс, и должен быть назначен новый DOI. Это входит в противоречие с п. 2, т.е. с пожеланием иметь единственный DOI, который ассоциируется со всем рядом прошлых и будущих наблюдений на определенной обсерватории. Теоретически можно представить, что изменение/обновление данных не должно сопровождаться изменением DOI, поскольку модифицируется только информация целевой страницы. Именно так делается в РА “figshare”. Однако неизменность DOI может породить проблемы с воспроизводимостью результатов и выводов, если, например, в данных будет обнаружена ошибка, и они будут заменены. Обязательно должно быть введено датирование набора данных.
6. РА “figshare” регистрирует большое количество имен DOI для разных типов цифровых данных, причем бесплатно. После того, как автор данных зарегистрировался на сайте, регистрация DOI-имени для заданного файла или набора файлов происходит как простая загрузка файлов и заполнение формы для основных метаданных (название, автор, ключевые слова, описание). После этого сразу назначается DOI и создается целевая страница. Сами данные размещаются на серверах “figshare”. Это весьма популярное РА часто используется исследователями, желающими размещать и делиться данными и рисунками, и оно уже интегрировано с примерно 1000 журналов открытого доступа. Таким образом, у “figshare” есть следующие преимущества: бесплатное, простое и быстрое получение DOI; поддержка обновлений данных без нового DOI; DOI может быть назначен для любого набора файлов (напр., все файлы окончательных минутных данных для обсерватории); данные легко найти и использовать; системой пользуется большое сообщество. К недостаткам можно отнести: не содержит метаданные, важные именно для геомагнетизма; в настоящее время метаданные не содержат таких полей как “издатель”, “институт-производитель” и т.п. РА “figshare” принадлежит частной компании, и их бизнес-модель, при которой DOI предоставляется бесплатно, не очень понятна. Принимаются любые форматы и типы данных независимо от их качества. Это может быть удобно для индивидуальных, но не для данных обсерваторий.
7. Хотя “figshare” не может соотнести обсерваторские

метаданные с DOI (кроме как в форме свободно-го текста), эта информация уже хранится в МЦД, и МЦД может сам создать целевую страницу для каждого DOI. Страница МЦД будет содержать все необходимые метаданные, включая ссылки на страницы институтов и ссылки на сами данные. С обычной целевой страницы на “figshare” можно создать ссылку на такую расширенную целевую страницу. Кроме того, МЦД может поддерживать связь между DOI и обозначаемым набором данных для всех обсерваторий, а не только для тех, которые используют МЦД как регистранта DOI. Поскольку версия набора данных, на который указывает DOI, может меняться, то в дополнение к DOI требуется еще информация о дате изменения/дополнения временного ряда. Если МЦД будет предоставлять такой сервис, то все обсерватории могут получать число цитирований своих данных бесплатно и без каких-либо действий со своей стороны.

Практические шаги по включению обсерваторских данных и геомагнитных индексов (*AE* и *Dst*) в систему DOI через национальное PA JaLC предпринимаются в Японии. В Европе Центр Гельмгольца предполагает получить DOI обсерваторий Niemeck (NGK) и Wingst (WNG), а также для *K*- и *Kp*-индексов. В ближайшее время в качестве пробного проекта BGS планирует, используя систему NGDC, определить DOI трем английским обсерваториям, участвующим в расчете *K*-индекса. Так как эти данные в свою очередь используются для получения *Kp*-индекса, это будет хороший тест для проверки связи между различными DOI: DOI для *K*-индекса будут входить в метаданные DOI для *Kp*-индекса. Другим пробным проектом будет назначение DOI для 1-минутных рядов данных одной из английских обсерваторий через МЦД Эдинбурга.

7. Новая культура цитирования данных

Для полного обеспечения обсерваторий информацией о том, насколько широко используются их данные, кроме создания самого механизма учета, требуется, чтобы (а) пользователи приучились цитировать данные корректно, и (б) обсерватории были включены в систему индексов цитирования и имели возможность отслеживать, где именно используются их данные. МЦД должны помочь в развитии культуры цитирования, обозначив, что цитирование является необходимым условием пользования данными, и предоставив удобные механизмы копирования ссылок.

Как можно видеть, некоторые МЦД уже приступили к практике применения DOI. Эту практику следует активно внедрять и в других местах. Важно также воспитывать культуру цитирования источников данных в тех случаях, когда эти данные используются в публикуемой работе. Процесс будет развиваться успешно, если такие организации как МЦД, INTERMAGNET и Международ-

ная служба геомагнитных индексов (ISGI) будут постоянно пропагандировать идею о том, что, по аналогии со ссылками на литературные источники, необходимо давать подобные ссылки и на данные обсерваторий, а также предоставлять DOI для своих наборов данных. Издатели ведущих геофизических научных журналов уже призывают редакторов и референтов обращать внимание на необходимость использования DOI во всех ссылках, где он имеется, включая ссылки на источники данных. Рекомендации о корректном цитировании источников данных должны быть включены в инструкции по подаче статей в научные журналы. Следует всячески подчеркивать преимущества ссылок на источники данных с помощью DOI для авторов статей, т.к. при этом читатели получают важную дополнительную информацию. Кроме того, DOI (по крайней мере теоретически) гарантирует постоянство ссылки, независимо от владельца и местоположения самих данных.

Для обсерваторий, управляющих институтами и их объединений полезно было бы давать описания своих продуктов в виде статей в научных журналах, которые автоматически учитываются в наукометрической системе (Thomson Reuter’s Web of Knowledge и т.п.). К тематическим реферируемым журналам относится недавно запущенное издание “Nature – Scientific Data”, а также “Earth System Science Data”, “Dataset Papers in Science”, “The CODATA Data Science Journal” и др. Надо признать, однако, что написание качественной статьи по обсерваторской тематике, в которой содержится необходимое количество оригинального научного материала, является нетривиальной задачей.

При развитой культуре цитирования данных повышается заинтересованность обсерваторий в качестве производимых данных, выкладывании их в Интернет и регулярной поставке в МЦД. Для случаев, когда в работе используется большой комплекс данных (обычно собираемый в МЦД), и список цитирований отдельных источников получился бы слишком длинным, должны быть предусмотрены соответствующие опции для краткого цитирования, которое тем не менее увеличивает счет каждой обсерватории.

8. Заключение

Система DOI является современным инструментом для увеличения эффективности использования и повышения цитируемости научных продуктов. С другой стороны, она способствует тому, чтобы произведенные продукты, в частности, ряды данных обсерваторских наблюдений, были реально доступны для пользователей как в текущий момент, так и в будущем. На современном этапе для поддержания конкурентоспособности российских обсерваторий и центров данных актуальной является задача начать практику присвоения DOI, т.е. начать создание листов метаданных, целевых страниц, выбрать PA, зарегистрировать свои данные, получить DOI. PA может быть любым на усмотрение института/обсерватории.

При этом стоит руководствоваться установкой, что один тип данных (дискретность временного ряда) должен иметь свой собственный DOI. Следует выработать способ, чтобы использование глобальной системы наблюдений цитировалось через один DOI, но приносило бы очки каждой из обсерваторий. Следуя общим тенденциям, необходимо заниматься развитием культуры цитирования данных, а также искать новые нестандартные под-

ходы к описанию данных для публикации статей в тематических журналах, входящих в системы цитирований.

Р. Ю. Лукьянова, председатель рабочей группы “Геомагнитные данные и индексы” Международной ассоциации геомагнетизма и аэрономии (IAGA); ФГБУ Геофизический центр Российской академии наук, ул. Молодежная 3, 119296, Москва, Россия. (r.lukianova@gcras.ru)