КРАТКИЕ СООБЩЕНИЕ

УДК 001.92
doi: 10.31140/j.vestnikib.2018.3(205).5

ZENODO И GBIF: ИНСТРУМЕНТЫ ДЛЯ ПУБЛИКАЦИИ НАБОРОВ ПЕРВИЧНЫХ ДАННЫХ

И.Ф. Чадин
Федеральное государственное бюджетное учреждение науки
Институт биологии Коми научного центра Уральского отделения РАН, Сыктывкар
E-mail: chadin@ib.komisc.ru

Аннотация. В работе представлены инструменты для публикации наборов первичных данных: репозитория Zenodo и Глобальной информационной системы по биологическому разнообразию – GBIF. Даны рекомендации по подготовке рукописей особого типа научных публикаций – статей о данных.

Ключевые слова: Zenodo, GBIF, открытая наука, управление данными

Развитие Интернета оказало и продолжает оказывать сильнейшее влияние на традиционные формы научной коммуникации, культуру научной деятельности в целом. Одним из проявлений этого является тенденция к снижению барьеров на пути распространения результатов научных исследований – «Открытая наука» (Point ..., 2016). Публикация в открытом доступе научных статей, первичных данных, научных компьютерных программ, лекций и руководств – основные составляющие этого явления. Несмотря на то, что предоставление открытого доступа ко всем упомянутым видам ресурсов и знаний, безусловно, требует соблюдения этических и юридических ограничений, в общем случае свободный доступ к результатам работы ученых позволяет повысить качество, ускорить и упростить проведение научных исследований.

Публикация первичных данных в независимых хранилищах (репозиториях) прежде всего приносит выгоду для самих авторов научной работы, такая публикация значительно повышает сохранность и доступность этих данных. Современная инфраструктура центров хранения данных по надежности намного превышает персональные компьютеры большинства исследователей. Намерение предоставить публичный доступ к своим первичным данным вынуждает ученого приложить значительные усилия к их аннотированию, описанию методов их получения. Это в свою очередь значительно упрощает их использование в будущем самим авторам данных, отвечает принципам воспроизводимости научной работы и позволяет корректно ссылаться на эти данные в других работах. Открытая публикация первичных данных подтверждает уверенность авторов работы в достоверности публикуемых материалов и служит формированию их научной репутации. Еще одно преимущество публикации своих результатов в открытых репозиториях – упрощение организации научного сотрудничества, так как опубликованные наборы данных и сопровождающее их описание служит исчерпывающей характеристикой возможностей и квалификации авторов этих данных.

В настоящей публикации будут кратко рассмотрены возможности двух инструментов для публикации результатов экспериментов и наблюдений, которые подходят для большинства биологических исследований: репозитория для первичных научных данных общего назначения – Zenodo (https://zenodo.org) и портала, специализирующегося на публикации данных по биологическому разнообразию международной организации Global Biodiversity Information Facility – GBIF (https://www.gbif.org).

Название репозитория Zenodo связано с именем Зенодота Эфесского – библиотекаря, управлявшего знаменитой Александрийской библиотекой, автора первого в истории применения метаданных для описания литературных источников (https://en.wikipedia.org/wiki/Zenodotus). Zenodo был создан на средства Европейской комиссии при технической поддержке ЦЕРНа. Сервис позволяет публиковать любые материалы по любым направлениям науки: научные статьи, материалы конференций, презентации докладов, первичные данные. Размер одного набора не должен превышать 50 Гб. Данные, отправленные в Zenodo, могут оставаться доступными только самому публикатору или быть открыты для доступа на произвольных условиях, например: доступ с явного разрешения владельца данных, свободный доступ для некоммерческого использования, доступ без ограничений.

После публикации каждому набору данных немедленно присваивается идентификатор цифрового объекта – DOI, который обеспечивает корректное цитирование опубликованных материалов в научной литературе. Предусмотрена функция наложения эмбарго – задания даты, после которой загруженный набор данных станет доступным любым пользователям. После публикации отредактировать сам набор данных уже невозможно, но можно вносить исправления и дополнения в метаданные. С недавнего времени Zenodo стал поддерживать возможность выпуска новых версий своих материалов с присвоением каждой версии отдельного DOI и сохранением специального DOI, относящегося ко всему набору данных. Примером публикации открытого набора данных может служить карта распространения борщевика на территории г. Сыктывкара (Dalke, 2018).

В отличие от Zenodo Глобальная информационная система по биологическому разнообразию GBIF является специализированным инструментом, позволяющим не просто публиковать наборы данных, но и интегрировать их, обеспечивая поиск и отображение информации об интересующем исследователя таксоне живых организмов одновременно по всем опубликованным наборам данных.

Развитие информационных технологий в области биоразнообразия уже давно прошло первый этап разработки отдельных узконаправленных (по региону или по таксономической группе) баз данных, прошло второй этап отрицания целесообразности разработки и ведения индивидуальных баз данных, на котором предпринимались попытки интеграции таких баз данных в единую информационную систему. Третий (современный) этап развития этих систем находится в фазе «отрицания отрицания»: разработан единый стандарт обмена данными в области биологического разнообразия Darwin Core (Darwin …, 2012), создано специально программное обеспечение – Integrated Publishing Toolkit (The GBIF …, 2014) – для публикации данных о биоразнообразии в единой информационной системе GBIF, но при этом разработчикам и кураторам индивидуальных баз данных предоставлена полная свобода в выборе технологии сбора, хранения, обработки своих данных. Важно, что в любое время разработчик (куратор) отдельной информационной системы по биоразнообразию может принять решение о публикации части или всех накопленных данных в системе GBIF.

Источником данных для публикации в GBIF могут быть любые хранилища данных: от простых текстовых файлов в формате csv до реляционных СУБД. Соответствие между внутренним стандартом хранения данных в индивидуальной информационной системе и стандартом Darwin Core устанавливается при помощи процедуры сопоставления (mapping), реализованной в Integrated Publishing Toolkit. Более подробно порядок подготовки и публикации наборов данных по биологическому разнообразию в информационной системе GBIF изложен в работе М.П. Шашкова с соавторами (Шашков, 2017). Отметим, что Институт биологии Коми НЦ УрО РАН зарегистрирован как публикатор данных в GBIF и поддерживает отдельный экземпляр IPT на своем вебсервере: http://ib.komisc.ru:8088/ipt.

Развитие системы GBIF привело к появлению нового типа статей в рецензируемых журналах – статьи о данных. Согласно концепции, предложенной в работе V. Chavan и L. Penev (Chavan, 2011), статья о данных (data paper) – это особый вид публикации в рецензируемых научных журналах, основной целью которой является описание набора (наборов) данных, но не описание результатов исследований. В отличие от классической научной статьи она содержит факты о данных, но не о гипотезах и результатах их проверки, которые были получены с использованием этих данных. В этой же работе указаны три причины появления данного типа статей: 1) дать возможность научному сообществу ссылаться в привычной форме на работу коллег, 2) обеспечить структурированное, пригодное для чтения человеком описание данных и 3) довести до широкой научной общественности факт существования набора данных.

Очевидно, что главная причина появления статей о данных – это временное отсутствие общепринятой практики признания факта создания и публикации набора данных самостоятельным результатом труда ученого наравне с публикацией статьи в научном журнале. В этом смысле статьи о данных напоминают пароход с парусами – временный гибрид технологий, характерный для ранних этапов внедрения технических новшеств. Уже сейчас ничто не мешает ссылаться на наборы данных, опубликованные, например, в GBIF, также ничто не мешает признать публикацию набора данных в этой базе результатом труда, который учитывается при карьерном росте и оплате труда ученого. Тем не менее, строгий процесс рецензирования, который проходит статья о данных перед публикацией, действительно помогает исследователю повысить качество опубликованных данных. Таким образом публикация статьи о своем наборе данных позволит не только заработать свои академические очки и баллы, но и, безусловно, окажет положительное влияние на ваши способы сбора, первичной обработки и описания данных.

Здесь мы рассмотрим особенности подготовки статьи о данных в журналы издательства Pensoft Publishers (https://pensoft.net), одним из основателей которого является автор концепции «data paper» Любомир Пенев (Chavan, 2011; Strategies …, 2017). Интегрированный инструмент для публикации данных в базе данных GBIF – IPT – предоставляет определенные средства автоматизации процесса подготовки рукописи статьи о данных. При подготовке статьи о данных основным руководством к действию должны служить, разумеется, правила для авторов. На сайтах журналов издательства Pensoft Publishers правила для авторов статей о данных изложены достаточно подробно (Data Publishing Guidelines, 2017). Остановимся на некоторых нюансах подготовки статьи о данных, которые трудно уловимы из текста правил для авторов либо вообще не отражены в них.

Перед началом работы над статьей нужно обратить особое внимание на права, которые вы как автор и публикатор данных передаете обществу. Если политика GBIF позволяет публиковать данные с лицензиями, которые, например, ограничивают право на коммерческое использование данных (CC-BY-NC, https://creativecommons.org/licenses/by-nc/4.0), то Pensoft Publishers не позволяет использовать лицензии, накладывающие подобные ограничения.

Если GBIF не накладывает никаких ограничений на объем, географический и временной охват публикуемых данных, то издатель статей о данных оставляет за собой право принимать к публикации рукописи, только содержащие «значимый для науки» набор данных.

Подготовку рукописи статьи целесообразно начинать с заполнения всех полей метаданных,

которые доступны в IPT. Разобраться с тем, какую именно информацию нужно вносить в каждое из полей каждого раздела метаданных, помогают подробные подсказки во всплывающих окнах и наличие русскоязычного интерфейса. В сложных случаях рекомендуется переключаться между языками интерфейса, чтобы точнее интерпретировать значение тех или иных терминов. Текст, вводимый в поля метаданных, должен соответствовать общепринятому стилю изложения научной информации на английском языке.

После публикации первой версии набора данных на сайте GBIF на странице с описанием набора данных используемой вами инсталляции IPT можно будет скачать файл в формате RTF, который будет оформлен как рукопись статьи с соблюдением правил для авторов Pensoft Publishers. Данный файл будет представлять собой черновик статьи о данных, который потребует ручной доработки, дополнения иллюстрациями, ссылками на литературу. Вероятнее всего, в процессе подготовки рукописи вам придется несколько раз вносить уточнения и в описание метаданных, и в сами публикуемые данные. IPT предоставляет удобную возможность публиковать новые версии набора данных и метаданных с сохранением истории изменений.

Обязательно скачайте свой набор данных в формате DwC-A и внимательно просмотрите, как выглядят ваши данные с точки зрения их конечного пользователя. В них не должно быть дублирующихся записей (строк), дублирующихся полей (столбцов). В текстовых данных не должно быть грамматических ошибок, пробелов в начале и конце строки, пробельных символов, не являющихся пробелом (неразрывный пробел, перевод каретки, разрыв строки, табулятор и т.п.). Загрузив координаты точек находок из вашего набора данных в какую-либо геоинформационную систему, проверьте их на отсутствие грубых ошибок в географической привязке данных (например: точки, находящиеся в море, для сухопутных видов, точки с координатами 0.0000, 0.0000). Примером статьи о данных может служить работа, посвященная распространению борщевика Сосновского на территории Республики Коми (Distribution ..., 2017).

ЛИТЕРАТУРА

Шашков, М. П. Методические рекомендации по стандартизации данных для публикации через глобальный портал GBIF.ORG и подготовке статьи о данных / М. П. Шашков, И. Ф. Чадин, Н. В. Иванова // Труды Кольского НЦ РАН. – 2017. – Т. 6/2017(8), № 5. – С. 22–36. doi: 10.1371/journal.pone.0102623

Chavan, V. The data paper: a mechanism to incentivize data publishing in biodiversity science / V. Chavan, L. Penev // BMC Bioinformatics. – 2011. – Т. 12, № 15. – P. 00121. doi: 10.1186/1471-2105-12-S15-S2. DOI: 10.1186/1471-2105-12-S15-S2

Dalke, I. Distribution Of Heracleum Sosnowskyi In Syktyvkar City. Vector (Polygon) Dataset In Shapefile. [Электроный ресурс] / I. Dalke, I. Chadin, I. Zakhozhiy. – 2018. – Режим доступа: https://zenodo.org/record/1203598.8. doi:10.5281/zenodo.1203598

Darwin Core: An Evolving Community-Developed Biodiversity Data Standard / J. Wieczorek, D. Bloom, R. Guralnick, S. Blum, M. Doring, R. Giovanni, T. Robertson, D. Vieglais // PLOS ONE. – 2012. – Vol. 7, N 1. – P. e29715. doi: 10.1371/journal.pone.0029715

Data Publishing Guidelines [Электроный ресурс] / Pensoft Publishers. – 2018. – Режим доступа: https://phytokeys.pensoft.net/about#DataPublishingGuidelines.

Distribution of the invasive plant species Heracleum sosnowskyi Manden. in the Komi Republic (Russia) / I. Chadin, I. Dalke, I. Zakhozhiy, R. Malyshev, E. Madi, O. Kuzivanova, D. Kirillov, V. Elsakov // PhytoKeys. – 2017. – V. 77. – P. 71. doi: 10.3897/phytokeys.77.11186

Point of View: How open science helps researchers succeed / E. C. McKiernan, P. E. Bourne, C. T. Brown, S. Buck, A. Kenall, J. Lin, D. McDougall, B. A. Nosek, K. Ram, C. K. Soderberg, J. R. Spies, K. Thaney, A. Updegrove, K. H. Woo, T. Yarkoni // eLife. – 2016. – T. 5. – C. e16800. doi: 10.7554/eLife.16800

Strategies and guidelines for scholarly publishing of biodiversity data / L. Penev, D. Mietchen, V. Chavan, G. Hagedorn, V. Smith, D. Shotton, E. O. Tuama, V. Senderov, T. Georgiev, P. Stoev, Q. Groom, D. Remsen, S. Edmunds // Research Ideas and Outcomes. – 2017. – T. 3. – P. e12431. doi: 10.3897/rio.3.e12431

The GBIF Integrated Publishing Toolkit: Facilitating the Efficient Publishing of Biodiversity Data on the Internet / T. Robertson, M. Doring, R. Guralnick, D. Bloom, J. Wieczorek, K. Braak, J. Otegui, L. Russell, P. Desmet // PLOS ONE. – 2014. – Vol. 9, N 8. – P. e102623.

ZENODO AND GBIF: TOOLS FOR SCIENTIFIC PRIMARY DATA PUBLICATION

I.F. Chadin

Institute of Biology of Komi Scientific Centre of the Ural Branch of the Russian Academy of Sciences, Syktyvkar

Summary. Open science is a new scientific communication culture phenomenon. Open publication of primary scientific data gives new opportunities for the data authors as well as accelerates the development of science. This publication ensures the safety of data, their reproducibility, contributes to the high quality of research makes it more easy to start new collaboration. In this paper, we briefly described the possibilities of two tools for the publication of experimental and observational datasets: the repository for primary scientific data of General purpose – Zenodo (https://zenodo.org) and Global Biodiversity Information Facility – GBIF (https://www.gbif.org). Zenodo allows you to publish any material in any field of science. The size of one record should not exceed 50 GB. Each data set is assigned a digital object identifier – DOI. An example of the such publication is the map of the distribution of Heracleum sosnowskyi on the territory of Syktyvkar (https://zenodo.org/record/1203598). Unlike Zenodo, GBIF is a specialized service that allows not only to publish datasets but also to integrate them, providing search and display information about the taxon of interest to the researcher simultaneously across all published datasets. Integration of heterogeneous data sources is provided by standard Darwin Core and special software – Integrated Publishing Toolkit. The development of the GBIF system led to the emergence of a new type of articles in peer-reviewed journals – articles about the data (data papers). In contrast to the classical primary resaerch paper, it contains facts about the data, but not about the hypotheses and the results of their verification, which were obtained with these data. One of data paper example can be found here: https://phytokeys.pensoft.net/articles.php?id=11186.

Key words: Zenodo, GBIF, open science, data management