Зоологические исследования / Zoologicheskie Issledovania. 2018. No. 20. С. / P. 169–174

ВОЗМОЖНОСТИ, РЕШЕНИЯ И ИНСТРУМЕНТЫ GBIF ДЛЯ ОЦИФРОВКИ И РАЗВИТИЯ ЕСТЕСТВЕННОНАУЧНЫХ КОЛЛЕКЦИЙ

М.П. Шашков[1,2], **Н.В. Иванова**[1,2], **Н.В. Филиппова**[3], **Д.С. Щигель**[4]

[1]*Институт физико-химических и биологических проблем почвоведения РАН, Пущино, Россия, max.carabus@gmail.com*
[2]*Институт математических проблем биологии РАН, филиал ИПМ имени М.В. Келдыша РАН, Пущино, Россия, natalya.dryomys@gmail.com*
[3]*Югорский государственный университет, Ханты-Мансийск, Россия, filippova.courlee.nina@gmail.com*
[4]*Секретариат GBIF, Копенгаген, Дания, dschigel@gbif.org*

Свободный доступ к данным коллекций через сеть Интернет повышает их научное использование и востребованность коллекционного и музейного дела в современной науке. GBIF.org является крупнейшей в мире универсальной поисковой системой по данным естественнонаучных коллекций и другим источникам сведений о распространении видов и включает более 1 миллиарда записей. GBIF — независимая межправительственная организация, растущее международное сообщество и распределённая система, которая обеспечивает свободный доступ к данным о видах. Использование международных стандартов и инструментов GBIF.org обеспечивают совместимость разнородных данных и их эффективный поиск. Наиболее распространённым стандартом GBIF является Darwin Core. Данные публикуются организациями с сохранением авторства и возможностью их цитирования через DOI. К 2018 г. 29 российскими организациями опубликованы более 1,3 млн записей.

PERSPECTIVES, SOLUTIONS AND TOOLS FROM GBIF FOR DIGITIZATION AND DEVELOPMENT OF NATURAL HISTORY COLLECTIONS

M. Shashkov[1,2], **N. Ivanova**[1,2], **N. Filippova**[3], **D. Schigel**[4]

[1]*Institute of Physicochemical and Biological Problems in Soil Sciences of RAS, Pushchino, Russia, max.carabus@gmail.com*
[2]*Institute of Mathematical Problems of Biology of RAS, Branch of the Keldysh Institute of Applied Mathematics of RAS, Pushchino, Russia, natalya.dryomys@gmail.com*
[3]*Ugra State University, Khanty-Mansiysk, Russia, filippova.courlee.nina@gmail.com*
[4]*GBIF Secretariat, Kopenhagen, Denmark, dschigel@gbif.org*

Open access to data increases the demand for and scientific use of collection and museum resources by modern science. GBIF.org is the world's largest universal index for biodiversity data on species distributions from the collections and other sources, and includes more than 1 billion records. GBIF is a voluntary network of governments and institutions, a growing international community of practice, and a distributed infrastructure providing free and open access to biodiversity data. International data standards and GBIF tools enable effective aggregation and efficient searching of disconnected data sources. The most common biodiversity data standard is Darwin Core. Data are published by the organizations, authorships is preserved, and data citation is enabled by DOIs. By 2018, 1.3 M records has been published by the 29 Russian organizations.

Введение

Оцифровка естественнонаучных коллекций является мощным инструментом для систематизации коллекционных фондов и повышения эффективности управления ими через быстрый доступ к образцам и связанной с ними информации. Доступ к данным ведёт к повышению научного использования коллекций и востребованности коллекционного и музейного дела в современной науке и к увеличению числа научных сотрудников, использующих подобные данные в процессе выполнения своих исследований. Доступ к коллекционным фондам через сеть Интернет позволяет избежать длительной (и зачастую малоэффективной) переписки с кураторами коллекций и экономит время и ресурсы исследователя, связанные непосредственно с посещением фондов и работой с образцами. Для решения многих биологических задач достаточно оцифрованной информации этикеток и изображений, а виртуальная копия коллекции способствует лучшей сохранности физических фондовых образцов.

Важно отметить, что перечисленные выше преимущества цифровых коллекций доступны только при условии соответствующего качества оцифровки, обработки и систематизации накопленных данных, а также обеспечения их доступности для исследователей. Для повышения результативности научного использования разрозненных данных недостаточно разместить электронные каталоги на веб-сайтах коллекций и музеев. Эффективным решением для обнаружения контактной информации коллекции, общего описания её фондов, данных об образцах или их изображений являются международные тематические порталы по биоразнообразию. Участие в таких глобальных системах позволяет отдельным коллекциям, а также полевой биологии и музейному делу как таковым выйти из цифровой, научной и грантовой тени, в которой эти области научной деятельности оказалась в связи с бурным развитием молекулярных дисциплин, работающих с первоначально цифровыми, «born digital», данными.

В России международные порталы по биоразнообразию пока остаются мало востребованными, а зачастую просто малоизвестными. Во многом это объясняется недостатком информации о принципах их работы, правилах использования данных и опыте применения подобных ресурсов в научных исследованиях. В публикации мы описываем основные принципы работы крупнейшего в мире портала по биоразнообразию GBIF (Global Biodiversity Information Facility), Глобальной информационной системы по биоразнообразию, и возможности для интеграции в него данных оцифрованных коллекций.

Стандарты представления данных естественнонаучных коллекций

Большинство отечественных и многие зарубежные коллекции используют для хранения данных собственные локальные системы и стандарты, появившиеся в процессе их разработки. При этом современные таксономические ревизии, филогенетические и биоклиматические модели, природоохранные проекты и другие научные продукты основываются на данных более чем одной коллекции (Patel et al., 2016; Del Olmo-Ruiz et al., 2017; Pagad et al., 2018). Объединение данных из разных источников — повседневная реальность современной научной деятельности. В масштабах планеты многообразие языков, номенклатур, форматов представления данных (например, дат, географических координат и др.) создавало проблемы для их объединения и масштабирования. Появилась необходимость разработки универсальных стандартов данных, использование которых позволяет объединять информацию, происходящую из разных источников. К настоящему времени наиболее значимым и распространённым для естественнонаучных коллекций является стандарт Darwin Core, сокращенно DwC (Wieczorek et al., 2012). DwC — это набор полей (терминов), с помощью которых представляется атрибутивная информация о находках видов или коллекционных образцах, и правила заполнения этих полей (http://rs.tdwg.org/dwc/terms). Иными словами, данные в DwC — это электронная таблица, заголовки столбцов которой соответствуют терминам, а строки — образцам. Обязательное наличие среди заголовков идентификатора и ограничения, накладываемые описанием терминов (типизация), делают данную таблицу небольшой базой данных, в которой термины DwC являются полями, а описания образцов — записями.

Возможности DwC позволяют подробно описывать как данные о коллекционных образцах, так и сведения о коллекциях, в которых они хранятся. Так, можно указать акроним коллекции, её идентификатор в международных музейных системах, название и веб-сайт

учреждения, в котором она располагается. Для образцов, помимо стандартной информации о таксономическом положении, исследователях, собравших и определивших данный образец, дате и месте сбора, можно привести данные о способе хранения, состоянии образца, методе и точности географической привязки места сбора, ссылки на изображения или другие связанные медиа-ресурсы и др. Использование единого стандарта обеспечивает совместимость данных и их эффективный поиск в глобальных и тематических информационных системах и не требует модификации структуры исходных локальных баз данных или информационных массивов. Существующий набор терминов DwC позволяет хранить не только формализованную информацию об образце, но и исходные «дословные» данные. На сегодняшний день стандарт DwC используется как основной в GBIF, Информационной биогеографической системе о морских организмах OBIS, Орнитологической информационной системе ORNIS и др.

Возможности GBIF для публикации, индексации и поиска данных

На настоящий момент крупнейшим в мире универсальным ресурсом по цифровой информации коллекций и других источников данных о распространении видов является портал GBIF.org. В июле 2018 г. число записей, доступных через GBIF, превысило 1 млрд. GBIF — независимая межправительственная организация и растущее международное сообщество, которое обеспечивает свободный доступ к данным коллекций и другим источникам информации о распространении видов. Это создаёт условия для сотрудничества и поддерживает экономически эффективный обмен, поиск и использование цифровых данных для научных исследований и практических решений в природопользовании. Кроме данных, предоставленных примерно 1200 отдельными организациями и коллекциями, в GBIF представлены данные, которые объединяются международными сообществами, такими как Охрана флоры и фауны Арктики (Conservation of Arctic Flora and Fauna, CAFF), Международная сеть наблюдений за птицами (eBird), Европейский совет по учётам птиц (European Bird Census Council, EBCC), Международная сеть наблюдений за организмами iNaturalist и др. Принципиальной основой работы GBIF является свободный и бесплатный доступ ко всем данным.

С 2014 г. в России наблюдается постоянный рост наборов данных, числа записей и публикующих организаций. В начале июня 2018 г. из более 2,5 млн записей по России более 1,3 млн опубликованы 29 российскими организациями в 42 наборах данных. Среди наиболее активных участников — научно-исследовательские институты РАН, МГУ имени М.В. Ломоносова, Югорский государственный университет, заповедники и национальные парки (https://www.gbif.org/publisher/search?offset=0&country=RU). Наблюдается рост интереса к публикации данных через Российские узлы IPT в других странах постсоветского пространства. Этот сегмент сообщества GBIF координируется командой GBIF.ru при поддержке Секретариата GBIF — авторами данной статьи.

В настоящее время данные в системе GBIF публикуются от имени организаций (музеев, коллекций и т. п.) в виде гомогенных наборов (datasets) с сохранением авторства. Организация (data publisher) самостоятельно принимает решение о подробности публикуемых ею данных и устанавливает однозначные правила (согласно выбранному типу лицензии) для повторного использования этой информации. Каждый опубликованный набор данных получает уникальный идентификатор цифрового объекта (DOI) и имеет постоянную веб-страницу на глобальном портале. Портал GBIF.org обеспечивает поиск по названиям таксонов, геопривязке, датам и другим параметрам. Результат каждого поискового запроса почти всегда является объединением данных из разных наборов, табличному результату запроса присваивается уникальный DOI. Начиная с 2015 г. эти DOI содержат ссылки на источники происхождения и авторства всех данных, использованных в запросе. Большая часть данных в GBIF публикуется по лицензиям Creative Commons (https://creativecommons.org/licenses/) CC BY («С указанием авторства») и CC BY-NC («С указанием авторства — Некоммерческая»). Авторы научных работ, запрашивающие данные через GBIF.org, должны цитировать полученные данные, указывая DOI, практика цитирования с помощью DOI постоянно расширяется, приближаясь к системе sequence accession numbers в биоинформатике. Секретариат GBIF ведёт систематическую работу с авторами публикаций, международными издательствами и журналами для дальнейшего внедрения практики цитирования данных с помощью DOI в научных текстах. Эта практика обеспечивает соблюдение условий

лицензии, заявленной публикующей организацией, независимо от размера вклада использованного набора в цитируемый запрос. Таким образом, наборы данных по биоразнообразию и т.н. статьи о данных, «data papers» (подробнее см. Chavan, Penev, 2011; Шашков и др., 2017; https://www.gbif.org/data-papers) становятся новым и востребованным цифровым продуктом научной деятельности естественнонаучных музеев и коллекций.

Публикация данных в GBIF осуществляется как вручную, так и с помощью специального инструмента Integrated Publishing Toolkit, IPT (Robertson et al., 2014), функционирующего как серверное приложение с визуальным web-интерфейсом. Удобство IPT заключается в том, что настройку соответствия структуры экспортируемых исходных данных в DwC (т. н. «mapping»), а также частоту автоматического обновления, нужно выполнить лишь один раз. На сегодняшний день в мире работает 229 IPT-инсталляций в 69 странах, 5 из которых находятся в России. К одной IPT может быть «привязано» несколько организаций, каждая из которых может иметь несколько аккаунтов для сотрудников с разными правами в отношении публикации данных. Важно отметить, что GBIF является распределённой системой, и публикация данных лишь делает их обнаружимыми для поисковых запросов, в то время как сами данные остаются под полным контролем публикующей организации, хранятся на собственном или самостоятельно выбранном сервере с IPT и могут быть обновлены или удалены ей на любом этапе. Помимо GBIF.org публикацию через IPT осуществляют тематические порталы OBIS, VertNet и Global Genome Biodiversity Network (GGBN).

Система GBIF также позволяет предоставлять сведения о массивах данных, которые ещё не оцифрованы и/или пока не опубликованы через глобальные порталы. Для этого существует специальный веб-инструмент Suggest a dataset (https://www.gbif.org/suggest-dataset), который создан с целью привлечения новых источников данных о биоразнообразии для заполнения «белых пятен» на цифровой карте GBIF. Воспользоваться им может любой зарегистрированный на портале пользователь. Достаточно указать название набора данных, а также их таксономический и географический охват.

Одним из перспективных путей активизации оцифровки коллекций и публикации данных в GBIF является использование для менеджмента данных готовых программных продуктов с открытым исходным кодом. Важным преимуществом таких программ является хранение данных в формате DwC или возможность их экспорта в этот формат, а также настройка автоматической публикации новых версий наборов данных, по мере роста и обновления исходного массива. Так, для работы с базами данных биологических коллекций широко используется платформа Specify (http://www.sustain.specifysoftware.org). Высокой популярностью для управления ботаническими коллекциями и анализа их данных пользуется система BRAHMS (https://herbaria.plants.ox.ac.uk/bol). Для организации порталов коллекций, электронных флор и интерактивных ключей применяется продукт Symbiota (Gries et al., 2014; http://symbiota.org/docs). С полным перечнем продуктов, рекомендуемых группой по оцифровке коллекций iDigBio, можно ознакомиться на странице https://www.idigbio.org/wiki/index.php/Electronic_Data_Capture.

Широкие перспективы для развития национальных порталов по биоразнообразию представляют программные решения Atlas of Living Australia, ALA (https://demo.gbif.org/programme/living-atlases), предоставляемые группой разработчиков из Австралии. При поддержке GBIF, ALA Tools используются такими странами как Испания, Португалия, Шотландия, Франция, Канада, Аргентина и Коста-Рика для развития на их основе собственных национальных систем.

Электронные коллекции в мире и в России

К настоящему времени в цифровой формат переведены многие известные мировые коллекции и более 200 из них обнаруживаемы через GBIF.org. Крупнейшими являются цифровые коллекции Британского музея естественной истории (3 673 699 записей об образцах, Natural…, 2018), Национального музея истории и науки Японии (4 093 085 записей, 342 набора данных), Американского музея естественной истории (1 492 196 записей, Trombone, 2013, 2016; Dickey, 2016; Hussaini, 2017) и др. Использование единого стандарта представления данных и свободный доступ к ним позволяют использовать эту информацию для анализа в составе объединённых массивов. Например, данные об образцах, опубликованные музеем

Университета провинции Альберта (Канада), процитированы в 50 научных публикациях, музея Австралии — в 36, Ботанического музея г. Лунд (Швеция) — в 42.

В России процесс перевода в цифровой формат национальных биологических коллекций находится на начальном этапе (Филиппова и др., 2017). Крупнейшей российской оцифрованной коллекцией является гербарий МГУ имени М.В. Ломоносова (MW; Серегин, 2017; также настоящий сборник); также оцифрована часть гербарных фондов Ботанического института имени В.Л. Комарова РАН (LE) и Института экологии растений и животных УрО РАН (SVER). Среди зоологических коллекций систематическая масштабная работа по их оцифровке проводится, по всей видимости, только в Зоологическом институте РАН и в Зоологическом музее МГУ. К настоящему времени в ЗИН РАН оцифрованы 3123 типовых образца (https://www.zin.ru/Collections/collections.html), часть которых доступна через портал GBIF (Milto et al., 2018; Sinev et al., 2018; Smirnov et al., 2018a, 2018b, Volkovitsh et al., 2018). Из фондов Зоологического музея МГУ в настоящее время через Интернет доступны данные о 131079 образцах животных, относящихся к 6497 видам (https://animal.depo.msu.ru).

Заключение

Организация доступа к оцифрованным данным фондов биологических коллекций через глобальные порталы повышает их доступность и востребованность для исследователей. Использование международных стандартов и веб-инструментов GBIF.org обеспечивают совместимость данных, полученных из разных источников, и их эффективный поиск в тематических и глобальных информационных системах.

Литература

Серегин А.П. 2017. Цифровой гербарий МГУ — крупнейшая российская база данных по биоразнообразию. — Известия РАН. Сер. биол., 6: 30–36.

Шашков М.П., Чадин И.Ф., Иванова Н.В. 2017. Методические рекомендации по стандартизации данных для публикации через глобальный портал GBIF.org и подготовке статьи о данных. — Известия Кольского научного центра РАН. Сер. Прикладная экология Севера, 3 (45): 22–35.

Филиппова Н.В., Филиппов И.В., Щигель Д.С., и др. 2017. Информатика биоразнообразия: мировые тенденции, состояние дел в России и развитие направления в Ханты-Мансийском Автономном Округе. — Динамика окружающей среды и глобальные изменения климата, 8 (2): 46–56. http://journals.eco-vector.com/EDGCC/article/view/7080.

Chavan V., Penev L. 2012. The data paper: a mechanism to incentivize data publishing in biodiversity science. — BMC Bioinformatics, 12 (15): S2. doi: 10.1186/1471-2105-12-S15-S2.

Dickey D. 2016. AMNH Herpetology Collections. American Museum of Natural History. Occurrence dataset. https://doi.org/10.15468/jfkgyh.

Gries C., Gilbert E., Franz N. 2014. Symbiota — A virtual platform for creating voucher-based biodiversity information communities. — Biodiversity Data Journal, 2: E1114. doi 10.3897/BDJ.2.e1114.

Hussaini B. 2017. AMNH Invertebrate Paleontology Collection. Version 1.10. American Museum of Natural History. Occurrence dataset. https://doi.org/10.15468/e3soei.

Natural History Museum. 2018. Natural History Museum (London) Collection Specimens. Occurrence dataset. https://doi.org/10.5519/0002965.

Milto K., Ananjeva N., Golikov A., Khalikov R. 2018. Catalogue of the type specimens of Bufonidae and Megophryidae (Amphibia: Anura) from research collections of the Zoological Institute, Russian Academy of Sciences. Ver. 1.23. Zoological Institute, Russian Academy of Sciences, St. Petersburg. Checklist dataset. https://doi.org/10.15468/crgfcq.

Robertson T., Döring M., Guralnick R., et al. 2014. The GBIF Integrated Publishing Toolkit: Facilitating the Efficient Publishing of Biodiversity Data on the Internet. — PLoS ONE, 9(8): e102623. https://doi.org/10.1371/journal.pone.0102623.

Sinev S., Golikov A., Khalikov R. 2018. Catalogue of the type specimens of Cosmopterigidae (Lepidoptera: Gelechioidea) from research collections of the Zoological Institute, Russian Academy of Sciences. Ver. 1.23. Zoological Institute, Russian Academy of Sciences, St. Petersburg. Checklist dataset. https://doi.org/10.15468/sbga6b.

Smirnov R., Golikov A., Khalikov R. 2018. Catalogue of the type specimens of Pogonophora (Annelida; seu Polychaeta: Siboglinidae) from research collections of the Zoological Institute, Russian Academy of Sciences. Ver. 1.15. Zoological Institute, Russian Academy of Sciences, St. Petersburg. Checklist dataset. https://doi.org/10.15468/1mlkdp. a

Smirnov I., Golikov A., Khalikov R. 2018. Ophiuroidea collections of the Zoological Institute Russian Academy of Sciences. Version 1.41. Zoological Institute, Russian Academy of Sciences, St. Petersburg. Occurrence dataset. https://doi.org/10.15468/ej3i4f. b

Trombone T. 2013. AMNH Bird Collection. American Museum of Natural History. Occurrence dataset. https://doi.org/10.15468/xvzdcm.

Trombone T. 2016. AMNH Mammal Collections. American Museum of Natural History. Occurrence dataset. https://doi.org/10.15468/wu3poe.

Volkovitsh M., Golikov A., Khalikov R. 2018. Catalogue of the type specimens of Polycestinae (Coleoptera: Buprestidae) from research collections of the Zoological Institute, Russian Academy of Sciences. Ver. 1.22. Zoological Institute, Russian Academy of Sciences, St. Petersburg. Checklist dataset. https://doi.org/10.15468/c3eork.

Wieczorek J., Bloom D., Guralnick R., et al. 2012. Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. — PLoS ONE, 7 (1): e29715. doi: 10.1371/journal.pone.0029715.