УДК 004.422.81+58.087

М. П. Шашков[1,2], И. Ф. Чадин[3], Н. В. Иванова[1]
[1] Институт математических проблем биологии РАН – филиал Федерального государственного учреждения «Федеральный исследовательский центр Институт прикладной математики им. М. В. Келдыша РАН»
[2] Институт физико-химических и биологических проблем почвоведения РАН
[3] Институт биологии Коми НЦ Уральского отделения РАН

МЕТОДИЧЕСКИЕ РЕКОМЕНДАЦИИ ПО СТАНДАРТИЗАЦИИ ДАННЫХ ДЛЯ ПУБЛИКАЦИИ ЧЕРЕЗ ГЛОБАЛЬНЫЙ ПОРТАЛ GBIF.ORG И ПОДГОТОВКЕ СТАТЬИ О ДАННЫХ

Аннотация

Обсуждены основные принципы и процедура публикации данных через GBIF.org, приводятся рекомендации по использованию стандарта Darwin Core и оценке качества данных. Впервые на русском языке описан опыт публикации статьи о данных.

Ключевые слова:

биоразнообразие, открытые данные, стандарт Darwin Core, качество данных, геопривязка.

M. P. Shashkov, I. F. Chadin, N. V. Ivanova

GUIDE TO BEST PRACTICES ON A DATA STANDARDIZATION FOR PUBLICATION VIA THE GLOBAL PORTAL GBIF.ORG AND PREPARING DATA PAPER

Abstract

The basic principles and procedure for datasets publishing through GBIF.org are discussed. Recommendations on the use of Darwin Core standard and data quality assessment are given. The experience of data paper publishing is described for first time in Russian.

Keywords:

biodiversity, open data, Darwin Core standard, data quality, georeference

Введение

Интеграция данных о находках биологических видов является актуальной задачей в области изучения биоразнообразия. Конвенция о Биологическом разнообразии (1992), ратифицированная Россией в 1995 г., а также Айтинские целевые задачи (2010) предполагают совместное использование данных о биоразнообразии странами-участницами конвенции, а также развитие научной базы и технологий, связанных с оценкой биологических ресурсов.

В настоящее время существует несколько крупных международных систем, объединяющих данные о биоразнообразии, крупнейшей из которых является Глобальная Информационная Система по Биоразнообразию GBIF (gbif.org), созданная в 2001 году на основе межправительственных соглашений как научная инфраструктура, обобщающая данные биологических коллекций. Структура GBIF представляет собой сеть взаимосвязанных национальных «узлов» (*Nodes*) стран-участниц, крупных международных организаций и отдельных институтов в странах, не являющихся официальными участниками

GBIF. На момент подготовки этой статьи 1064 организации (*Data Publishers*) через GBIF.org опубликовали 35885 наборов данных (*Datasets*), со сведениями о 797353488 находках (*Occurrences*) 1727442 видов. Все данные, опубликованные через GBIF.org, являются открытыми, т. е. доступны для загрузки любому зарегистрированному на портале пользователю.

Потребность «открытия» информации, т. е. публикации первичных данных все более осознается мировым научным сообществом. Постоянно растущий объем «открытых» данных способствует повышению качества исследований, электронная публикация через глобальные порталы позволяет демонстрировать использование и цитирование данных, увеличивает их потенциал для повторного использования в междисциплинарных исследованиях в составе объединенных массивов (Penev et al., 2017).

Информация, опубликованная в электронном виде, также позволяет опубликовать описание набора данных в виде отдельной научной статьи — т.н. статьи о данных (*Data paper*). Согласно концепции, предложенной в работе V. Chavan и L. Penev (2011), статья о данных – это особый вид публикации в рецензируемых научных журналах, основной целью которой является не анализ результатов исследований, а подробное описание набора (наборов) данных. В отличие от классической научной статьи она содержит факты о данных и их описание, информацию о методах сбора и др., но не о гипотезах и результатах их проверки, которые были получены с использованием этих данных. Такие публикации дают возможность научному сообществу ссылаться в привычной форме на работы коллег, обеспечивают структурированное описание информации и доводят до широкой научной общественности факт ее существования. В настоящее время такие статьи публикуются в 27 международных журналах, в т. ч. индексированных в реферативных базах данных публикаций (Data…, 2017).

В России процесс «открытия» национальных данных о биоразнообразии и их интеграции в международные проекты находится на начальном этапе. Большинство российских ресурсов о биоразнообразии до настоящего времени остаются закрытыми (Иванова, Шашков, 2014), существует лишь несколько открытых тематических информационных систем. Наиболее успешными среди них являются электронный портал гербария МГУ им. М. В. Ломоносова (Seregin, 2016), информационная система по биоразнообразию криптогамных организмов CRIS (Мелехин и др., 2013) и краудсорсинговая информационная система «Фаунистика», объединяющая данные о находках хищных птиц (Веб-ГИС «Фаунистика», 2012). Число российских научных и природоохранных организаций-участников GBIF за предыдущие три года возросло от двух до одиннадцати, был опубликован 21 набор данных (более 140 тыс. находок видов), в 2017 г. опубликована первая российская статья о данных (Chadin et al., 2017). Все это свидетельствует о заинтересованности российского научного сообщества в «открытии» данных и их публикации через международные порталы. В то же время, имеет место ряд технических препятствий. Электронная публикация через GBIF.org осуществляется на основе единого стандарта описания и представления информации. Это облегчает обмен данными между разными ресурсами и позволяет полностью сохранить сведения о находках видов и их правообладателях. Однако в России международные стандарты для описания данных о биоразнообразии используются редко, а национальные стандарты отсутствуют (Иванова, Шашков, 2014), поэтому возникает

необходимость предварительной обработки и форматирования данных (т. е. приведения их к соответствующему стандарту). Имеющийся у авторов опыт проведения в России практических семинаров по работе с GBIF.org показывает, что англоязычных руководств, доступных через глобальный портал (Create…, 2011 и др.), и кратких инструкций на неофициальном русскоязычном сайте GBIF (gbif.ru) недостаточно для подготовки набора данных к публикации.

В работе описывается общая схема публикации через GBIF.org и приводятся рекомендации по стандартизации данных для их мобилизации через глобальный портал и подготовке статьи о данных.

Основные принципы публикации данных через GBIF.org

Большая часть данных, опубликованных через GBIF.org, соответствует единому стандарту Darwin Core, сокращенно DwC (Wieczorek et al., 2012), используются также стандарты ABCD (Access…, 2015) и EML (Ecological…, 2017). Для приведения информации к DwC и публикации через GBIF.org необходим специальный онлайн инструмент Integrated Publishing Toolkit, IPT (Robertson et al., 2014).

Объединение данных в GBIF происходит на основе таксономической системы Backbone Taxonomy (GBIF Backbone…, 2016), которая представляет собой синтез из многих таксономических баз данных, а также некоторых списков видов, опубликованных через GBIF.org. Крупнейшей таксономической базой в ее составе является Catalogue of Life, CoL (2015), содержащий более 3 млн таксонов или около 60% всех данных Backbone. Для каждого таксона могут быть указаны синонимы, а также названия на национальных языках. Важно отметить, что в настоящее время многие российские эндемичные виды отсутствуют в Backbone, поэтому при публикации данных о находках таких видов их таксономическое положение будет автоматически указано до ближайшего известного таксона более высокого ранга. Для внесения новых видов в Backbone необходимо сначала опубликовать список, содержащий новые для системы виды (*Checklist Data*, см. ниже), после чего связаться с Секретариатом с просьбой включения этого списка в набор ресурсов, из которых формируется Backbone. В случае, если он будет включён в новую версию системы, то названия таксонов ранее опубликованных находок будут актуализированы с учетом обновлений Backbone. Отметим, что обновления происходят довольно редко (2–4 раза в год).

Публикацию через глобальный портал осуществляют только зарегистрированные в GBIF организации, которые принимают решение об объеме и подробности публикуемых данных, контролируют их качество, а также имеют возможность отозвать любой из опубликованных ими наборов данных (Costello, 2009). Публикующая организация устанавливает однозначные правила повторного использования информации, которые должны быть понятны потенциальным пользователям. Для описания этих правил существуют различные лицензии (Penev et al., 2017). Секретариат GBIF рекомендует использовать открытые лицензии сообщества Creative Commons (О лицензиях…, 2017), которые позволяют авторам и правообладателям данных объявить об отсутствии ограничений на их использование (CC0), указать необходимость цитирования источника (CC-BY), или заявить о запрете их коммерческого использования (CC-BY-NC).

Каждый набор данных, опубликованный через GBIF.org, получает уникальный идентификатор цифрового объекта (DOI) и имеет постоянную веб-страницу на глобальном портале. Все данные однозначно указывают на организацию, от имени которой они опубликованы и роли лиц, причастных к публикации. Также автоматически формируется ссылка для цитирования данных, в которую включаются как авторы самих данных, так и метаданных. Авторы и представители организации могут отслеживать использование данных по списку публикаций, в которых часть или весь набор данных цитируются с указанием DOI. Тем не менее, механизм контроля за соблюдением пользователями правил использования и цитирования данных, доступных через GBIF.org, отсутствует, т. е. случаи плагиата и несоблюдения лицензионных требований теоретически возможны. Очевидно, что такая ситуация может вызывать беспокойство правообладателей и авторов данных. Однако, важно понимать, что корректное использование и цитирование «обычных» литературных источников также находится под контролем цитирующего и, как показывает практика, в большинстве случаев эти правила соблюдаются. Благодаря наличию DOI у каждого набора данных, случаи недобросовестного их использования выявить несложно, также, как и доказать авторство в отношении собственных данных.

Основные этапы публикации данных через GBIF.org

Процедура электронной публикации данных через GBIF.org при помощи IPT состоит из следующих этапов: (1) регистрация организации в GBIF, (2) выбор типа публикации, (3) составление описания набора данных – метаданных, (4) стандартизация набора данных, если публикация не ограничивается метаданными, и (5) собственно публикация набора данных через IPT как технический процесс.

Регистрация организации в GBIF

Публикацию через GBIF.org следует предварительно согласовать с руководителем организации, от имени которой будут опубликованы данные (письменного подтверждения при этом не требуется). Для регистрации новой организации в GBIF, необходимо заполнить специальную веб-форму (*Request endorsement*), доступную на GBIF.org (Become…, 2017), в которой необходимо указать информацию об организации, а также административные и технические контакты. Все поля заполняются на английском языке. Заявка рассматривается Секретариатом, после чего организация одобряется участниками GBIF. Как правило, процедура одобрения (*Endorsement*) новых участников занимает 1–2 недели. Для каждой организации на портале автоматически создается отдельная веб-страница, где размещается информация о ней и об опубликованных наборах данных. Все данные об организации вносятся на основе информации регистрационной формы, изменить их можно, связавшись с Секретариатом через helpdesk@gbif.org.

Выбор типа публикации

В настоящее время GBIF.org поддерживает публикацию данных, содержащих описания коллекций или тематических электронных ресурсов, таксономические сводки, данные о находках видов (коллекционных образцах), а также результаты площадных и маршрутных учетов и данные мониторинга.

Общая схема организации данных в наборах разных типов с обязательными для заполнения полями представлена на рис. 1, подробное описание приводится ниже.

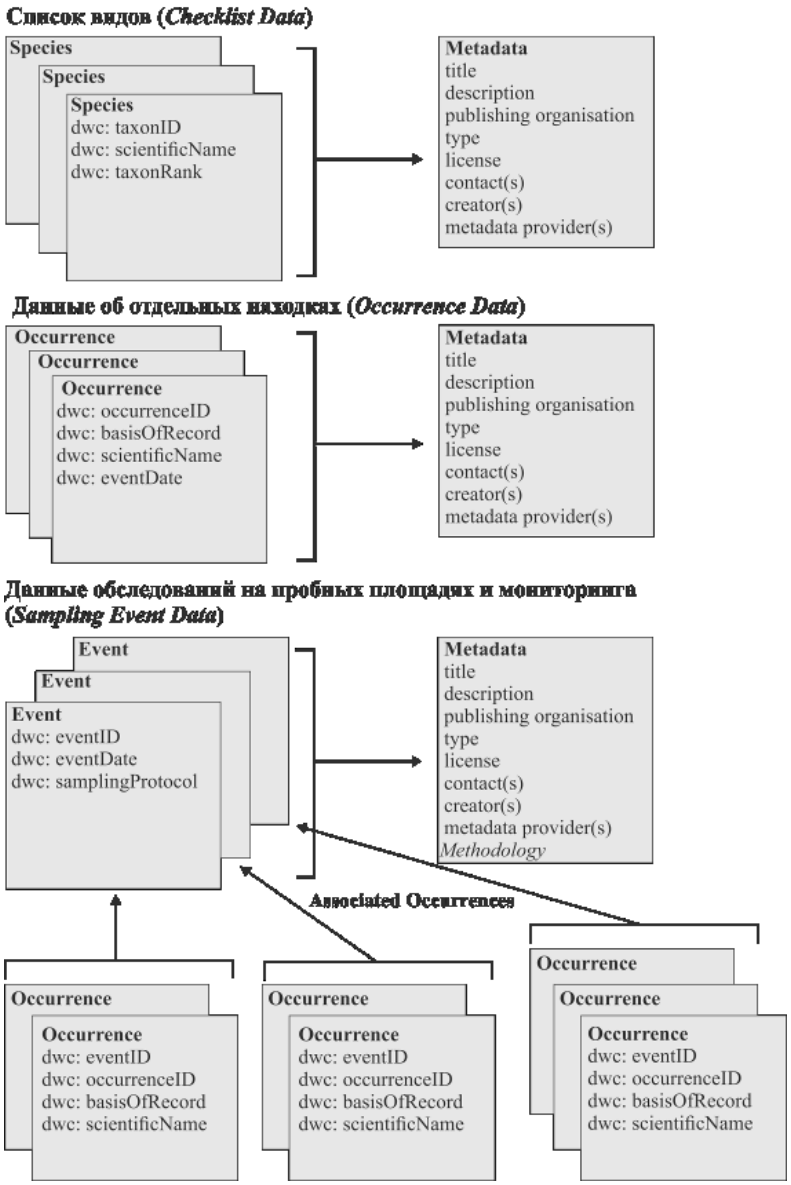

Рис. 1. Типы данных, доступных через GBIF.org,
и обязательные для заполнения поля

1. Метаданные (*Meta Data Only*), или «данные о данных», предполагают только описание набора данных без предоставления исходных сведений. Наборы метаданных представляют собой ценный источник базовой информации о массивах данных, которые недоступны через Интернет. Публикация метаданных позволяет объявить о существовании того или иного массива данных и сделать его потенциально доступным для научного сообщества.

Например, таким способом можно опубликовать описание коллекции, не имеющей электронного каталога, описать узкоспециализированный тематический ресурс, к которому сложно применить стандарты GBIF, или ресурс, данные которого раскрывать неправомерно, но информация о нем является ценной для специалистов. В первую очередь, это касается сводок о распространении редких и охраняемых видов, а также видов, представители которых имеют коммерческую ценность и могут легко стать объектом браконьерства.

Технические возможности IPT позволяют авторам данных предоставить метаданные в объеме, который они считают необходимым (Resource…, 2017). Обязательной информацией являются название, аннотация и указание языка, на котором описаны метаданные, а также имена создателя GBIF-ресурса и автора метаданных. Отметим, что эти функции могут быть определены как для одного, так и для нескольких специалистов. Помимо этого, можно привести ключевые слова, описать географический, таксономический и временной охват данных, методы их сбора, привести список публикаций и проектов, связанных с этими данными. Метаданные рекомендуется приводить на английском языке, хотя публикация метаданных на других языках технически возможна. На сайте GBIF также создана возможность уведомить научное сообщество о существующих неопубликованных данных без использования IPT при помощи веб-инструмента *Suggest a dataset* (https://gbif.org/suggest-dataset). Остальные типы наборов данных, помимо обязательных метаданных, предполагают публикацию сведений о находках видов или списков видов.

2. Список видов (*Checklist Data*) предполагает публикацию таксономического списка. Это может быть список видов для определенной территории (Khanina et al., 2016 и др.), таксономический список рода или более крупного таксона, каталог типовых образцов коллекции (Smirnov et al., 2017; Volkovitsh et al., 2017 и др.) и т. д. Поскольку не все образцы могут быть определены с точностью до вида, то помимо названия таксона указывается его ранг. Также наборы данных этого типа могут включать другую дополнительную информацию, например, синонимы, названия видов на национальных языках и др.

3. Данные об отдельных находках (*Occurrence Data*). Наборы данных этого типа содержат информацию о нахождении того или иного вида в определенном месте в определенное время, т. е. предполагают наличие сведений о дате находки и географической привязке места нахождения вида (или сбора образца). В отличие от следующего типа, данные о находках могут быть не связаны между собой. Эти наборы данных составляют основную часть данных, опубликованных через GBIF.org, и могут происходить из различных источников. В качестве *Occurrence Data* могут быть представлены как непосредственно данные о полевых находках (включая отдельные наблюдения, полученные с помощью автоматических фото- и видеорегистраторов; для серий наблюдений см. *Sampling Event* ниже), так и наборы данных, описывающие коллекционные образцы (Melechin, 2016), материалы летописей природы (Seyfulina, Buyvolov, 2016), литературные данные о находках видов, данные тематических ресурсов (Chadin et al., 2016; Shashkov, Ivanova, 2016; Petrosyan, Kuzmin, 2017) и т. д. Существует также возможность опубликовать список видов со связанными с ним *Occurrence Data*, используя список видов как центральную таблицу (*Core*).

4. Данные обследований на пробных площадях и мониторинга (*Sampling Event Data*). Этот тип публикации является относительно новым для

GBIF и позволяет мобилизовать данные, собранные в результате площадных и маршрутных учетов или мониторинговых исследований в виде методологически связанных блоков информации. Кроме информации об отдельных находках, в таких наборах данных приводятся сведения, относящиеся сразу ко многим находкам, например, это могут быть данные о характеристиках сообщества или методы проведения исследований. Также можно описать обилие какого-либо вида на различных участках или его динамику в течение определенного периода времени. Такие данные собираются с применением тех или иных стандартных методов учетов, сборов или наблюдений. Методы сбора данных кратко описываются в соответствующих полях таблицы с исходными данными. Объем проделанных работ можно оценить, указав способ сбора данных (*samplingProtocol*[2]), площади учета или протяженность маршрутов (*sampleSizeValue* и *sampleSizeUnit*), а также, при необходимости, объем или длительность наблюдений (*sampling Effort*). В метаданных следует приводить подробное описание методов сбора. Особая ценность подобных данных заключается в возможности сопоставления результатов, полученных одинаковыми методами на разных территориях и / или в разное время (Dataset…, 2017).

Стандартизация данных в соответствии Darwin Core

Как было описано выше, данные, публикуемые через GBIF.org, должны соответствовать международным стандартам. Под стандартом мы понимаем набор полей (терминов), описывающих свойства объектов и правила их использования. В настоящее время в GBIF основным является стандарт DwC, разработанный группой Biodiversity Informatics Standards, ранее известной как Taxonomic Databases Working Group (TDWG…, 2017).

Для публикации данные необходимо представить в виде электронной таблицы MS Excel или файла CSV, в которых содержание и заголовки полей (столбцов) соответствуют терминам DwC. При подготовке *Checklist Data* одной записи (строке) в таблице должен соответствовать один таксон, для *Occurrence Data* – одна находка (одновидовой коллекционный образец). В случае *Sampling Event Data* необходимо подготовить две таблицы: первая (*Sampling Events*) содержит описание проб или пробных площадей (одна запись соответствует одной пробной площади, маршруту и т. д.), методов и объемов сборов, а вторая (*Associated Occurrences*) – список таксонов, обнаруженных на каждой пробной площади и их характеристики (одна запись соответствует одной находке). Метаданные набора данных приводятся в отдельном текстовом файле.

Через веб-сервис github.com для каждого типа наборов данных доступны шаблоны (Checklist..., 2017; Occurrence…, 2017; Sampling…, 2017), которые представляют собой электронные таблицы, содержащие DwC-заголовки полей и примеры их заполнения. Выделены обязательные и рекомендуемые к заполнению поля; приводятся примеры для разных объектов. Также в шаблонах содержатся краткие инструкции по подготовке данных.

При стандартизации набора данных необходимо соблюдать правила использования терминов DwC (Darwin…, 2015). Для каждого термина приводится название (*Term Name*), уникальный идентификатор (*Identifier*),

[2]Здесь и далее термины DwC приводятся в оригинальном написании.

категория (*Class*), содержание термина (*Definition*), краткие рекомендации по его использованию (*Comment*) и ссылка на подробное описание (*Details*). Содержание всех терминов DwC и примеры их использования доступны на сайте TDWG, русскоязычное описание наиболее часто употребительных терминов и примеры их использования – на сайте gbif.ru (Иванова, Шашков, 2017).

Текущая версия DwC содержит около 200 терминов, позволяющих описать данные, полученные из разных источников. Более 30 расширений (*extensions*) для специфических данных DwC позволяют описывать характеристики отобранных проб, результаты молекулярно-генетического анализа, аудио- и видеоматериалы, и многие другие атрибуты, обычно не применимые к большинству данных, но в определенных областях исследований имеющие большую ценность. С помощью терминов DwC можно описывать данные о локалитетах, где целевые виды не были обнаружены, что важно для решения задач мониторинга популяций охраняемых видов и изучения динамики их ареалов. Очевидно, что для описания каждого образца невозможно использовать абсолютно все термины DwC, некоторые термины вообще являются взаимоисключающими. В то же время, при описании специфических данных часто возникают ситуации, когда терминов DwC не хватает для описания всей атрибутивной информации. В таких случаях рекомендуется вносить такие данные в примечания (см. ниже) или использовать наиболее близкие по смыслу термины с обязательным указанием того, какие именно данные приводятся в соответствующих полях.

Для удобства пользователей термины DwC сгруппированы в несколько тематических разделов. Раздел *Record-level Terms* используется для описания общих данных об объектах. В нем приводится информация об организации, в которой хранятся данные (образцы) и описываются правила их использования. Если в данных содержится какая-то специфическая атрибутивная информация, которая доступна на стороннем ресурсе и не публикуется через глобальный портал, об этом также стоит указать в этом разделе (термин *informationWithheld*). В разделе *Occurrence* приводятся сведения о состоянии организма во время находки или сбора (численность, поведение, жизненное состояние и др.). Время и методы сборов описываются терминами раздела *Event* | *HumanObservation* | *MachineObservation*. В разделе *Location* можно охарактеризовать географическую привязку места находки. Таксономическое положение вида (образца) описывается в разделе *Taxon*, а источники, использованные для его определения – в разделе *Identification*. Описание особенностей конкретного организма или таксономически однородной группы организмов приводится в разделе *Organism*. Для каждого раздела предусмотрены примечания, которые заносятся в поля *Remarks*. В них можно приводить любую дополнительную информацию, касающуюся соответствующего раздела.

Концепция DwC предполагает хранение как исходной, так и формализованной информации с описанием методов, которыми эта формализация была выполнена. Для хранения исходной информации используются термины группы *verbatim*, в буквальном переводе «дословно», (например, *verbatimEventDate*, *verbatimLocality* и др.); информация в них должна в точности соответствовать первоисточнику (гербарной этикетке, полевому дневнику, литературной публикации и т.п.). Термины, предназначенные для хранения формализованной информации, предполагают фиксированное значение из словаря или использование определенного стандарта представления данных.

В описании терминов, имеющих словарь, содержится фраза «*Recommended best practice is to use a controlled vocabulary*», список возможных значений доступен по ссылке в разделе *Comment*. Например, для термина *disposition*, описывающего актуальное состояние коллекционного образца, возможны значения «*in collection*», «*missing*», «*voucher else where*», «*duplicates else where*». Информация о форматах представления данных также приводится в описании терминов. Например, для описания даты и времени используется стандарт ISO 8601:2004(E). Некоторые термины DwC могут содержать перечисление нескольких значений, например, в *identificationReferences* может быть перечислено несколько источников, использованных для определения таксономического положения объекта. В таких случаях значения рекомендуется отделять друг от друга вертикальной чертой (|). Данные в полях *verbatim* следует приводить на языке оригинала, а формализованную информацию – желательно на английском языке.

Часто при подготовке данных к публикации возникают трудности при указании названий географических объектов. Проблема заключается в том, что в России отсутствует общепринятый перечень географических названий, а их написание в английской транслитерации в общедоступных картографических сервисах может различаться. В связи с этим, необходимо не только приводить англоязычные названия географических объектов, но и указывать источники, из которых они получены, а также приводить топонимы на языке оригинала. Эти правила касаются также трансформированных географических координат, унифицированных названий видов и др.

Важным этапом стандартизации данных является оценка их качества. Прежде всего, это касается определения точности географических координат места находки того или иного вида. В начале XXI в. исследователям стали широко доступны средства спутниковой навигации, позволяющие фиксировать координаты с точностью до нескольких метров. В течение большей части XX в. координаты, как правило, указывали по топографическим картам общедоступных масштабов 1:200 000 или 1:500 000, вследствие чего точность привязки, в лучшем случае, составляла сотни метров. Для большинства находок в качестве ориентира указывался ближайший географический объект или населенный пункт, а также административная принадлежность территории. При этом за последние 100 лет границы регионов неоднократно менялись, многие населенные пункты были переименованы (некоторые неоднократно) или в настоящее время уже не существуют. Очевидно, что в такой ситуации оценка точности географической привязки находок очень важна, и от нее зависят возможности использования данных для решения тех или иных задач.

Для оценки качества геоданных рекомендуется сначала загрузить координаты точек находок в какую-либо геоинформационную систему (ГИС) и проверить данные на отсутствие грубых ошибок в географической привязке (например, точки, находящиеся в море для сухопутных видов, ошибки привязки из-за замены местами значений широты и долготы, точки с нулевыми координатами и др.). После этого желательно также оценить точность геопривязки, если это не было сделано ранее в процессе сбора или оцифровки данных. Термины DwC позволяют указать точность определения координат с помощью GPS-навигатора в долях градуса (термин *coordinatePrecision*) или в метрах при определении места находки на основе вербального описания (термин *coordinateUncertaintyInMeters*). В последнем случае рекомендуется

следовать стандартной методике («точка-радиус»), описанной в соответствующем руководстве (Guide…, 2006). Оценка точности привязки может быть выполнена при помощи онлайн-калькулятора (Wieczorek, Wieczorek, 2015). В случае оценки точности привязки другими методами необходимо приводить данные об источниках и описание этих методов в соответствующих полях DwC. Если по какой-то причине точные координаты находок заведомо не приводятся или точность привязки преднамеренно снижается, например, в случаях публикации данных о распространении охраняемых видов, то это также необходимо указывать (термин *informationWithheld*).

Важно также обратить внимание на технические требования к качеству данных. Не должно быть дубликатных записей (строк) и повторяющихся полей (столбцов). В текстовых данных не допускаются грамматические ошибки, пробелы в начале и конце строки, пробельные символы, не являющиеся пробелом (неразрывный пробел, перевод каретки, разрыв строки, знак табуляции и т. п.). Также рекомендуется изучить отчет об ошибках (*interpretation issues*), доступный на странице набора данных, и загрузить уже опубликованный набор в формате DwC-A, чтобы оценить, как он выглядит с точки зрения их конечного пользователя.

Публикация данных через IPT

Непосредственно процедура публикации данных через GBIF.org осуществляется с помощью специального программного обеспечения IPT, написанного на языке Java и функционирующего как серверное приложение с визуальным интерфейсом, доступным через браузер. IPT работает под управлением веб-сервера Apache и службы веб-приложений TomCat 7. Изначально IPT имеет англоязычный интерфейс, начиная с версии 2.3.3, вышедшей в начале 2017 г., доступна русскоязычная версия. Все опубликованные данные хранятся непосредственно на сервере с установленным IPT, на глобальном портале размещаются только метаданные. IPT может быть связан с веб-сайтом национального портала, на котором визуализируются опубликованные через эту инсталляцию данные. К одной IPT-инсталляции может быть «привязано» несколько организаций, каждая из которых, в свою очередь, может иметь несколько аккаунтов для сотрудников с разными правами в отношении публикации данных. Для того, чтобы опубликовать подготовленный набор данных нужно установить собственный IPT или воспользоваться уже существующей инсталляцией, связавшись с ее администратором. В настоящее время в России функционируют 4 IPT-инсталляции: в Зоологическом институте РАН (г. Санкт-Петербург, администратор Халиков Р. Г.), Институте математических проблем биологии — филиале ИПМ им. М. В. Келдыша РАН (г. Пущино, администратор Шашков М. П.), Институте биологии Коми научного центра УрО РАН (г. Сыктывкар, администратор Чадин И. Ф.) и в Институте растениеводства им. Н. И. Вавилова РАН (г. Санкт-Петербург, администратор Лоскутов И. Г.).

Для публикации необходимо загрузить в IPT электронную таблицу с данными, проверить их на соответствие DwC и разместить метаданные. После окончания технической части публикации (нажатия кнопки *Register*) набор данных индексируется в глобальной системе и в течение нескольких часов становится доступным через портал GBIF.org.

Подготовка и публикация статьи о данных

Как было описано выше, статья о данных представляет собой подробное описание массива, опубликованного через какой-либо тематический электронный портал. Как и в случае «обычной» исследовательской статьи, структура рукописи должна соответствовать определенным критериям, которые описаны в правилах для авторов. Рукописи проходят обязательное рецензирование. Важно учитывать, что издатели статей о данных оставляют за собой право принимать к публикации рукописи, содержащие описание только «значимых для науки» данных, в то время как открытые порталы (GBIF.org и др.) не накладывают никаких ограничений на объем, географический и временной охват публикуемой информации. В процессе публикации набора данных важно следить, чтобы все необходимые поля метаданных были заполнены, а текст метаданных соответствовал общепринятому стилю изложения научной информации на английском языке.

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. В первую очередь, необходимо обратить особое внимание на лицензию, которую автор может присвоить своему набору данных. Pensoft Publishers, в отличие от GBIF, не позволяет использовать лицензии, ограничивающие коммерческое использование данных (CC-BY-NC). Подготовка рукописи осуществляется автоматизировано, с помощью средств IPT. На основе введенных метаданных формируется черновик (текстовый файл в формате RTF), соответствующий правилам оформления статей. Он становится доступным для загрузки в используемой IPT-инсталляции на странице набора данных после его публикации. При необходимости этот файл можно дополнить иллюстрациями и ссылками на литературные источники.

Как правило, в процессе подготовки рукописи, требуется несколько раз вносить уточнения, как в исходные, так и в метаданные, что, в свою очередь, требует обновления уже опубликованного набора данных. Благодаря функциональным возможностям IPT, обновленные версии публикуются с сохранением истории изменений.

Заключение

Представленные материалы разработаны на основе трехлетнего опыта проведения авторами тематических семинаров и деятельности неформального сообщества GBIF в России. Приведенные рекомендации никоим образом не являются исчерпывающим руководством в силу специфичности биологических данных и постоянного совершенствования стандартов их описания. Они представляют основные принципы публикации данных и ключевые этапы их стандартизации. Для уточнения деталей при работе с каждым конкретным набором данных необходимо обращаться к источникам, приведенным в списке литературы.

Объединение данных на основе международных стандартов способствует повышению их качества и стимулирует развитие методов их интеграции, хранения, визуализации и анализа. Публикация данных через GBIF.org будет способствовать вовлечению российских исследователей в международные проекты и увеличению числа публикаций в высокорейтинговых изданиях.

Благодарности

Авторы благодарят участников неформального сообщества GBIF в России и участников семинаров, посвященных публикации данных через GBIF.org, за активное участие и вопросы, которые послужили основой этой публикации. Также авторы благодарны рецензентам, замечания и предложения которых способствовали улучшению рукописи.

Литература

Айтинские целевые задачи. 2010. URL: https://www.cbd.int/sp/targets/ (28.08.2017).

Веб-ГИС «Фаунистика». 2012. URL: http://rrrcn.ru/ru/birdwatching/web-gis (28.08.2017).

Иванова Н. В., Шашков М. П. Перспективы создания открытого всероссийского информационного ресурса по биоразнообразию на основе международного стандарта GBIF // Математическая биология и биоинформатика, 2014. Т. 9. Вып. 2. С. 396–405. doi: 10.17537/2014.9.396.

Иванова Н. В., Шашков М. П. Спецификация Darwin Core. 2017. URL: http://gbif.ru/DwC_spec (28.08.2017).

Конвенция о биологическом разнообразии. 1992. URL: https://www.cbd.int/doc/legal/cbd-ru.pdf (28.08.2017).

Мелехин А. В., Давыдов Д. А., Шалыгин С. С., Боровичев Е. А. Общедоступная информационная система по Биоразнообразию цианопрокариот и лишайников CRIS (Cryptogamic Russian Information System) // Бюллетень Московского общества испытателей природы. Отдел биологический, 2013. Т. 118. Вып. 6. С. 51–56.

О лицензиях Creative Commons. 2017. Для чего созданы наши лицензии. URL: https://creativecommons.org/licenses/ (28.08.2017).

Access to Biological Collections Data — ABCD. 2015. URL: http://www.tdwg.org/activities/abcd/ (28.08.2017).

Become a publisher. 2017. URL: https://www.gbif.org/become-a-publisher (28.08.2017).

Catalogue of Life. 2015. URL: http://www.catalogueoflife.org/ (28.08.2017).

Chadin I., Dalke I., Zakhozhiy I., Malyshev R., Madi E., Kuzivanova O., Kirillov D., Elsakov V. Occurrences of the invasive plant species Heracleum sosnowskyi Manden. in the Komi Republic territory (European North-East Russia) // PhytoKeys, 2017. Vol. 77. P. 71–80. https://doi.org/10.3897/phytokeys.77.11186.

Chadin I., Dalke I., Zakhozhiy I., Malyshev R., Madi E., Kuzivanova O., Kirillov D. Occurrences of the invasive plant species *Heracleum sosnowskyi* Manden. in the Komi Republic (Russia). Version 1.8. Institute of Biology of Komi Scientific Centre of the Ural Branch, Russian Academy of Sciences. Dataset / Occurrence. 2016. https://doi.org/10.15468/zo2svq (28.08.2017).

Checklist Data. 2017. URL: https://github.com/gbif/ipt/wiki/checklistData (28.08.2017).

Chavan V., Penev L. The data paper: a mechanism to incentivize data publishing in biodiversity science // BMC Bioinformatics, 2011. Vol. 12(15): S2. doi: 10.1186/1471-2105-12-S15-S2.

Costello M. J. Motivating Online Publication of Data // BioScience, 2009. Vol 59. N. 5. P. 418–427. doi:10.1525/bio.2009.59.5.9.

Create your own Darwin Core Archive, How-To Guide (contributed by Remsen D.P., Braak K., Döring M., Robertson T.). Copenhagen: Global Biodiversity Information Facility, 2011. 20 pp. URL: http://www.gbif.jp/v2/pdf/gbif_dwc-a_how_to_guide_en_v1.pdf (28.08.2017).

Darwin Core Terms: A quick reference guide. 2015. URL: http://rs.tdwg.org/dwc/terms/ (28.08.2017).

Data papers. 2017. URL: https://www.gbif.org/data-papers (28.08.2017).

Dataset classes. 2017. URL: https://www.gbif.org/dataset-classes (28.08.2017).

Ecological Metadata Language (EML). 2017. URL: https://knb.ecoinformatics.org/#external//emlparser/docs/index.html (28.08.2017).

GBIF Backbone Taxonomy. GBIF Secretariat. Checklist Dataset. 2016. URL: https://doi.org/10.15468/39omei(28.08.2017).

Guide to Best Practices for Georeferencing / eds. Chapman A.D., Wieczorek J. Global Biodiversity Information Facility: Copenhagen, 2006. 90 pp. URL: http://herpnet.org/herpnet/documents/biogeomancerguide.pdf (28.08.2017).

Khanina L., Zaugolnova L., Smirnova O., Shovkun M., Glukhova E. Dataset «Flora of vascular plants in the Central European Russia» Institute of Mathematical Problems of Biology, Russian Academy of Sciences. Dataset Species checklist, 2016. doi: 10.15468/96gqtn.

Melechin A. Specimens of lichen herbarium KPABG. Version 1.3. Polar-Alpine Botanical Garden-Institute of N. A. Avrorin KSC RAS. Occurrence Dataset. 2016. URL: https://doi.org/10.15468/nctfm2 (28.08.2017).

Occurrence Data. 2017. URL: https://github.com/gbif/ipt/wiki/occurrenceData. (28.08.2017).

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

Petrosyan V., Kuzmin S. Amphibians of the Former USSR. Version 1.11. A. N. Severtsov Institute of Ecology and Evolution, Russian academy of sciences. Occurrence Dataset, 2017. URL: https://doi.org/10.15468/wxz3yj (28.08.2017).

Resource Metadata. 2017. URL: https://github.com/gbif/ipt/wiki/resourceMetadata (28.08.2017).

Robertson T., Döring M., Guralnick R., Bloom D., Wieczorek J., Braak K., Otegui O., Russell L., Desmet P. The GBIF Integrated Publishing Toolkit: Facilitating the Efficient Publishing of Biodiversity Data on the Internet // PLoS ONE, 2014. Vol. 9(8): e102623. https://doi.org/10.1371/journal.pone.0102623.

Sampling Event Data. 2017. URL: https://github.com/gbif/ipt/wiki/samplingEventData (28.08.2017).

Seregin A. P. Making the Russian flora visible: Fast digitization of the Moscow University Herbarium (MW) in 2015 // Taxon, 2016. Vol. 65. № 1. P. 205–207. http://dx.doi.org/10.12705/651.29.

Seyfulina R., Buyvolov Y. List of Spiders of Prioksko-Terrasnyi Biosphere Reserve. Prioksko-Terrasnyi Biosphere Reserve. Occurrence Dataset. 2016. URL: https://doi.org/10.15468/3cbyt7 (28.08.2017).

Shashkov M., Ivanova N. Database of finds of rare lichen species *Lobaria pulmonaria* in Russia. Version 1.4. Institute of Mathematical Problems of Biology, Russian Academy of Sciences. Occurrence Dataset, 2016. URL: https://doi.org/10.15468/uennht (28.08.2017).

Smirnov R., Golikov A., Khalikov R. Catalogue of the type specimens of Pogonophora (Annelida; seu Polychaeta: Siboglinidae) from research collections of the Zoological Institute, Russian Academy of Sciences. Version 1.1. Zoological Institute, Russian Academy of Sciences, St. Petersburg. Dataset/Checklist. 2017. URL: https://doi.org/10.15468/1mlkdp (28.08.2017).

TDWG. Biodiversity Information Standards. 2017. URL: http://www.tdwg.org/ (28.08.2017).

Volkovitsh M., Golikov A., Khalikov R. Catalogue of the type specimens of Polycestinae (Coleoptera: Buprestidae) from research collections of the Zoological Institute, Russian Academy of Sciences. Version 1.10. Zoological Institute, Russian Academy of Sciences, St. Petersburg. Checklist Dataset. 2017. URL: https://doi.org/10.15468/c3eork (28.08.2017).

Wieczorek J., Bloom D., Guralnick R., Blum S., Dorring M., Gilovanni R., Robertson T., Vieglais D. Darwin core: An Evolving Community-Developed Biodiversity Data Standard // PLoS ONE, 2012. Vol. 7. №. 1. P. 1–8. doi.org/10.1371/journal.pone.0029715.

Wieczorek C., Wieczorek J. Georeferencing Calculator (version 20160929). Museum of Vertebrate Zoology, University of California, Berkeley. 2015. URL: http://manisnet.org/gci2.html (28.08.2017).

Сведения об авторах

Шашков Максим Петрович,
Магистр биологии, старший научный сотрудник Института математических проблем биологии РАН – филиала Федерального государственного учреждения «Федеральный исследовательский центр Институт прикладной математики им. М. В. Келдыша РАН», научный сотрудник Федерального государственного бюджетного учреждения науки института физико-химических и биологических проблем почвоведения РАН, Пущино; Max.carabus@gmail.com
Чадин Иван Федорович,
Кандидат биологических наук, заместитель директора по научной работе Федерального государственного бюджетного учреждения науки Института биологии Коми НЦ Уральского отделения РАН, Сыктывкар; chadin@ib.komisc.ru
Иванова Наталья Владимировна,
Магистр биологии, научный сотрудник Института математических проблем биологии РАН – филиала Федерального государственного учреждения «Федеральный исследовательский центр Институт прикладной математики им. М. В. Келдыша РАН», Пущино; Natalya.dryomys@gmail.com

Shashkov Maxim Petrovich,
Master of science (biology), Senior Researcher of Institute of Mathematical Problems of Biology RAS – the Branch of Keldysh Institute of Applied Mathematics of RAS, Researcher of Institute of Physicochemical and Biological Problems in Soil Science Russian Academy of Science, Pushchino; Max.carabus@gmail.com
Chadin Ivan Fedorovich
PhD, Deputy Director of Institute of Biology of Komi Scientific Centre of the Ural Branch of RAS, Syktyvkar; chadin@ib.komisc.ru
Ivanova Natalya Vladimirovna,
Master of science (biology), Researcher of Institute of Mathematical Problems of Biology RAS – the Branch of Keldysh Institute of Applied Mathematics of RAS, Pushchino; Natalya.dryomys@gmail.com