***Доклады Международной конференции "Математическая биология и биоинформатика"**. Под ред. В.Д. Лахно. Том 7. Пущино: ИМПБ РАН, 2018. Статья № e98. doi: 10.17537/icmbb18.55*

Стандарты и веб-инструменты для публикации данных через глобальные порталы по биоразнообразию

Шашков М.П.[1, 2], Иванова Н.В.[1]

[1]*ИМПБ РАН – филиал ИПМ им. М.В. Келдыша РАН*
[2]*Институт физико-химических и биологических проблем почвоведения РАН*

max.carabus@gmail.com

Разработка методик и инструментов для хранения, управления и обмена цифровыми данными о находках видов является актуальным направлением развития информатики биоразнообразия. Использование унифицированных стандартов обеспечивает совместимость данных и их эффективный поиск в тематических информационных системах. Крупнейшей из них на сегодняшний день является Глобальная информационная система по биоразнообразию GBIF.org, объединяющая более миллиарда записей о находках видов, происходящих из разных источников. В данной работе описаны стандарты и готовые веб-инструменты, которые могут быть использованы как для интеграции данных в глобальные тематические порталы, так и для менеджмента локальных массивов. Наиболее распространенным для хранения данных о находках видов является стандарт Darwin Core, для представления данных геоботанических описаний активно развивается стандарт Veg-X. Для публикации данных через глобальные и тематические порталы служит Java веб-приложение Integrated Publishing Toolkit (IPT). Программные решения портала Atlas of Living Australia (ALA) предоставляют открытый набор готовых модулей и инструментов для разработки тематических порталов о биоразнообразии.

Ключевые слова: *информатика биоразнообразия, GBIF, Darwin Core, Veg-X, IPT, ALA.*

Standards and tools for data publishing through international biodiversity portals

Shashkov M.P.[1, 2], Ivanova N.V.[1]

[1]*IMPB RAS – Branch of KIAM RAS*
[2]*Institute of Physicochemical and Biological Problems in Soil Science RAS*

The development of the methods and tools for storing, managing and sharing digital data on species occurrences is an essential direction of the biodiversity informatics. The application of unified standards provides the data compatibility and their effective search in thematic information systems. The Global Biodiversity Information Facility GBIF.org is the world's largest portal, which includes more than 1billion occurrences from many different sources. Standards and web tools for data integration into global thematic portals and for local datasets managing are described in this article. Darwin Core is the most common biodiversity data standard for occurrence data storing. The Veg-X standard was created for vegetation relevés data, and now it is under active developing. The Java web-application Integrated Publishing Toolkit (IPT) is used for data publishing through different biodiversity portals. Atlas of Living Australia (ALA) software solutions provide a set of open source modules and tools for the development of biodiversity portals.

Key words: biodiversity informatics, GBIF, Darwin Core, Veg-X, IPT, ALA.

1. Введение

Современные таксономические ревизии, филогенетические и биоклиматические модели, как правило, являются результатом анализа данных, полученных из множества источников. Для их объединения необходимы универсальные стандарты, позволяющие не только хранить информацию, но и быстро обновлять уже существующие массивы.

В данной работе мы описываем основные стандарты для представления данных по биоразнообразию и инструменты для их публикации через глобальные тематические порталы.

2. Глобальные данные и информационные системы о биоразнообразии

На мировом уровне разработка инструментов для хранения и управления цифровыми данными о

биоразнообразии, а также протоколов для обмена информацией, полученной из разных источников, является динамично развивающейся областью науки [1–4]. Направление получило название Biodiversity informatics или информатика биоразнообразия [5]. Ведущей в этой области является группа Biodiversity Informatics Standards (альтернативное название Taxonomic Databases Working Group, TDWG – http://www.tdwg.org). За более чем 30 лет деятельности ею были разработаны открытые стандарты для хранения, объединения и распространения данных о биоразнообразии, наиболее известными из которых являются Darwin Core (DwC) и ABCD, а также протокол обмена данными TAPIR [1]. На основе этих универсальных решений были созданы глобальные тематические порталы, интегрирующие данные о распространении живого на Земле. На сегодняшний день крупнейшей такой структурой является Глобальная информационная система по биоразнообразию GBIF.org [6, 7], объединяющая более миллиарда записей о находках видов. Кроме данных 1225 научно-исследовательских организаций и биологических коллекций, в GBIF представлены данные, объединяемые международными сообществами, такими как Сохранение флоры и фауны Арктики (Conservation of Arctic Flora and Fauna, CAFF), Европейский совет по учетам птиц (European Bird Census Council, EBCC), Международная сеть любительских наблюдений iNaturalist и др. Таким образом, GBIF является универсальной поисковой системой данных о биоразнообразии. Использование универсальных стандартов для представления данных и инструментов для их публикации позволяет авторам отслеживать использование данных в составе объединенных массивов и их цитирование в научных публикациях при помощи DOI. Все это, в конечном счете, повышает востребованность и эффективность использования этой информации.

Большинство отечественных исследователей используют для хранения данных собственные локальные системы и стандарты, появившиеся в процессе их разработки, международные стандарты для представления данных пока остаются мало известными и используются редко [8]. Между тем, с 2014 г. наблюдается устойчивый рост объема данных, опубликованных российскими исследователями через портал GBIF. На момент подготовки этой публикации в GBIF было зарегистрировано 29 российских организаций, которые опубликовали 44 набора данных, содержащих 1369450 стандартизированных записей о находках видов. Среди них крупнейшим массивом является оцифрованный гербарий Московского университета (MW) [9].

3. Стандарты для представления данных о биоразнообразии

3.1. Стандарты для представления данных об отдельных находках

Наиболее распространенным для хранения данных о находках видов является стандарт DwC [1]. Он представляет собой набор терминов (*terms*) и правил их использования. Полный перечень и подробное описание терминов доступны на сайте TDWG (http://rs.tdwg.org/dwc/terms/), краткое – в нашей предыдущей работе [10].

Первоначально DwC разрабатывался для хранения данных об образцах естественнонаучных коллекций, позднее он был расширен для описания литературных данных, полевых наблюдений и др. Текущая версия DwC содержит около 200 основных терминов для описания атрибутивных данных о находках, а также 30 расширений (*extensions*), т.е. наборов дополнительных терминов, для представления (формализации) специфической информации, например, характеристик отобранных проб, результатов молекулярно-генетического анализа и др. Для удобства пользователей термины DwC сгруппированы в несколько тематических разделов. Раздел *Record-level Terms* используется для описания общих данных о находках. В разделе *Occurrence* приводятся сведения о состоянии организма (ов) во время наблюдения или сбора (численность, поведение, жизненное состояние и др.). Дата наблюдений и методы отбора образцов описываются терминами раздела *Event.* В разделе *Location* можно подробно 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 вносятся в поля группы *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*; 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 определенной формы представления данных.

Стандарт DwC является основным в сети GBIF, а также используется в информационной системе о

распространении морских видов OBIS (http://www.iobis.org/), международном проекте, обобщающем находки позвоночных животных VertNet (http://vertnet.org/) [11], портале Encyclopedia of Life (http://eol.org/) и др.

Широко известные готовые программные продукты для управления данными естественнонаучных коллекций, такие как Symbiota (http://symbiota.org/docs) [12], BRAHMS (https://herbaria.plants.ox.ac.uk/bol), Specify (http://www.sustain.specifysoftware.org) и др. изначально хранят данные в DwC, либо имеют функцию экспорта в этот стандарт. Для хранения данных естественнонаучных коллекций существует также специализированный стандарт ABCD (Access to Biological Collections Data) [13]. Аналогично DwC он построен на отдельных находках видов, но предполагает более детальное описание образцов. ABCD используется в Международной сети биологических коллекций BioCASe (http://www.biocase.org/), GBIF и др.

3.2. Стандарты для представления описаний площадных учетов

Принципиальной особенностью данных, которые собираются на пробных площадях является то, что те или иные совокупности видов обнаружены в определенное время и в определенном месте. Представление такой информации должно быть основано не на описании отдельных находок, а на описании учетных площадей. В связи с этим использование DwC и подобных ему стандартов, для хранения данных площадных описаний затруднительно. Технически система GBIF позволяет публиковать данные площадных учетов, в DwC (т.н. *sampling event data*) [14], в которых описания характеристик исследуемых площадей представляются как события (*event Core*), а данные о видах и их характеристиках – как отдельные находки, связанные через ID с соответствующими площадями (*occurrence Core*). Разработана процедура экспорта данных в DwC из широко используемой для хранения геоботанических описаний базы TurboVeg [15]. В то же время наш собственный опыт публикации данных геоботанических описаний [16] показывает, что существующих терминов DwC недостаточно для представления стандартных характеристик ярусов растительности.

Рабочей группой по экологической информатике Международной ассоциации науки о растительности (International Association for Vegetation Science, IAVS) при поддержке TDWG был разработан стандарт Veg-X, предназначенный специально для хранения данных геоботанических описаний. Описание пробной площади в Veg-X включает следующие компоненты [17]:

- *Plots*. Базовые свойства пробной площади, которые не зависят от времени проведения исследований, например, место нахождения, положение в рельефе, геометрическая форма и др. Здесь приводятся сведения о локальной системе координат, в которой описано положение объектов;
- *Plot observations*. Выполненные на пробной площади измерения, результаты которых зависят от времени проведения исследований (например, учет обилия видов, изучение свойств почвы и т.д.). Такие измерения могут включать несколько групп записей: флористические, климатические, эдафические и др., они группируются по времени или участку пробной площади, на котором проведены исследования;
- *Organism observations*. Наблюдения за организмами (как живыми, так и мертвыми) на пробной площади. Это могут быть характеристики как отдельных особей (например, высота или диаметр конкретного дерева), так и групп особей (биомасса, покрытие и др.);
- *Organism and community identifications*. Информация о таксономии, использованной для названий видов и классификации для названий сообществ;
- *Strata and stratum observations*. Описание характеристик групп организмов, выделенных по определенному протоколу. Например, это могут быть характеристики ярусов растительности;
- *Projects*. Информация об исследовательском проекте и месте данной пробной площади в нем.

С 2018 г. разрабатывается пакет для среды R, осуществляющий экспорт данных в Veg-X [18], что является удобным инструментом для хранения локальных тематических массивов данных.

По мнению разработчиков, использование Veg-X также облегчает публикацию данных через универсальные репозитории, такие как Figshare (https://figshare.com/) и DRYAD (https://datadryad.org/), а также интеграцию этих сведений в глобальные тематические системы. Между тем, большинство массивов геоботанических описаний, включая крупнейшие международные базы данных о растительности GIVD (http://www.givd.info/) и sPlot (https://www.idiv.de/sdiv/working_groups/wg_pool/splot.html) пока остаются закрытыми, поэтому на данный момент невозможно оценить широту применения Veg-X для хранения геоботанических данных.

4. Веб-инструменты для представления данных в сети Интернет

4.1. Программные решения для поддержки публикации данных через глобальные порталы о биоразнообразии

Среди готовых инструментов, осуществляющих электронную публикацию данных, наиболее популярным является GBIF Integrated Publishing Toolkit (IPT) [2]. Это ПО с открытым исходным кодом, написанное на языке Java, разработано для

публикации наборов данных через глобальные порталы по биоразнообразию. Основные функции IPT – проверка данных на соответствие DwC, их последующая архивация в DwC Archive (набор файлов, соответствующий стандартному формату [19]) и публикация через глобальный портал. Для работы с IPT используется визуальный веб-интерфейс, доступный через браузер. Серверная часть приложения работает под управлением службы TomCat, обычно в сочетании с веб-сервером Apache.

Публикацию через IPT осуществляют тематические порталы GBIF, OBIS (http://www.iobis.org/), VertNet (http://vertnet.org/) и Global Genome Biodiversity Network GGBN (http://www.ggbn.org/). Важно отметить, что публикация данных через IPT лишь делает их обнаружимыми для поисковых запросов, в то время как сами данные остаются под полным контролем публикующей организации и хранятся на определенном сервере с IPT, они могут быть произвольно обновлены, сняты с публикации или удалены. К одной инсталляции IPT может быть «привязано» несколько организаций, каждая из которых может, в свою очередь, иметь несколько аккаунтов для сотрудников с разными правами в отношении публикации данных.

На сегодняшний день в мире работает 229 GBIF IPT-инсталляций в 69 странах, 5 из которых находятся в России. IPT-инсталляция поддерживаемая ИМПБ РАН – филиал ИПМ им. М.В. Келдыша предоставляет хостинг для 22 российских научно-исследовательских организаций и двух – из Республики Узбекистан. Фактически эта структура выполняет функции национального узла сети GBIF.

4.2. Инструментарий для создания тематических порталов ALA

Бесплатные инструменты для создания тематических порталов по биоразнообразию предоставляет открытая программная инфраструктура, разработанная коллективом проекта Atlas of Living Australia, ALA (https://demo.gbif.org/programme/living-atlases, https://www.ala.org.au/) [20]. Это интегрированное веб-приложение, состоящее из модульного набора инструментов и компонентов, соединенных вместе через архитектуру микросервисов, каждый из которых имеет определенную роль. Набор компонентов ALA дает возможность создавать многопользовательские ресурсы, включающие интерактивные карты находок видов с имеющейся атрибутивной информацией, справочные материалы о биологии и филогении отдельных видов и галереи их изображений, а также описания фондов естественнонаучных коллекций. Модульная архитектура ALA позволяет в короткие сроки запустить работоспособный ресурс, обладающий минимальным набором компонентов и, впоследствии, по мере необходимости, добавлять новые функции [21, 22]. Использование международных стандартов (Darwin Core, DwC-Archive, ISO и др.) обеспечивает совместимость данных с другими глобальными системами.

Картографический сервис ALA имеет широкие возможности для визуализации и пространственного анализа данных о находках видов. Система поддерживает работу не только со встроенными слоями, но и позволяет пользователям загружать собственные данные в формате CSV. Имеются встроенные инструменты, позволяющие проводить широко распространенные виды пространственного анализа, также возможен экспорт данных для их последующей обработки в других программных продуктах.

С 2014 г. при поддержке Секретариата GBIF ведется работа по распространению инструментов ALA для создания национальных порталов о биоразнообразии в странах-участницах GBIF. На сегодняшний день системы, основанные на ALA инструментарии, уже работают в Испании, Португалии, Швеции, Франции, Андорре, Эстонии, Шотландии, Уэльсе и Великобритании, Канаде, Бразилии, Аргентине, Коста-Рике, Острове Мен; ведется разработка информационной системы о биоразнообразии Бенина [23]. В России работы по созданию национального портала на основе ALA инструментария осуществляются на базе ИМПБ РАН – филиал ИПМ им. М.В. Келдыша.

Компоненты ALA и их спецификация и доступны через хостинг IT-проектов github [24], с 2013 года участники проекта ежегодно проводят практические обучающие семинары.

5. Заключение

Использование унифицированных стандартов и инструментов для менеджмента локальных массивов данных существенно сокращает трудозатраты по их представлению в табличном виде и позволяет проводить совместный анализ сведений, собранных в разное время или на разных территориях. Электронная публикация данных через глобальные порталы по биоразнообразию обеспечивает обнаружимость информации, повышая эффективность ее использования в составе объединенных массивов. Благодаря готовым инструментам, осуществляющим публикацию и индексацию данных, авторы могут обновлять уже опубликованные массивы и отслеживать их цитирование в научных публикациях.

6. Список литературы

1. Wieczorek J., Bloom D., Guralnick R., Blum S., Döring M., Giovanni R., Robertson T., Vieglais D. Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. *PLoS ONE*. 2012. V. 7. № 1. P. e29715. doi: 10.1371/journal.pone.0029715.

2. Robertson T., Döring M., Guralnick R., Bloom D., Wieczorek J., Braak K., Otegui J., Russell L., Desmet P. The GBIF Integrated Publishing Toolkit: Facilitating the Efficient Publishing of Biodiversity Data on the Internet. *PLoS ONE*. 2014. V. 9. № 8. P. e102623. doi: 10.1371/journal.pone.0102623 .
3. 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. V. 3. P. e12431. doi: 10.3897/rio.3.e12431.
4. Pennisi E. Boom in digital collections makes a muddle of management. *Science*. 2005. V. 308. P. 187–189.
5. Bisby F.A. The quiet revolution: Biodiversity informatics and the Internet. *Science*. 2000. V. 289. P. 2309–2312.
6. Wheeler Q.D. What if GBIF? *Bioscience*. 2004. V. 54. P. 717.
7. Yesson C., Brewer P.W., Sutton T., Caithness N., Pahwa J.S., Burgess M., Gray W.A., White R.J., Jones A.C., Bisby F.A., Culham A. How global is the Global Biodiversity Information Facility? *Plos ONE*. 2007. V. 2. № 11. P. e1124. doi: 10.1371/journal.pone.0001124.
8. Филиппова Н.В., Филиппов И.В., Щигель Д.С., Иванова Н.В., Шашков М.П. Информатика биоразнообразия: мировые тенденции, состояние дел в России и развитие направления в Ханты-Мансийском Автономном Округе. *Динамика окружающей среды и глобальные изменения климата*. 2017. Т. 8. №. 2. С. 46–56. doi: 10.17816/edgcc8246-56.
9. Seregin A. *Moscow University Herbarium (MW). Version 1.37. Lomonosov Moscow State University. Occurrence dataset*. 2018. URL: https://doi.org/10.15468/cpnhcc (дата обращения: 18.07.2018).
10. Шашков М.П., Чадин И.Ф., Иванова Н.В. Методические рекомендации по стандартизации данных для публикации через глобальный портал GBIF.org и подготовке статьи о данных. *Труды Кольского научного центра РАН. Серия Прикладная экология Севера*. 2017. Т. 5. № 3. С. 22–35.
11. Constable H., Guralnick R., Wieczorek J., Spencer C., Peterson A.T. VertNet: A New Model for Biodiversity Data Sharing. *Plos ONE*. 2010. V. 8. № 2. P. e1000309. doi: 10.1371/journal.pbio.1000309 .
12. Gries C., Gilbert E., Franz N. Symbiota – A virtual platform for creating voucher-based biodiversity information communities. *Biodiversity Data Journal*. 2014. V. 2. P. E1114. doi: 10.3897/BDJ.2.e1114.
13. *Access to Biological Collection Data* (ABCD). URL: https://www.tdwg.org/standards/abcd/ (дата обращения: 18.07.2018).
14. *Best Practices in Publishing Sampling-event data*: Version 2.0. URL: https://github.com/gbif/ipt/wiki/BestPracticesSamplingEventData#sampling-event-data (дата обращения: 18.07.2018).
15. Hennekens S. *Turboveg for Windows*: Version 2. 2017. URL: https://www.synbiosys.alterra.nl/turboveg/tvwin.pdf (дата обращения: 18.07.2018).
16. Ivanova N., Shanin V., Grozovskaya I., Khanina L. *Forest vegetation of the northeastern part of the Kostroma region (European Russia)*: Sampling event dataset. (Version 1.1.) Institute of Mathematical Problems of Biology, Russian Academy of Sciences, 2018. doi: 10.15468/qemuyc.
17. Wiser S., De Cáceres M., Kleikamp M., Boyle B., Peet R.K. Veg-X – an exchange standard for plot-based vegetation data. *Journal of Vegetation Science*. 2011. V. 22. P. 598–609. doi: 10.1111/j.1654–1103.2010.01245.x.
18. *The Veg-X exchange standard. IAVS Ecoinformatics Working Group*. URL: https://miquelcaceres.github.io/VegX/articles/VegXStandard.html#data-structure-of-veg-x-ver--2-0 (дата обращения: 18.07.2018).
19. *Darwin Core Archive. How-To Guide*. (Version 1.0). Copenhagen: Global Biodiversity Information Facility, 2011. 23 p. URL: http://www.gbif.jp/v2/pdf/gbif_dwc-a_how_to_guide_en_v1.pdf (дата обращения: 18.07.2018).
20. Belbin L., Williams K. Towards a national bioenvironmental data facility: experiences from the Atlas of Living Australia. *International Journal of Geographical Information Science*. 2016. V. 30. № 1. P. 108–125. doi: 10.1080/13658816.2015.1077962.
21. *Atlas of Living Australia*: ALA Infrastructure Implementation. 2016. URL: https://www.ala.org.au/wp-content/uploads/2017/01/ALA-Infrastructure-Implementation-overview-October-2016-final.pdf (дата обращения: 18.07.2018).
22. *Atlas of Living Australia South Australian User Support:* DEWNR Information Sheet. 2016. URL: https://data.environment.sa.gov.au/Content/Publications/ALA-DEWNRUserSupport.pdf (дата обращения: 18.07.2018).
23. *Living Atlases. Participants*. URL: https://living-atlases.gbif.org/ (дата обращения: 18.07.2018).
24. *Atlas of Living Australia*. URL: https://github.com/AtlasOfLivingAustralia (дата обращения: 18.07.2018).