Доклады Международной конференции "Математическая биология и биоинформатика".
Под ред. В.Д. Лахно. Том 10. Пущино: ИМПБ РАН, 2024. Статья № e55. doi: 10.17537/icmbb24.53

Опыт реализации региональных проектов оцифровки литературного наследия

Иванова Н.В.[1], Шашков М.П.[1], Созонтов А.Н.[2], Филиппова Н.В.[3], Ермолов С.А.[4], Соколова С.С.[2,5], Устинова А.Л.[6], Плакхина Е.В.[6]

[1]*ИМПБ РАН – филиал ИПМ им. М.В. Келдыша РАН*
[2]*Институт экологии растений и животных УрО РАН*
[3]*Югорский государственный университет*
[4]*Центр по проблемам экологии и продуктивности лесов РАН*
[5]*ЮУ ФНЦ МиГ УрО РАН*
[6]*Пермский национальный исследовательский университет*

Natalya.dryomys@gmail.com

Научное наследие, накопленное поколениями отечественных исследователей до сих пор остается разрозненным и в основном сосредоточено в серой литературе. Оцифровка этих материалов и открытое размещение их в в мировых репозиториях существенно повысит доступность первичных данных о биоразнообразии, а, следовательно, и качество модельных прогнозов динамики биоразнообразия в условиях глобальных климатических изменений. В работе проанализирован опыт трех проектов, посвященных оцифровке научной литературы по разным группам организмов. Определены этапы оцифровки, сформулированы основные проблемы, возникающие в процессе работы, описаны пути их решения. Показано, что подготовка к оцифровке (сбор полнотекстовых публикаций) и географическая привязка оцифрованных данных являются трудозатратными этапами, которые необходимо учитывать при планировании работ. Описаны возможности привлечения волонтеров к проектам по оцифровке. Обоснованы преимущества использования международного стандарта для обмена данными Darwin Core для оцифровки литературных источников.

Ключевые слова: *GBIF, моделирование биоразнообразия, грибы, пауки, дождевые черви*.

Experience of regional projects for digitisation of literature heritage

Ivanova N.V.[1], Shashkov M.P.[1], Sozontov A.N.[2], Filippova N.V.[3], Ermolov S.A.[4], Sokolova S.S.[2,4], Ustinova A.L.[5], Plakkhina E.V.[5]

[1]*Institute of Mathematical Problems of Biology of RAS, M.V. Keldysh Institute of Applied Mathematics*
[2]*Institute of Plant and Animal Ecology UB RAS*
[3]*Югорский государственный университет*
[4]*Center for Forest Ecology and Productivity*
[5]*SU FRC MG UB RAS*
[6]*Perm State University*

The scientific heritage accumulated by generations of national researchers is still scattered and mainly available from grey literature. Digitisation and publishing of these materials through open digital repositories could significantly increase the availability of primary biodiversity data and, consequently, improve the quality of species distribution models under global climate change. Our paper analyses the experience of three literature digitisation projects on different organism groups. The stages of digitisation are defined, the main problems arising of data digitisation process are specified, and the ways of solving them are described. It is shown that pre-digitisation activities (collection of full-text publications) and georeferencing of digitised occurrences are labour-consuming stages that should be taken into account when planning the project. The possibilities of involving volunteers in data digitisation projects are described. The advantages of using Darwin Core international standard for literature data digitisation are substantiated.

Key words: GBIF, biodiversity modelling, fungi, spiders, earthworms.

1. Введение

Экологическое моделирование ареалов, Species Distribution Modelling (SDM), – направление, позволяющее реконструировать потенциальные ареалы целевых видов, количественно оценивать взаимосвязь их распространения с условиями среды, а также прогнозировать динамику ареалов при разных сценариях изменений климата. SDM получило бурное развитие в мире благодаря появлению новых методов и алгоритмов построения моделей, реализованных в виде бесплатных программных продуктов [1, 2], а также доступности массивов исходных данных [3–5]. В то же время, несмотря на разнообразие ресурсов, предоставляющих сведения о характеристиках среды, используемых в моделировании ареалов в качестве предикторов, информаиция о точках встреч видов до сих пор сильно ограничена для многих территорий и таксонов. Построение моделей на основе таких неполных данных может привести к смещенным оценкам потенциальных ареалов и их динамики.

Общепринятым источником данных о точках встреч видов является Глобальный портал о биоразнообразии GBIF (Global Biodiversity Information Facility). На сентябрь 2024 г. через GBIF для территории России доступно 13.8 млн. записей о точках находок видов, имеющих географические координаты. Тем не менее, полнота и представленность данных для разных регионов нашей страны существенно различается [6]. Кроме того, более половины (56 %) этих записей происходят из систем для сбора любительских наблюдений (iNaturalist, RU-Birds, eBird). Эти данные имеют ряд специфических особенностей, таких как ненадежность определения, тяготение наблюдений к населенным пунктам и дорожной сети и др., поэтому требуют дополнительной проверки и обработки [7]. Кроме того, как правило, они отражают сведения о современных встречах видов и не позволяют реконструировать их распространение в прошлом.

В то же время, советскими и российскими исследователями в литературе накоплен колоссальный объем данных о распространении биологических видов (в виде аннотированных списков, монографических обработок, публикаций новых и интересных находок и др.). Оцифровка этих источников и их публикация через портал GBIF может существенно расширить объем доступных данных, что в свою очередь, позволит более точно прогнозировать реакции биоты на глобальные изменения климата.

В отсутствии национальной программы по оцифровке научного наследия в настоящее время эта задача решается отдельными коллективами для некоторых таксономических групп. К настоящему времени из литературных источников оцифровано около 1 млн. записей [8–10]. В данной работе мы обобщаем накопленный нашим коллективом опыт в виде формализации задач, решаемых в ходе оцифровки, возникающих при этом сложностей и описания путей их решения. Основой обобщения стали три проекта, посвященные оцифровке литературы по разным таксономическим группам и географическим регионам: 1) грибы Западной Сибири (завершенный) [11, 12]; 2) дождевые черви Северной Евразии (завершенный) [13]; 3) пауки Урала (продолжающийся) [14].

2. Этапы реализации проекта по оцифровке литературы

В результате анализа выполненных видов работ были выделены следующие этапы реализации проекта по оцифровке литературы:

1. Подготовка к оцифровке. Этот этап включает в себя составление требующих оцифровки источников в виде библиографического списка или базы данных, а также поиска полнотекстовых версий публикаций.

2. Непосредственно оцифровка. Перенесение информации согласно определенному стандарту из литературы в электронные таблицы или базы данных.

3. Обработка данных после оцифровки. На этом этапе решаются задачи верификации таксономических названий, ручной географической привязки мест сборов, проверки полученного массива данных и исправление технических ошибок (data cleaning).

4. Публикация оцифрованных данных онлайн в тематическом репозитории.

5. Публикация статьи о данных [15] в рецензируемом журнале (не рассматривается в этой работе).

Перечисленные этапы могут быть реализованы последовательно или параллельно в зависимости от числа сотрудников и используемых технических решений. Каждый участник может быть вовлечен как во все, так и только в некоторые этапы оцифровки. Перед началом работы важно определить “глубину” оцифровки, т.е. подробность, с которой будут представлены данные относительно первоисточника.

Особенности реализации рассматриваемых нами проектов, согласно выделенным этапам, описаны в таблице 1.

Обобщение литературных источников – это первая задача, которую необходимо решить на этапе подготовки к оцифровке. Далеко не по всем таксономическим группам существуют готовые библиографические сводки. Например, среди рассматриваемых нами проектов полная актуальная библиография имелась только для пауков [16, 17]. В двух других проектах потребовалось обобщение имеющихся источников и составление библиографических списков. Кроме того, не всегда доступны полнотекстовые версии публикаций.

Особенно это актуально для статей советского периода. Многие публикации доступны только на бумажных носителях и до настоящего времени сохранились в единичных экземплярах. Вероятно, некоторые публикации уже утрачены. Поиск полнотоекстовых версий в библиотеках или по запросу к авторам может потребовать значительных временных ресурсов, что важно учитывать при планировании работ по оцифровке.

Таблица 1. Методы реализации и результаты рассматриваемых проектов по оцифровке литературы

Этапы реализации и показатели	Проект по оцифровке литературы		
	Грибы	Дождевые черви	Пауки
	Подготовка к оцифровке		
Библиография	требовалось расширить предварительный список	требовалось составить	готовый
Хранение библиографических данных	Zotero	JabRef	собственное веб-приложение
Хранение публикаций	аккаунт Zotero или облачное хранение	локальное файловое хранилище	собственное веб-приложение
	Оцифровка		
Инструмент введения данных	Microsoft Excel, Google Sheets	Google Sheets	собственное веб-приложение
Используемый стандарт	Darwin Core	Darwin Core	Darwin Core
Кем ведется оцифровка	специалисты, волонтеры	специалисты	волонтеры
	Обработка данных после оцифровки		
Соответствие таксономическому справочнику GBIF	соответствует	соответствует частично	соответствует
Процент записей с ручной геопривязкой	75	68	>85 (оценочно)
	Результаты оцифровки		
Число оцифрованных источников	250	159	20 (на 09.2024) запланировано 400
Число оцифрованных записей	35 000	5304	600 (на 09.2024) запланировано 80 000
Доступность результатов	GBIF, статьи о данных [11, 12]	GBIF, статья о данных (на рецензии)	GBIF, статьи о данных (планируются)

Для оптимизации поиска и хранения (локально или онлайн) собранные публикации полезно организовать в виде библиографической базы. Для этого существует множество готовых технических решений (например, использованные нами JabRef и Zotero). Также рекомендуется составить базу метаданных собранных публикаций, которая будет влючать сведения о предварительной оценке числа записей в источнике, наличии географических координат мест сборов, изображений и другую информацию. База метаданных полезна для установления приоритетов в последовательности оцифровки, а также для выбора публикаций для оработки участниками проекта (см. ниже).

При оцифровке вне зависимости от таксономической группы из пуликаций извлекается и формализуется следующая обязательная информация: научное название биологического вида, дата сбора (наблюдения), сведения о месте проведения исследований и ученых, собравших и (или) определивших экземпляр(ы). При наличии рекомендуется включать информацию о методе сбора, продолжительности полевых работ, типе биотопа, субстрата и т.д.

Оцифровка литературных данных в электронный формат может быть реализована на основе различных программных решений. При небольшом числе специалистов, ведущих оцифровку, достаточный функционал предоставляют электронные таблицы. При значительном числе участников (как в случае проекта о пауках) требуется специализированное веб-приложение.

Во время планирования этого этапа работ важно обратить внимание на следующие положения:

1. Создание фильтров или выпадающих списков для стандартизации ввода данных.

2. Необходимость хранения исходных данных случае их формализации. Это важно при неверной интерпретации или обнаружении более подробных сведений о конкретной находке.

3. Возможность выбора оцифровщиком источников, которые ему более интересны или удобны для ввода.

4. Расстановка приоритетов. В некоторых случаях в первую очередь важнее оцифровать более поздние аннотированные списки видов по определенной территории, чем ранние разрозненные статьи. Часть библиографических источников может содержать только косвенные упоминания находок и их оцифровка может быть отложена.

5. В случае вовлечения в проект волонтеров, особое внимание необходимо уделить обратной связи и возможностям поощрения и поддержки. Например, волонтер может получать сообщения о личной или общей статистике проекта, поощрения в виде благодарственных писем или сертификата участника и др.

Во всех рассматриваемых проектах для представления оцифрованных данных использовали стандарт Darwin Core [18]. Этот общепринятый на мировом уровне стандарт содержит достаточное число терминов для представления литературных сведений и позволяет хранить как первичную, так и генерализованную информацию о находках. Кроме того, Darwin Core является основным стандартом для обмена данными в GBIF, что позволяет автоматизировать процесс публикации в этом репозитории.

При объединении данных, полученных из различных источников, опубликованных в разное время, неизбежно возникает необходимость сопоставления устаревших и современных таксономических названий. В процессе этой работы могут быть выявлены несоответствия между таксономическим справочником GBIF Backbone Taxonomy и принятыми по целевой группе таксономическими списками. Например, в ходе оцифровки литературы по дождевым червям было выяснено, что в GBIF Backbone отсутствуют некоторые названия из сводок Т.С. Перель [19, 20], которыми в основном руководствуются отечественные специалисты. Также выявлены несоответствия GBIF Backbone современному таксономическому списку [21].

Ручная географическая привязка находок видов на основе текстовых описаний мест сборов является важным этапом процесса оцифровки. Даже в современных публикациях авторы далеко не всегда приводят географические координаты мест проведения исследований. Например, в процессе оцифровки данных о распространении дождевых червей было получено 5304 записи, из которых только для 310 авторами были указаны координаты.

Общепринятым для ручной геопривязки является метод “точка-радиус” [22]. В процессе работы помимо популярных картографических онлайн сервисов часто требуется привлечение дополнительных картографических и исторических источников. Особенно это актуально для сборов, проведенных в советский период. Многие населенные пункты с тех пор были переименованы, а некоторые уже не существуют. Исследователям также доступны готовые веб-приложения для геопривязки и оценки радиуса, например недавно представленное GeoPick [23]. Однако возможности его применения для постсоветского пространства требуют дополнительной оценки.

Таким образом, ручная геопривязка – трудозатратный и длительный процесс, который может занять время, сопоставимое с непосредственно оцифровкой (этап 2). В то же время, именно качественно геопривязанные находки видов имеют наибольшую ценность для моделирования.

Для публикации данных в открытом доступе целесообразно использовать репозиторий GBIF. Все наборы данных (datasets) в GBIF имеют авторство и DOI (Digital Object Identifier), а также связаны с организациями, от имени которых были опубликованы. Функционал портала позволяет отслеживать выгрузки данных другими пользователями и их цитирование в литературе. При этом актуальной остается задача организации доступа к оцифрованным данным через национальный ресурс.

3. Заключение

Оцифровка научного литературного наследия является актуальным в мире направлением, которое в последние годы получило развитие и в России. Однако ввиду отсутствия тематической национальной программы, охват этих работ в нашей стране остается фрагментарным. В данной публикации на основе опыта трех региональных проектов, направленных на оцифровку отечественной научной литературы по разным группам организмов, сформулированы основные этапы этого процесса, обозначены потенциальные проблемы и описаны пути их решения. Обобщенный опыт трех проектов показывает, что работы по оцифровке требуют тщательного планирования всех ее этапов. При этом, поиск полнотекстовых публикаций и географическая привязка оцифрованных записей могут потребовать сопоставимые с оцифровкой данных трудозатраты. Наибольшей ценностью обладают оцифрованные находки видов с геопривязкой, основанной на указании точного места нахождения организма. Такие данные могут быть наиболее эффективно использованы в моделировании, описании ареалов распространения, оценке риска исчезновения видов.

4. Список литературы

1. Phillips S.J., Anderson R.P., Schapire R.E. Maximum entropy modeling of species geographic distributions. *Ecological Modelling*. 2006. V. 190. № 3–4. P. 231–259. doi: 10.1016/j.ecolmodel.2005.03.026
2. Hallgren W., Beaumont L., Bowness A., et al. The Biodiversity and Climate Change Virtual Laboratory: Where ecology meets big data. *Environmental Modelling & Software*. 2016. V. 76. № 4. P. 182–186. doi: 10.1016/j.envsoft.2015.10.025
3. Edwards J.L., Meredith A.L., Nielsen E.S. Interoperability of Biodiversity Databases: Biodiversity Information on Every Desktop. *Science*. 2000. V. 289. № 5488. P. 2312–2314. doi: 10.1126/science.289.5488.2312
4. Graham C.H., Ferrier S., Huettman F., et al. New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology & Evolution*. 2004. V. 19. № 9. P. 497–503. doi: 10.1016/j.tree.2004.07.006
5. Heberling J.M., Noesgaard D., Weingart S.B., et al. Data integration enables global biodiversity synthesis. *PNAS*. 2021. V. 118. № 6. P. 1–7. doi: 10.1073/pnas.2018093118
6. Ivanova N.V., Shashkov M.P. The possibilities of GBIF data use in ecological research. *Russian Journal of Ecology*. 2021. V. 52. № 1. P. 1–7. doi: 10.1134/S1067413621010069
7. Aceves-Bueno E., Adeleye A.S., Feraud M., et al. The Accuracy of Citizen Science Data: A Quantitative Review. *The Bulletin of the Ecological Society of America*. 2017. V. 98. № 4. P. 278–290. doi: 10.1002/bes2.1336.
8. Bochkov D.A., Seregin A.P. Local floras of Russia: records from literature. Version 1.75. *Occurrence dataset*. Lomonosov Moscow State University. URL: https://doi.org/10.15468/rxtjt2 (accessed 18.09.2024).
9. Bolshakov S., Kalinina L., Palomozhnykh E., et al. Agaricoid and boletoid fungi of Russia: the modern country-scale checklist of scientific names based on literature data. *Biological Communications*. 2021. V. 66. № 4. P. 316–325. doi: 10.21638/spbu03.2021.404
10. Seregin A.P., Basov Y.M. Fleroff goes digital: georeferenced records from "Flora des Gouvernements Wladimir" (Fleroff, 1902). *Biodiversity Data Journal*. 2021. V. 9. № e75299. doi: 10.3897/BDJ.9.e75299
11. Filippova N., Arefyev S., Zvyagina E., et al Fungal literature records database of the Northern West Siberia (Russia). *Biodiversity Data Journal*. 2020. V. 8. № e52963. doi: 10.3897/BDJ.8.e52963
12. Filippova N., Ageev D., Bolshakov S., et al. The fungal literature-based occurrence database for southern West Siberia (Russia). *Biodiversity Data Journal*. 2021. V. 9. № e76789. doi: 10.3897/BDJ.9.e76789
13. Shashkov M., Ivanova N., Ermolov S.Filling Gaps in Earthworm Digital Diversity in Northern Eurasia from Russian-language Literature. *BISS*. 2023. V. 7. Article No. e112957. doi: 10.3897/biss.7.112957
14. Созонтов А.Н. Мобилизация данных о распространении пауков (Araneae) России с привлечением возможностей citizen science. В: *XVI съезд Русского энтомологического общества* (Москва, 22–26 августа 2022 г.). М.: Т-во научных изданий КМК, 2022. С. 153.
15. Chavan V., Penev L. The data paper: a mechanism to incentivize data publishing in biodiversity science. *BMC Bioinformatics*. 2011. V. 12. No. 15. doi: 10.1186/1471-2105-12-S15-S2
16. Esyunin S.L., Efimik V.E. *Catalogue of the spiders (Arachnida, Aranei) of the Urals*. M.: KMK Scientific Press Ltd., 1996. 229 p.
17. Mikhailov K.G. Bibliographia Araneologica Rossica (1770-2022). Bibliography on spiders of Russia and post-Soviet Republics. *Zoologicheskie Issledovania*. 2024. № 22. 227 p.
18. Wieczorek J., Bloom D., Guralnick R., et al. Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. *PLoS ONE*. 2012. V. 7. № 1. Article No. e75299. doi: 10.1371/journal.pone.0029715
19. Перель Т.С. *Распространение и закономерности распределения дождевых червей фауны СССР*. М.: Наука, 1979. 272 с.
20. Всеволодова-Перель Т.С. *Дождевые черви фауны России. Кадастр и определитель*. М.: Наука, 1997. 102 с.
21. Brown G.G., James S.W., Csuzdi C., et al. A checklist of megadrile earthworm (Annelida: Clitellata) species and subspecies of the world. *Zenodo*. 2023. doi: 10.5281/zenodo.7301848
22. Wieczorek J., Guo Q., Hijmans R. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International Journal of Geographical Information Science*. 2004. V. 18. № 8. P. 745–767. doi: 10.1080/13658810412331280211
23. Marcer A., Escobar A., Chapman A.D., Wieczorek J.R. GeoPick – A web application for georeferencing natural history collections following best practices. *Ecography*. 2024. doi: 10.1111/ecog.07431