БАЗА ЛИТЕРАТУРНЫХ ДАННЫХ О НАХОДКАХ ГРИБОВ ЗАПАДНОЙ СИБИРИ

АЛГОРИТМ ОЦИФРОВКИ ДАННЫХ

В процессе оцифровки данных был использован следующий алгоритм, позволяющий структурировать и формализовать операции. Он может быть использован как шаблон для повторения оцифровки данных по другим группам организмов в регионе.

1. Формирование библиографического списка научных работ разного типа (статьи, монографии, диссертации, главы в книгах и др.), имеющих отношение к изучению грибов в регионе. При поиске мы опирались на знакомый нам список авторов, работающих на территории, которые лично подготовили списки своих работ и известные им другие источники по разным группам. Дополнительно происходил поиск по цитируемым источникам, региональным Красным Книгам, библиографическим работам, и основным журналам. Найденные источники вносились в библиографический менеджер (в нашем случае Zotero).
2. Поиск, сканирование и распознавание полных текстов работ и добавление их в Zotero. В результате была создана электронная библиотека, которая размещена на сайте проекте в открытом доступе.
3. Создание базы данных библиографических источников, которая будет служить для мониторинга процесса оцифровки. В этой базе данных, кроме основной библиографической информации, содержатся колонки о количестве находок видов (если имеются), дате и авторе их внесения в базу, общем описании источника данных (аннотированный список, региональная сводка, теоретическая работа, экологическое исследование, описание нового вида и пр.), качестве геопривязки, наличии даты наблюдения/сбора находок, наличие гербарных номеров находок. Такая база данных позволяет вести трек процесса оцифровки и качества получаемых в итоге данных.
4. Создание самой базы данных находок на основе Google Sheets (или Excel). Формирование структуры базы данных (полей) происходило в соответствии с форматом Darwin Core и типом поступающих данных. Формат данных также определялся DwC и договоренностью между авторами базы данных.
5. Внесение находок видов в базу. Мы не использовали фильтр на качество данных и вносили все упоминания, будь то указание конкретной находки с координатой, или просто сообщение о нахождении вида в регионе. Далее опишем алгоритм заполнения каждого поля базы данных:
    a. bibliographicCitation – указание на ссылку источника данных.
    b. originalNameUsage – исходное название таксона, как оно сообщалось в публикации. Для коррекции грамматических ошибок после заполнения этого поля, оно проверялось через GBIF Species matching tool (www.gbif.org/tools/species-lookup).
    c. identificationQualifier – для указаний на сомнительные определения, если об этом сообщается в публикации (cf., aff.).
    d. habitat – в это поле вносилась информация об особенностях местообитания, типе растительности и субстрате. Для более детального анализа этой информации, имеет смысл вносить ее в разные поля и вводить классификаторы, чего не было сделано на настоящем этапе.
    e. verbatimLocality – поле для внесения географического описания уровнем ниже района.
    f. Locality – перевод вышеназванного поля на английский язык, часто в сокращенном виде.

g. year – год, в случае если публикация не содержала информации о дате находки, в это поле вносился год публикации.

h. month – месяц, если публикация содержит дату находки.

i. day – день, если публикация содержит дату находки.

j. fieldNumber – полевой номер, или номер в коллекции, если указан в публикации.

k. basisOfRecord – основание находки, имеет значение HumanObservation для сообщения находок без номеров гербария, PreservedSpecimen – для находок с сообщением номеров.

l. countryCode – код страны, RU.

m. stateProvince – административная единица уровнем ниже страны (регион, республика, область, край, или округ). Названия указывались на латинице, как в Google maps.

n. county – административная единица уровнем ниже stateProvince (район). Названия указывались латиницей, как в Google maps.

o. decimalLatitude – широта. В зависимости от качества геопривязки в источнике, она дополнялась в момент внесения ее в базу. Для видов, где указывалось географическое описание места сбора, координаты извлекались с использованием Google maps. Для сообщений о находках в регионе бралась центральная координата региона (из википедии) с радиусом неточности, покрывающем его границы. Все координаты приводились к общему формату ГГ.ГГГГГ.

p. decimalLongitude – долгота. Детали см. выше.

q. coordinateUncertaintyInMeters – радиус неточности, т.е. расстояние в пределы которого может попасть координата от указанной точки, определялось следующим образом: 1) указание конкретных координат находок видов или площадок соответствует неточности GPS (в среднем 3-10 м), 2) указание координат или географического описания места работы соответствует неточности от 500 м до 5 км и зависит от длины маршрутов, 3) сообщение о нахождении вида в регионе или на территории ООПТ дает неточность, равную радиусу круга, в который попадает вся территория (в FuNWS до 100 км).

r. georeferenceSources – источник геопривязки, в нашем случае Google или Yandex maps.

s. rank – таксономический ранг находки, заполняется автоматически с помощью GBIF Species matching tool на последней стадии заполнения базы данных.

t. kingdom – царство, заполняется автоматически с помощью GBIF Species matching tool на последней стадии заполнения базы данных. То же самое относится к остальным шести полям. При заполнении этих полей из таксономической системы GBIF автоматически происходит синонимизация таксонов. Эти поля в дальнейшем использовались для таксономического анализа данных, загруженных в FuNWS.

u. phylum

v. class

w. order

x. family

y. genus

z. species

6. На последнем этапе производится верификация данных, т.е. поиск возможных ошибок в сформированном объеме базы данных. Для этого подходят как простые методы фильтров и сортировки, так GIS и специальные инструменты GBIF.

7. Обновление набора данных и метаданных (его описания) в GBIF по мере появления новых публикаций или других правок в базе данных.

8. По мере роста базы данных возможна публикация «статьи о данных» для формализации выполненной работы в научной печати. Подробнее см. в статье: http://gbif.ru/datapaper

База данных FuNWS не является законченным продуктом и будет пополняться по мере выхода новых публикаций. Также будет продолжена оцифровка уже опубликованных работ, которые не были найдены или показались нам второстепенными на первом этапе формирования источников. Обновленные версии базы будут регулярно загружаться в GBIF.