DOI 10.31489/2023BMG3/73-83

УДК 57.087

Н.В. Иванова*

Институт математических проблем биологии РАН — филиал Федерального государственного учреждения «Федеральный исследовательский центр Институт прикладной математики имени М.В. Келдыша Российской академии наук», Пущино, Россия
**Автор для корреспонденции: Natalya.dryomys@gmail.com*

Рекомендации по оценке качества данных *iNaturalist*

Данные, собранные волонтерами (*citizen science*), составляют значительную долю сведений о биоразнообразии, доступных через глобальный портал GBIF. Тем не менее, многие профессиональные исследователи скептически относятся к возможности их использования для научного анализа. Целью исследования стала разработка методики для оценки качества данных любительских наблюдений на примере системы *iNaturalist*. Для достижения цели через портал GBIF получены данные *iNaturalsit* в табличном виде. Выполнена оценка полноты и качества данных. Показано, что до трети наблюдений в рассматриваемых выборках содержали некорректные или подозрительные значения, однако большинство наблюдений имели высокое качество и могут быть использованы для научного анализа. На основе полученных результатов сформулированы следующие критерии, которым должны соответствовать наблюдения *iNaturalist*: (1) наблюдение должно быть определено экспертом, имеющим соответствующую квалификацию, (2) дата наблюдения должна быть указана корректно, (3) указанные географические координаты должны соответствовать реальному месту наблюдения. Предложены методы для выявления некорректных значений. Показано, что на основе доступных через GBIF данных можно самостоятельно проверить правильность определения изучаемого таксона, оценить научную квалификацию экспертов, определивших наблюдение в *iNaturalist* и верифицировать дату наблюдения. Оценка корректности географических координат мест наблюдений является более сложной задачей. Для проверки этих данных требуется наиболее тщательный анализ с привлечением дополнительных источников информации.

Ключевые слова: любительские наблюдения, GBIF, Darwin Core, анализ массивов объединенных данных, геоданные, ORCID.

Введение

Анализ объединенных данных, полученных из разных источников, является мировым трендом в науке о биоразнообразии [1, 2]. Такой подход позволяет решать задачи на макрорегиональном и глобальном уровнях, используя современные методы статистического и пространственного анализа [3–8]. Прогрессу в этой области способствовала оцифровка крупнейших мировых научных коллекций и их размещение в открытом доступе в сети Интернет [9–11]. В последние годы всё больший вклад в открытые данные о распространении видов вносят не профессиональные исследователи, а волонтеры-натуралисты [12].

На мировом уровне практика привлечения волонтеров к трудоемкому и зачастую дорогостоящему сбору научных данных [13] распространена довольно широко и позволяет существенно дополнить сведения, собираемые профессиональными исследователями [14–19]. В англоязычной литературе такая деятельность получила название *citizen science* (или *community science, public participation in scientific research*) [14], а волонтеров, которые помогают исследователям собирать или анализировать данные, называют *citizen scientists*. Для обозначения таких активностей в русскоязычной среде широко используется термин «гражданская наука» [20–22], хотя такой прямой перевод не является корректным. Во-первых, он неправильно передает изначальный смысл (здесь «citizen» употребляется как «горожанин», а не «гражданин»), а во-вторых, данное сочетание уже зарезервировано для обозначения исследований не военной тематики. По мнению автора, более удачным переводом *citizen science* является сочетание «любительские наблюдения», которое используется в тексте статьи.

Важно отметить, что объем данных, собираемых через системы любительских наблюдений, в мире возрастает существенно быстрее по сравнению с «научными» источниками. Так, крупнейшим ресурсом в Международном репозитории о биоразнообразии GBIF (см. ниже) является массив, собранный любителями птиц – eBird Observation Dataset [23], включающий >1 млрд. наблюдений. Число же образцов из научных коллекций, доступных через GBIF, составляет чуть более 213 млн.

Очевидно, что любительские наблюдения имеют большой потенциал для научного анализа. Часто они являются единственным источником данных о биоразнообразии той или иной территории [24]. В то же время многие ученые выражают сомнения относительно надежности таких данных и возможностей их использования для научного анализа наряду с материалами цифровых научных коллекций и публикаций.

Традиционно анализ объединенных данных начинается с оценки их качества, то есть проверки и верификации (*data cleaning*). Методика оценки качества данных в целом хорошо разработана и представлена в ряде работ [25–28], некоторые общепринятые методы реализованы в пакетах для среды статистического программирования R [29, 30]. В то же время существующие методики были разработаны в основном на основе материалов цифровых научных коллекций. Любительские наблюдения имеют определенную специфику, которую необходимо учитывать при оценке их качества и пригодности для решения задач конкретного исследования.

Целью данной работы стала разработка методики оценки качества данных из систем для сбора любительских наблюдений. Исследование выполнено на основе сведений из системы iNaturalist, доступных через репозиторий GBIF.

Методы и материалы

Описание системы iNaturalist. Система для сбора любительских наблюдений iNaturalist (https://www.inaturalist.org/) разработана выпускниками университета Беркли (США) в 2008 г. [31]. На январь 2023 г. она включает >125 млн наблюдений >411 тыс. видов со всего мира, сделанных >2.5 млн натуралистов. Каждый пользователь iNaturalist имеет личный аккаунт, через который он загружает наблюдения. Делать это можно при помощи мобильного приложения iNaturalist или через вебсайт. Основанием для наблюдения может быть фотография (серия фотографий) или аудиозапись. При загрузке наблюдения система предлагает автоматическое определение таксона по фотографии, основанное на работе нейронной сети, пользователь может согласиться с ним, или указать собственное. Все новые наблюдения получают статус «требуется определение» (NeedsID). Если определение подтверждается экспертами (то есть любыми зарегистрированными участниками сети iNaturalist), имеет географическую привязку и наблюдаемый объект не является культурным, он получает «исследовательский» уровень (ResearchGrade). Для видов, точные координаты которых не подлежат разглашению (sensitive data), есть функция генерализации геоданных.

Популярность iNaturalist обусловлена глобальным пространственным и универсальным таксономическим охватом, а также возможностью создавать тематические проекты, обобщающие наблюдения целевых таксонов, или наблюдения, сделанные за определенный период времени, либо же на определенной территории. Например, проект «Флора России и Крыма» [22] обобщает>2.2 млн наблюдений сосудистых растений.

Данные iNaturalist, доступные через глобальный портал GBIF. Global Biodiversity Information Facility (GBIF) — крупнейший Международный репозиторий открытых данных о биоразнообразии [32]. На начало 2023 г. через GBIF доступно >2.2 млрд записей о находках видов, происходящих из>80 тыс. источников (наборов данных). Все данные бесплатно доступны для анализа при соблюдении правил их использования. Данные из системы iNaturalist представлены в виде отдельного набора данных — iNaturalist Research-grade Observations [33]. В GBIF экспортируются не все наблюдения, а только имеющие «исследовательский» уровень, а также совместимую с GBIF лицензию (CC-0, CC-BY или CC-BY-NC). В отличие от «научных» наборов данных, представленных в GBIF, контроль над данными и качеством наблюдений практически полностью лежит на участниках сети iNaturalist. Разработчики платформы обеспечивают только ее техническую работу и автоматический экспорт данных в GBIF, при этом не оценивают качество наблюдений, которые загружают пользователи.

*Разработка методики оценки качества данных iNaturalist, доступных через GBIF***.** Для обеспечения совместимости данных, происходящих из разных источников, GBIF использует единый обменный стандарт Darwin Core [34], разработанный научно-образовательной ассоциацией Biodiversity Information Standards (TDWG). Все данные, как публикуемые, так и выгружаемыми пользователями портала для дальнейшего анализа, приведены к этому стандарту. Информация хранится в таблицах со строго определенным набором полей (терминов, заголовков столбцов), правила заполнения которых регламентированы. В то же время в силу разнородности источников данных, индексируемых GBIF, информация может быть представлена с разной подробностью, содержимое некоторых полей в разных наборах данных может отличаться.

Поэтому на первом этапе работы была проанализирована полнота данных iNaturalist, доступных через GBIF. Для этого через интерфейс GBIF получены выборки наблюдений, сделанных автором статьи [35] и определенных автором [36] в формате Darwin Core Archive. По этим таблицам оценена доступность следующей информации: авторство наблюдения, таксономия, сведения о дате и месте наблюдения, ссылка на исходное наблюдение в системе iNaturalist. На следующем этапе проанализированы возможности использования доступных сведений для верификации определения таксонов, оценки научной квалификации экспертов, корректности даты наблюдения и возможности проверки геоданных. На основе полученных результатов сформулированы рекомендации по оценке качества данных iNaturalist.

Результаты и их обсуждение

Особенности организации данных iNaturalist в таблицы DarwinCore

Авторство наблюдений. Ссылка на исходное наблюдение (в системе iNaturalist) доступна в полях dwc: occurrence ID и dwc: reference. Для указания автора наблюдения используется поле dwc: recordedBy, для определившего наблюдение — dwc: identified By. Натуралисты могут связать свой профиль с идентификатором ORCID (Open Researcher and Contributor ID). Через GBIF информация об ORCID пользователей iNaturalist доступна в полях dwc: recorded By ID и (или) dwc: identified By ID.

Сведения об определении наблюдения. Информация об идентификации таксонов представлена в нескольких полях. Научное название вида (или таксона более высокого ранга, до которого удалось определить объект) хранится в поле dwc: scientific Name, соответствующий ранг таксона — в dwc: taxon Rank. Дополнительные данные о таксономии доступны в полях dwc:kingdom, dwc: phylum, dwc: class, dwc: order, dwc: family, dwc: subfamily, dwc: genus, dwc: subgenus, dwc: specific Epithet, dwc: infraspecific Epithet. При необходимости исследователи могут самостоятельно проверить правильность определения, поскольку через GBIF доступны фотографии и аудиозаписи, послужившие основанием для наблюдений.

Сведения о дате наблюдения. Сведения о дате и времени наблюдения согласно ISO 8601-1:2019 представлены в поле dwc: eventDate, в поле dwc: verbatim EventDate — согласно EXIF фотографии. Информация о времени наблюдения также доступна в поле dwc: event Time, о годе, месяце и дне — в dwc: year, dwc: month и dwc: day соответственно.

Геоданные. При загрузке наблюдений в iNaturalist координаты экспортируются автоматически из EXIF фотографии либо вводятся натуралистами вручную. Также есть возможность указать погрешность определения координат. Встроенные GPS-приемники некоторых моделей смартфонов фиксируют погрешность автоматически. Остальные геоданные (страна и топонимика места наблюдения) генерируется автоматически на основе координат и используемого в iNaturalist картографического сервиса. Информация о методе привязки (автоматическая или ручная) в iNaturalist отсутствует.

В таблицах Darwin Core координаты места наблюдения, широта и долгота, доступны в полях dwc: decimal Latitude и dwc: decimal Longitude соответственно, погрешность определения координат — в dwc: coordinate Uncertainty In Meters. Для указания страны используются dwc: country и dwc: country Code, для указания топонимов — dwc: state Province, dwc: verbatim Locality и др.

Таким образом, доступные через GBIF данные iNaturalist включают большой объем атрибутивной информации о наблюдениях, которая может быть использована для оценки качества данных. Важным преимуществом является доступность фотографий и аудиозаписей, а также исходных наблюдений.

Оценка качества данных

Оценка научной квалификации экспертов. В выборке наблюдений, сделанных автором статьи, содержалось 11756 записей. Из них экспертами iNaturalist определены 2244 наблюдения, остальные были определены автором самостоятельно и подтверждены экспертами. В определении наблюдений, которые автор не смог идентифицировать, участвовали 79 экспертов, из которых 61 имеют ORCID. Доля наблюдений, определенных экспертами, имеющими ORCID, составляет 70,9 % (1593 записи). Таким образом, определения 11105 наблюдений можно рассматривать как корректные, 651 наблюдение нуждается в проверке специалистами.

Верификация дат наблюдений. Выборка наблюдений, определенных автором статьи, включала 9702 записи, из которых 190 — наблюдения пользователей iNaturalist, а остальные — собственные наблюдения автора. Из этих 190 записей 187 относились к сосудистым растениям. Все они являются

обычными, широко распространенными видами. Большинство наблюдений сделано в течение вегетационного сезона, сроки встреч соответствуют общепринятым представлениям о фенологии наблюдаемых видов в районах их произрастания (рис. 1). Ручная проверка наблюдений не выявила несоответствий между указанной датой и сезоном наблюдения на фотографиях. При этом для 83 записей (44,4%) дата наблюдения отличалась от даты загрузки в iNaturalist. Как правило, эта разница составляла от одного до нескольких дней, но в некоторых случаях достигала нескольких (максимум 12) лет. В целом, можно заключить, что все указанные даты корректны.

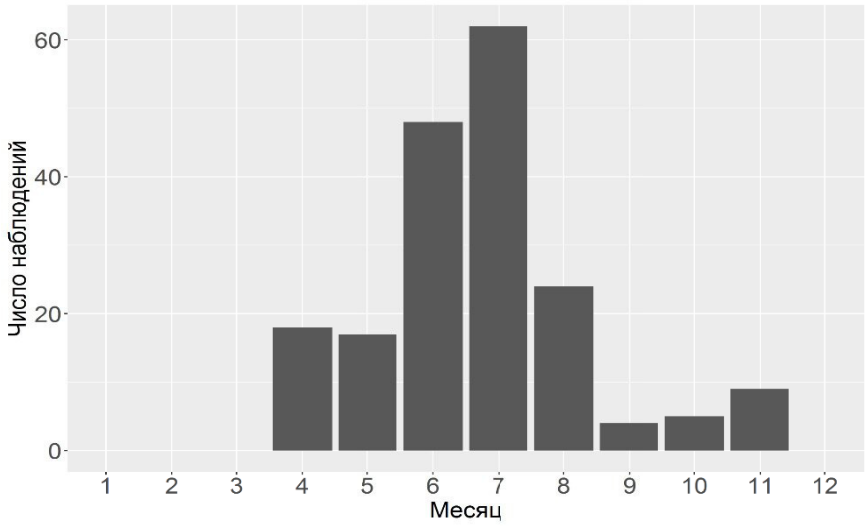

Рисунок 1. Распределение наблюдений iNaturalist по месяцам в анализируемой выборке

Оценка качества геоданных. Для оценки корректности географических координат использовали наблюдения сосудистых растений, анализируемые на предыдущем этапе. В целом, все точки наблюдений соответствовали представлениям об ареалах рассматриваемых видов. Результаты ручной верификации показали, что в одном наблюдении точка обнаружения сухопутного папоротника *Pteridium pinetorum* находилась в акватории пруда, что явно указывает на ошибочные координаты. Для остальных наблюдений подобных несоответствий не выявлено.

Также проанализированы доступные данные о погрешностях определения координат. Выяснено, что погрешность указана для 142 записей. При этом для 76,1 % наблюдений (109 записей) она не превышала 50 м, а для 79,6 % (113 записей) — 100 м (рис. 2). Среди наблюдений, для которых погрешность не была указана, 19 сделаны на территории населенных пунктов, то есть в зоне уверенного покрытия мобильной связи. Скорее всего, погрешность определения координат этих наблюдений невелика (первые десятки метров).

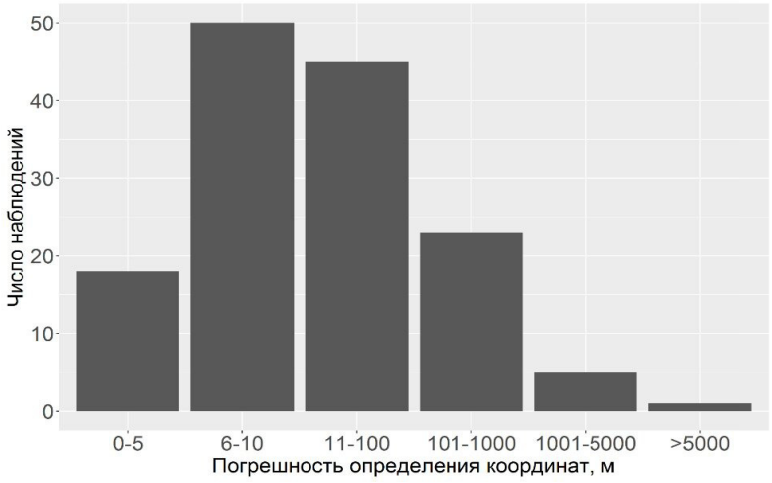

Рисунок 2. Погрешность определения координат в анализируемой выборке

Также оценено взаимное расположение точек наблюдений за одну дату относительно времени, в которое они были сделаны. В качестве примера приведем наблюдения пользователя max_carabus, сделанные 30 апреля 2018 г. в лесном массиве заповедника «Калужские засеки» (Россия, Калужская обл.). Все отмеченные виды являются обычными для флоры широколиственных лесов. Для всех наблюдений указана погрешность определения координат, которая составляет от 6 до 19 м. Как видно из рисунка 3, точки образуют кластеры, которые хорошо согласуются со временем, в которое были сделаны наблюдения. На основании этого можно заключить, что координаты всех рассматриваемых наблюдений с высокой вероятностью были определены корректно.

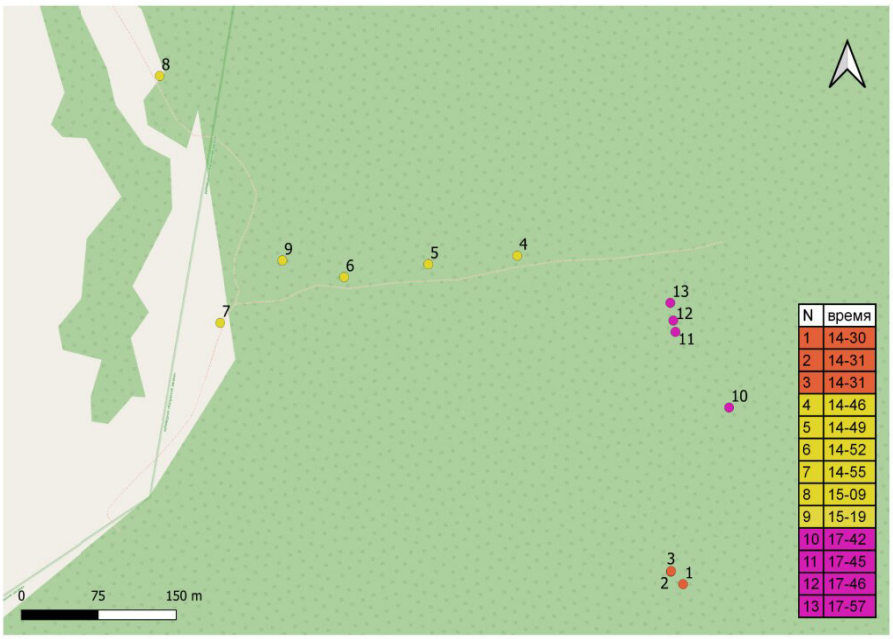

Рисунок 3. Наблюдения, сделанные натуралистом в течение дня. Подложка схемы — Open Street Map

Таким образом, в анализируемых выборках в зависимости от критерия оценки до трети наблюдений содержали подозрительные или некорректные значения. Их использование может существенно исказить результаты дальнейшего анализа и выводы. В то же время большинство наблюдений характеризовались высоким качеством данных и потенциально могут быть использованы для научного анализа.

Рекомендации по оценке качества данных iNaturalist, доступных через GBIF

На основе обобщения результатов, полученных на предыдущем этапе, разработаны рекомендации по оценке качества данных iNaturalist (см. табл.). Предлагаемая методика также может быть использована для проверки и верификации данных из других систем для сбора любительских наблюдений.

Прежде всего исследователям, планирующим использовать данные iNaturalist (или другой citizen science платформы), необходимо убедиться, что целевые таксоны можно с уверенностью идентифицировать по фотографиям или аудиозаписям. Если это не представляется возможным, то данные из систем для сбора любительских наблюдений следует исключить из анализа.

Оценку качества данных iNaturalist целесообразно начать с оценки квалификации экспертов, определивших наблюдения. Хотя все наблюдения из iNaturalist содержат однозначное указание авторства, выявить наблюдения, сделанные или идентифицированные профессиональными исследователями, в общем массиве данных часто затруднительно. Натуралисты (вне зависимости от их научной квалификации) могут указывать как свое настоящее имя, так и произвольно выбранный никнейм. Также необходимо отметить, что участники iNaturalist не связаны со своими организациями. Хотя аффилиация может быть указана в личном аккаунте, она не экспортируется в GBIF. В качестве критерия для поиска наблюдений профессиональных исследователей можно использовать наличие ORCID (как это было сделано в данной работе). Тем не менее далеко не все ученые указывают свой ORCID в iNaturalist. Поэтому рекомендуется ознакомиться с основными экспертами по целевым так-

сонам и их квалификацией. Помимо наличия ORCID, следует обратить внимание на число сделанных экспертом определений. Эта информация доступна на личных страницах натуралистов.

Таблица

Предлагаемые рекомендации по оценке качества наблюдений iNaturalist

Критерий	Индикатор высокого качества данных	Метод оценки и выявления некорректных значений
Наблюдение определено экспертом, имеющим соответствующую квалификацию	Наличие у эксперта большого числа определений целевого таксона и (или) ORCID	Оценка квалификации эксперта на основе его профиля в iNaturalist
Дата наблюдения указана корректно	Соответствие даты наблюдения сезону на фотографии	Анализ распределения наблюдений по месяцам с учетом биологии изучаемых таксонов. Ручная верификация наблюдений
Указанные географические координаты должны соответствовать реальному месту наблюдения	Положение точек наблюдений согласуется с представлениями о биологии изучаемых таксонов Наличие данных о погрешности определения координат. Величина погрешности соответствует требованиям исследования. Наблюдения сделаны в пределах крупных населенных пунктов (в зоне уверенного покрытия мобильной связи)	Анализ пространственного расположения точек наблюдений с учетом имеющихся знаний о биотопической приуроченности изучаемых таксонов Сопоставление координат с другими наблюдениями, сделанными пользователем в конкретную дату Анализ согласованности времени наблюдения и расположения точек наблюдений относительно друг друга Анализ значений погрешности определения координат и приуроченности наблюдений к урбанизированным или природным территориям

Тщательность проверки корректности определений может быть разной и зависит от простоты идентификации исследуемых таксонов по фотографиям или аудиозаписям, а также от представленности этих таксонов в системе iNaturalist. Для широко распространенных видов, которые легко определяются и представлены в iNaturalist большим числом наблюдений (несколько тысяч) верификация может быть минимальной. Если объектами исследования являются редкие или плохо определяемые виды, рекомендуется проверять корректность определения каждого наблюдения, используя доступные фотографии или аудиозаписи. При обнаружении ошибок в определении строго рекомендуется предлагать корректное определение в iNaturalist. Также рекомендуется обращаться к исходным наблюдениям (в iNaturalist), чтобы ознакомиться с историей определений, поскольку в GBIF экспортируются только сведения о пользователе, первым сделавшим признанное правильным определение.

Во многих исследованиях принципиально важной является информация о дате наблюдения, в таких случаях настоятельно рекомендуется тщательно верифицировать соответствующие сведения. Ошибки связаны с тем, что часто натуралисты указывают не дату наблюдения объекта в природе, а дату загрузки наблюдения в iNaturalist. Временной промежуток между этими событиями может составлять от нескольких дней до нескольких лет. Поэтому рекомендуется всегда обращаться к изображениям или аудиозаписям для оценки корректности указанной даты наблюдения.

Наиболее проблемным вопросом при оценке качества данных iNaturalist является верификация географических координат, то есть проверка их соответствия реальному месту наблюдения. В силу выявленных особенностей геоданных iNaturalist (см. раздел «Оценка полноты данных») о корректности определения координат можно судить только по косвенным признакам. Традиционная в таких случаях проверка соответствия координат описанию места наблюдения [25] не является корректной. Поэтому рекомендуется начинать оценку качества геоданных с визуализации точек наблюдений и сравнения результатов с имеющимися сведениями об ареалах и биотопической приуроченности изу-

чаемых таксонов. На следующем этапе следует использовать доступные сведения о погрешности определения координат, а также приуроченности наблюдений к урбанизированным или природным территориям (ручная проверка). По опыту автора, значения погрешности в несколько десятков метров и более в случае определения координат встроенным в смартфон GSP-приемником могут свидетельствовать о неправильной геолокации. В то же время в случае ручной геопривязки наблюдений значение погрешности в первые сотни метров–несколько километров является вполне обычным. Пороговое значение погрешности для исключения наблюдений из анализа должно определяться исследователем самостоятельно. Наблюдения без указания погрешности, сделанные в пределах крупных населенных пунктов в зоне уверенного приема мобильной связи с высокой вероятностью геолоцированы корректно, поэтому могут быть использованы для дальнейшего анализа.

Важно также учитывать, что если объектом исследования является широко распространенный, обычный вид, то, вероятно, даже значительная погрешность определения координат (до нескольких километров) не внесет существенных искажений в результаты. Если же целевой вид является редким или связан с определенными типами местообитаний, то данные iNaturalist нуждаются в более тщательной проверке. Кроме того, полезно провести анализ расположения точек наблюдений относительно других наблюдений, сделанных в этот день натуралистом. Если координаты целевого наблюдения значительно отличаются от других наблюдений, сделанных в тот же день, скорее всего в геоданных имеется ошибка.

Заключение

Для оценки качества данных любительских наблюдений требуются специфические методы, учитывающие их особенности. На основе анализа данных из системы iNaturalist, полученных через портал GBIF, разработаны критерии, которым должны соответствовать наблюдения, пригодные для научного анализа. Предложены методы для оценки качества данных по каждому критерию. Показано, что наиболее сложной задачей является верификация геоданных. Предложенная методика может быть использована для оценки качества данных из других систем для сбора любительских наблюдений.

Исследование выполнено за счет гранта Российского научного фонда № 23-24-00112, https://rscf.ru/project/23-24-00112/

Список литературы

1 De Prins J. Global Open Biodiversity Data: Future Vision of FAIR Biodiversity Data Access, Management, Use and Stewardship / De Prins J. // Biodiversity Information Science and Standards. — 2019. — 3. — e37190. https://doi.org/10.3897/biss.3.37190

2 Wilkinson M.D. The FAIR Guiding Principles for scientific data management and stewardship / M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg. Scientific Data. — 2016. — 3(1). — 160018. https://doi.org/10.1038/sdata.2016.18

3 Phillips H.R.P. Global distribution of earthworm diversity / H.R.P. Phillips, C.A. Guerra, M.L.C. Bartz. — Science. — 2019. — 366(6464). — 480–485. https://doi.org/10.1126/science.aax4851

4 Delgado M. Differences in spatial versus temporal reaction norms for spring and autumn phenological events / M. Delgado, T. Roslin, G. Tikhonov // Proceedings of the National Academy of Sciences. — 2020. — 117(49); 31249–31258. https://doi.org/10.1073/pnas.2002713117

5 Jayathilake D.R.M. A modeled global distribution of the kelp biome / D.R.M. Jayathilake, M.J. Costello // Biological Conservation. — 2020. — 252. — 108815. https://doi.org/10.1016/j.biocon.2020.108815

6 Polidori C. Environmental niche and global potential distribution of the giant resin bee Megachile sculpturalis, a rapidly spreading invasive pollinator / C. Polidori, D. Sánchez-Fernándezc, D. // Global Ecology and Conservation. — 2020. — 24. — e01365. https://doi.org/10.1016/j.gecco.2020.e01365

7 Roslin T. Phenological shifts of abiotic events, producers and consumers across a continent / T. Roslin, L. Antão, M. Hällfors // Nature climate change. — 2021. — 11. — 241–248. https://doi.org/10.1038/s41558-020-00967-7

8 Tamme R. Global macroecology of nitrogen-fixing plants / R. Tamme, M. Pärtel, U. Kõljalg, L. Laanisto, J. Liira, Ü. Mander, M. Moora, Ü. Niinemets, M. Öpik, I. Ostonen, L. Tedersoo, M. Zobel // Global ecology and biogeography. — 2021. — 30(2). — 514–526. https://doi.org/10.1111/geb.13236

9 Hedrick B.P. Digitization and the Future of Natural History Collections. / B.P. Hedrick, J.M. Heberling, E.K. Meineke, K.G. Turner, C.J. Grassa, D.S. Park, J. Kennedy, J.A. Clarke, J.A. Cook, D.C. Blackburn, S.V. Edwards, C.C. Davis // BioScience. — 2020. — 70 (3). — 243–251. https://doi.org/10.1093/biosci/biz163

10 Groom Q. Improved standardization of transcribed digital specimen data / Q. Groom, M. Dillen, H. Hardy, S. Phillips, L. Willemse, Z. Wu // Database, 2019, baz129. https://doi.org/10.1093/database/baz129

11 Санданов Д.В. Современные подходы к моделированию разнообразия и пространственному распределению видов растений: перспективы их применения в России / Д.В. Санданов // Вестн. Том. гос. ун-та. Биология. — 2019. — № 46. — С. 82-114. — [Электронный ресурс]. — https://doi.org/10.17223/19988591/46/5

12 Gura T. Citizen science: amateur experts / T. Gura // Nature. — 2013. — 496. — 259–261. https://doi.org/10.1038/nj7444-259a

13 Bonney R. Citizen science: a developing tool for expanding science knowledge and scientific literacy / R. Bonney, C.B. Cooper, J. Dickinson, S. Kelling, T. Phillips, K.V. Rosenberg, J. Shirk // Bioscience. — 2009. — 59. — 977–984. https://doi.org/10.1525/bio.2009.59.11.9

14 Theobald E.J. Global change and local solutions: Tapping the unrealized potential of citizen science for biodiversity research / E.J. Theobald, A.K. Ettinger, H.K. Burgess, L.B. De Bey, N.R. Schmidt, H.E. Froehlich, C. Wagner, J.H.R. Lambers, J. Tewksbury, M.A. Harsch, J.K. Parrish // Biological Conservation. — 2015. — 181. — 236–244. https://doi.org/10.1016/j.biocon.2014.10.021

15 Chandler M. Contribution of citizen science towards international biodiversity monitoring. / M. Chandler, L. See, K. Copas, A.M.Z. Bonde, B.C. López, F. Danielsen, J.K. Legind, S. Masinde, A.J. Miller-Rushing, G. Newman, A. Rosemartin, E. Turak // Biological Conservation. — 2017. — 213. — 280–294. https://doi.org/10.1016/j.biocon.2016.09.004

16 Soroye P. Opportunistic citizen science data transform understanding of species distributions, phenology, and diversity gradients for global change research / P. Soroye, N. Ahmed, J.T. Kerr // Global change biology. — 2018. — 24(11). — 5281–5291. https://doi.org/10.1111/gcb.14358

17 Young B.E. Using citizen science data to support conservation in environmental regulatory contexts / B.E. Young, N. Dodge, P.D. Hunt, M. Ormes, M.D. Schlesinger, H.Y. Shaw // Biological conservation. — 2019. — 237. — 57–62. https://doi.org/10.1016/j.biocon.2019.06.016

18 Fan F. Citizen, science, and citizen science / F. Fan, S-L. Chen // East Asian Science, Technology and Society: An International Journal. — 2019. — 13. — 181–193. https://doi.org/10.1215/18752160-7542643

19 Johnson B.A. Citizen science and invasive alien species: an analysis of citizen science initiatives using information and communications technology (ICT) to collect invasive alien species observations / B.A. Johnson, A.D. Mader, R. Dasgupta, P. Kumar // P. Global Ecology and Conservation. — 2020. — 21. — E00812. https://doi.org/10.1016/j.gecco.2019.e00812

20 Фаталиев Т.Х. Вопросы обеспечения информационной безопасности в проектах гражданской науки / Т.Х. Фаталиев, Н.Н. Бердиева // Информационные технологии. Проблемы и решения. — 2019. — № 4 (9). — С. 50–55.

21 Рябова Л.А. Гражданская наука как инструмент информационного обеспечения принятия решений в Российской Арктике в условиях изменения климата / Л.А. Рябова, Е.М. Ключникова, Е.А. Боровичев, В.А. Маслобоев // Север и рынок: формирование экономического порядка. — 2020. — № 3 (69). — С. 40–55.

22 Серегин А.П. «Флора России» на платформе INaturalist: большие данные о биоразнообразии Большой страны / А.П. Серегин, Д.А. Бочков, Ю.В. Шнер и др. Журнал общей биологии. — 2020. — Т. 81. — № 3. — С. 223–233.

23 Auer T. EOD – eBird Observation Dataset. Occurrence dataset / T. Auer, S. Barker, K. Borgmann // Cornell Lab of Ornithology. — 2010. — https://doi.org/10.15468/aomfnb.

24 Ivanova N. Contribution of citizen science to biodiversity data mobilization in Russia / N. Ivanova, M. Shashkov // Biodiversity Information Science and Standards, — 2020. — 4. — e59197. https://doi.org/10.3897/biss.4.59197

25 Chapman A. Principles of Data Quality, version 1.0. / A. Chapman. Copenhagen: GBIF Secretariat. — 2005. https://doi.org/10.15468/doc.jrgg-a190

26 Mesibov R. An audit of some processing effects in aggregated occurrence records / R. Mesibov // ZooKeys. — 2018. — 751. — 129–146. https://doi.org/10.3897/zookeys.751.24791

27 Chapman A.D. Developing standards for improved data quality and for selecting fit for use biodiversity data / A.D. Chapman, L. Belbin, P.F. Zermoglio // Biodiversity Information. Science and Standards. — 2020. — 4. — e50889. https://doi.org/10.3897/biss.4.50889

28 Chapman A.D. Georeferencing best practices. / A.D. Chapman, J.R. Wieczorek // Copenhagen: GBIF Secretariat. — 2020. https://doi.org/10.15468/doc-gg7h-s853

29 Zizka A. Coordinate Cleaner: Standardized cleaning of occurrence records from biological collection databases / A. Zizka, D. Silvestro, T. Andermann, J. Azevedo, C.D. Ritter, D. Edler, H. Farooq, A. Herdean, M. Ariza, R. Scharn, S. Svantesson, N. Wengström, V. Zizka, A. Alexandre Antonelli // Methods in Ecology and Evolution. — 2019. — 5. — 744–751. https://doi.org/10.1111/2041-210X.13152

30 Robertson M.P. Biogeo: An R package for assessing and improving data quality of occurrence record datasets / M.P. Robertson, V. Visser, C. Hui // Ecography. — 2016. — 39. — 394–401. http://doi.org/10.1111/ecog.02118

31 Seltzer C. Making biodiversity data social, shareable, and scalable: reflections on iNaturalist & citizen science / C. Seltzer // Biodiversity Information Science and Standards. — 2019. — 3. — e46670. http://10.3897/biss.3.46670

32 Edwards, J.L. Research and societal benefits of the Global Biodiversity Information Facility / J.L. Edwards. —BioScience. — 2004. — 54 (6). — 485–486. https://doi.org/10.1641/0006-3568(2004)054[0486:RASBOT]2.0.CO;2

33 iNaturalist contributors& iNaturalist. iNaturalist Research-grade Observations. iNaturalist.org. Occurrence dataset. — 2012. https://doi.org/10.15468/ab3s5x

34 Wieczorek J. Darwin Core: An evolving community-developed biodiversity data standard / J. Wieczorek, D. Bloom, R. Guralnick, S. Blum, M. Döring, R. Giovanni, T. Robertson, D. Vieglais // PLoS ONE. — 2012. — 7(1). — e29715. https://doi.org/10.1371/journal.pone.0029715

35 GBIF.org. GBIF Occurrence Download. — 2023. https://doi.org/10.15468/dl.p7pxwb

36 GBIF.org. GBIF Occurrence Download. — 2023. https://doi.org/10.15468/dl.7rmd9

Н.В. Иванова

iNaturalist деректер сапасын бағалау бойынша ұсынымдар

Еріктілер жинаған деректер (citizen science) GBIF жаһандық порталы арқылы қолжетімді биоәртүрлілік туралы ақпараттың айтарлықтай үлесін құрайды. Дегенмен, көптеген кәсіби зерттеушілер оларды ғылыми талдау үшін пайдалану мүмкіндігіне күмәнмен қарайды. Зерттеудің мақсаты — iNaturalist жүйесінің мысалында әуесқойлық бақылаулар деректерінің сапасын бағалау әдістемесін әзірлеу. Мақсатқа жету үшін GBIF порталы арқылы iNaturalsit деректері кесте түрінде алынды. Деректердің толықтығы мен сапасын бағалау орындалды. Қарастырылып отырған үлгілердегі бақылаулардың үштен біріне дейін қате немесе күдікті мәндер болғаны көрсетілген, бірақ бақылаулардың көпшілігі жоғары сапалы болды және ғылыми талдау үшін пайдаланылуы мүмкін. Алынған нәтижелер негізінде iNaturalist бақылаулары сәйкес келетін келесі критерийлер тұжырымдалады: (1) бақылауды тиісті біліктілігі бар сарапшы анықтауы керек; (2) бақылау күні дұрыс көрсетілуі тиіс; (3) көрсетілген географиялық координаттар нақты бақылау орнына сәйкес келуі қажет. Қате мәндерді анықтау әдістері ұсынылған. GBIF арқылы қол жетімді мәліметтер негізінде зерттелетін таксонды анықтаудың дұрыстығын өз бетінше тексеруге, iNaturalist-те бақылауды анықтаған сарапшылардың ғылыми біліктілігін бағалауға және бақылау күнін тексеруге болатындығы көрсетілген. Бақылау орындарының географиялық координаттарының дұрыстығын бағалау қиынырақ. Бұл деректерді тексеру үшін қосымша ақпарат көздерін тарта отырып, ең мұқият талдау қажет.

Кілт сөздер: әуесқойлық бақылаулар, GBIF, Darwin Core, біріктірілген деректер массивтерін талдау, геодеректер, ORCID.

N.V. Ivanova

iNaturalist Data Quality Guidelines

Data collected by volunteers (citizen science) significantly contributes to the biodiversity data available through the GBIF global portal. However, many professional researchers are skeptical about the possibility of using citizen science data for scientific analysis. The aim of the study was to develop guidelines for the quality assessment of iNaturalist amateur observation data. We obtained iNaturalsit data through the GBIF portal. Completeness and data quality were assessed. It is shown that up to a third of the observations contained in correct or suspicious values, however, most of the observations were of high quality and can be used for scientific analysis. Based on the obtained results, the following criteria for high quality iNaturalist observations were developed: (1) the observation must be identified by a qualified expert, (2) the observation date must be indicated correctly, (3) geographical coordinates must correspond to the real place of observation. Methods for detecting incorrect values are proposed. It is shown that GBIF users can check the identification of observations, evaluate the scientific qualifications of experts who identified the observation in iNaturalist, and verify the date of observation. Estimating the geo data correctness is more difficult. Verification of these data requires the most careful analysis with the involvement of additional information sources.

Keywords: citizen science, GBIF, Darwin Core, consolidated data analysis, geo data, ORCID.

References

1 De Prins, J. (2019). Global Open Biodiversity Data: Future Vision of FAIR Biodiversity Data Access, Management, Use and Stewardship. *Biodiversity Information Science and Standards*, *3,* e37190. https://doi.org/10.3897/biss.3.37190

2 Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

3 Phillips, H.R.P., Guerra, C.A., Bartz, M.L.C. et al. (2019). Global distribution of earthworm diversity. *Science, 366*(6464), 480–485. https://doi.org/10.1126/science.aax4851

4 Delgado, M., Roslin, T., Tikhonov, G. et al. (2020). Differences in spatial versus temporal reaction norms for spring and autumn phenological events. *Proceedings of the National Academy of Sciences*, *117*(49), 31249–31258. https://doi.org/10.1073/pnas.2002713117

5 Jayathilake, D.R.M., & Costello, M.J. (2020). A modeled global distribution of the kelp biome. *Biological Conservation, 252*, 108815. https://doi.org/10.1016/j.biocon.2020.108815

6 Polidori, C. & Sánchez-Fernándezc, D. (2020). Environmental niche and global potential distribution of the giant resin bee Megachile sculpturalis, a rapidly spreading invasive pollinator. *Global Ecology and Conservation, 24*;, e01365. https://doi.org/10.1016/j.gecco.2020.e01365

7 Roslin, T., Antão, L., Hällfors, M. et al. (2021). Phenological shifts of abiotic events, producers and consumers across a continent. *Nature climate change, 11*, 241–248. https://doi.org/10.1038/s41558-020-00967-7

8 Tamme, R., Pärtel, M., Kõljalg, U., Laanisto, L., Liira, J., Mander, Ü., Moora, M., Niinemets,Ü., Öpik, M., Ostonen, I., Tedersoo, L., & Zobel, M. (2021). Global macroecology of nitrogen-fixing plants. *Global ecology and biogeography*, *30*(2), 514–526. https://doi.org/10.1111/geb.13236

9 Hedrick, B.P., Heberling, J.M., Meineke, E.K., Turner, K.G., Grassa, C.J., Park, D.S., Kennedy, J., Clarke, J.A., Cook, J.A., Blackburn, D.C., Edwards, S.V., & Davis, C.C. (2020). Digitization and the Future of Natural History Collections. *BioScience*, *70*(3), 243–251. https://doi.org/10.1093/biosci/biz163

10 Groom, Q., Dillen, M., Hardy, H., Phillips, S., Willemse, L., & Wu, Z. (2019). Improved standardization of transcribed digital specimen data. *Database, 2019,* baz129. https://doi.org/10.1093/database/baz129

11 Sandanov, D.V. (2019). Sovremennye podkhody k modelirovaniu raznoobraziia i prostranstvennomy raspredeleniu vidov rastenii: perspektivy ikh primeneniia v Rossii [Modern approaches to modeling plant diversity and spatial distribution of plant species: Implication prospects in Russia]. *Vestnik Tomskogo gosudarstvennogo universiteta. Biologiia — Tomsk State University Journal. Biology, 46,* 82–114. https://doi.org/10.17223/19988591/46/5 [in Russian].

12 Gura, T. (2013). Citizen science: amateur experts. *Nature, 496,* 259–261. https://doi.org/10.1038/nj7444-259a

13 Bonney, R., Cooper, C.B., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K.V., & Shirk, J. (2009). Citizen science: a developing tool for expanding science knowledge and scientific literacy. *Bioscience, 59,* 977–984. https://doi.org/10.1525/bio.2009.59.11.9

14 Theobald E.J., Ettinger A.K., Burgess H.K., De Bey, L.B., Schmidt, N.R., Froehlich, H.E., Wagner, C., Lambers, J.H.R., Tewksbury, J., Harsch, M.A., & Parrish, J.K. (2015). Global change and local solutions: Tapping the unrealized potential of citizen science for biodiversity research. *Biological Conservation, 181,* 236–244. https://doi.org/10.1016/j.biocon.2014.10.021

15 Chandler, M., See, L., Copas, K., Bonde, A.M.Z., López, B.C., Danielsen, F., Legind, J.K., Masinde, S., Miller-Rushing, A.J., Newman, G., Rosemartin, A., & Turak, E. (2017). Contribution of citizen science towards international biodiversity monitoring. *Biological Conservation*, *213*, 280–294. https://doi.org/10.1016/j.biocon.2016.09.004

16 Soroye, P., Ahmed, N., & Kerr, J.T. (2018). Opportunistic citizen science data transform understanding of species distributions, phenology, and diversity gradients for global change research. *Global change biology, 24*(11),; 5281–5291. https://doi.org/10.1111/gcb.14358

17 Young, B.E., Dodge, N., Hunt, P.D., Ormes, M., Schlesinger, M.D., & Shaw, H.Y. (2019). Using citizen science data to support conservation in environmental regulatory contexts. *Biological conservation*, *237,* 57–62. https://doi.org/10.1016/j.biocon.2019.06.016

18 Fan, F., & Chen, S-L. (2019). Citizen, science, and citizen science. *East Asian Science, Technology and Society: An International Journal, 13,* 181–193. https://doi.org/10.1215/18752160-7542643

19 Johnson, B.A., Mader, A.D., Dasgupta, R., & Kumar, P. (2020). Citizen science and invasive alien species: an analysis of citizen science initiatives using information and communications technology (ICT) to collect invasive alien species observations. *Global Ecology and Conservation*, *21,* E00812. https://doi.org/10.1016/j.gecco.2019.e00812

20 Fataliev, T.Kh., & Verdieva, N.N. (2019). Voprosy obespecheniia informatsionnoi bezopasnosti v proektakh grazhdanskoi nauki [Problems of providing information security in citizen science projects]. *Informatsionnye tekhnologii. Problemy i resheniia – Information Technology. Problems and Solutions, 4*(9), 50–55 [in Russian].

21 Ryabova, L.A., Klyuchnikova, E.M., Borovichev, E.A., & Masloboev, V.A. (2020). Grazhdanskaia nauka kak instrument informatsionnogo obespecheniia priniatiia reshenii v Rossiiskoi Arktike v usloviiakh izmeneniia klimata [Citizen science as a tool for information support of decision-making in the Russian Arctic under conditions of climate change]. *Sever i rynok: formirovanie ekonomicheskogo poriadka — The North and the Market: Formation of the Economic Order*, *3*(69), 40–55 [in Russian]. https://doi.org/10.37614/2220-802X.2.2020.69.003

22 Seregin, A.P., Bochkov, D.A., & Shner, J.V. et al. (2020). «Flora Rossii» na platforme iNaturalist: bolshie dannye o bioraznoobrazii Bolshoi strany ["Flora of Russia" on iNaturalist: big data on biodiversity of a big country]. *Zhurnal obshchei biologii – Biology Bulletin Reviews*, *81*(3), 223–233 [in Russian]. https://doi.org/10.31857/S0044459620030070

23 Auer, T., Barker, S., Borgmann, K. et al. (2010). EOD — eBird Observation Dataset. Occurrence dataset. *Cornell Lab of Ornithology*. https://doi.org/10.15468/aomfnb.

24 Ivanova, N., & Shashkov, M. (2020). Contribution of citizen science to biodiversity data mobilization in Russia. *Biodiversity Information Science and Standards, 4*, e59197. https://doi.org/10.3897/biss.4.59197

25 Chapman, A. (2005). *Principles of Data Quality*, version 1.0. Copenhagen: GBIF Secretariat. https://doi.org/10.15468/doc.jrgg-a190

26 Mesibov, R. (2018). An audit of some processing effects in aggregated occurrence records. *ZooKeys, 751,* 129–146. https://doi.org/10.3897/zookeys.751.24791

27 Chapman, A.D., Belbin, L., Zermoglio, P.F. et al. (2020). Developing standards for improved data quality and for selecting fit for use biodiversity data. *Biodiversity Information Science and Standards, 4,* e50889. https://doi.org/10.3897/biss.4.50889

28 Chapman, A.D., & Wieczorek, J.R. (2020). Georeferencing best practices. *Copenhagen: GBIF Secretariat*. https://doi.org/10.15468/doc-gg7h-s853

29 Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Ritter, C.D., Edler, D., Farooq, H., Herdean, A., Ariza, M., Scharn, R., Svantesson, S., Wengström, N., Zizka, V., & Alexandre Antonelli, A. (2019). Coordinate Cleaner: Standardized cleaning of occurrence records from biological collection databases. *Methods in Ecology and Evolution*, *5,* 744–751. https://doi.org/10.1111/2041-210X.13152

30 Robertson, M.P., Visser, V., & Hui, C. (2016). Biogeo: An R package for assessing and improving data quality of occurrence record datasets. *Ecography, 39,* 394–401. http://doi.org/10.1111/ecog.02118

31 Seltzer, C. (2019). Making biodiversity data social, shareable, and scalable: reflections on iNaturalist & citizen science. *Biodiversity Information Science and Standards, 3,* e46670. http://10.3897/biss.3.46670

32 Edwards, J.L. (2004). Research and societal benefits of the Global Biodiversity Information Facility. *BioScience, 54*(6), 485–486. https://doi.org/10.1641/0006-3568(2004)054[0486:RASBOT]2.0.CO;2

33 iNaturalist contributors& iNaturalist (2012). iNaturalist Research-grade Observations. *iNaturalist.org. Occurrence dataset.* https://doi.org/10.15468/ab3s5x

34 Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring M., Giovanni, R., Robertson, T., & Vieglais, D. (2012). Darwin Core: An evolving community-developed biodiversity data standard. *PLoS ONE, 7*(1), e29715. https://doi.org/10.1371/journal.pone.0029715

35 GBIF.org (2023). GBIF Occurrence Download. https://doi.org/10.15468/dl.p7pxwb

36 GBIF.org (2023). GBIF Occurrence Download. https://doi.org/10.15468/dl.7rmd9