Guidelines and best practices for the scientific use of global iNaturalist data

Corey T. Callaghan[1,*], CJ Campbell[2,3], Jane Widness[4], Benjamin R. Goldstein[5], Thomas Mesaglio[6], Carrie Seltzer[7], Tony Iwane[7], Robert P. Guralnick[8], Brittany M. Mason[1,*]

[1]Department of Wildlife Ecology and Conservation, Fort Lauderdale Research and Education Center, Institute of Food and Agricultural Sciences, University of Florida, Davie, FL
[2]Bat Conservation International, Austin TX
[3]Department of Forest and Wildlife Ecology, University of Wisconsin-Madison, Madison WI
[4]Department of Anthropology, Yale University, New Haven, Connecticut, United States of America
[5]International Union for Conservation of Nature (IUCN), Washington, District of Columbia, United States of America
[6]Evolution & Ecology Research Centre, School of Biological, Earth and Environmental Sciences, UNSW Sydney, Sydney, Australia
[7]iNaturalist
[8]Florida Museum of Natural History, University of Florida, Gainesville, Florida
*Denotes equal contribution

Preprint Notice

This manuscript is a preprint version of an article published in *BioScience*. The final, peer-reviewed Version of Record should be used for citation.

Callaghan, C. T., Campbell, C. J., Widness, J., Goldstein, B. R., Mesaglio, T., Seltzer, C., Iwane, T., Guralnick, R. P., & Mason, B. M. (2026). Guidelines and best practices for the scientific use of global iNaturalist data. *BioScience*.

https://doi.org/10.1093/biosci/biag135

This preprint is made available in the interest of broad and equitable access to scientific knowledge. The final, peer-reviewed Version of Record represents the authoritative version of this work, and readers are encouraged to consult and cite the published article.

Rationale and scope

The majority of global biodiversity data is now contributed by participatory citizen science platforms, with at least 80% of recent contributions to the Global Biodiversity Information Facility (GBIF) generated by such platforms (Callaghan et al. 2021). iNaturalist, in particular, generates the broadest geographic and taxonomic coverage of biodiversity data (Di Cecco et al. 2021), with the volume of contributed data growing at a mean rate of 50% per year over the last decade. The volume of scientific research leveraging data from iNaturalist is also increasing at rapid rates, with a tenfold increase in publications using iNaturalist data in the last 5 years (Mason et al. 2025). This aligns with an increase in policy-usage of iNaturalist data (e.g., environmental impact statements; Callaghan et al. 2025). Despite this rapid uptake, researchers do not always have training in the best practices for handling and understanding iNaturalist data structure, knowledge of how to efficiently and appropriately access and attribute data, and guidance in how to work with the data most effectively. As a result, there is a growing need for guidance to ensure that the unprecedented growth of participatory biodiversity data from iNaturalist translates into advances in ecological understanding, synthesis, and decision-relevant insight.

As with most participatory citizen science datasets, iNaturalist data are uploaded by many different observers from different parts of the world. Attributes inherent to participatory science (e.g., variation in observation effort, identification accuracy, and metadata completeness) can impact and inform the resulting utility of the data. This includes variation in the spatiotemporal coordinates of a species (Feldman et al. 2021), the image metadata (Pernat et al. 2024a), or spatial coverage of the observations themselves (Geurts et al. 2023). Because the data are used

across disciplines—ranging from ecology to computer science to education—there are heterogeneous and diverse uses of, and caveats to, the data depending on the research questions. There is therefore a need to summarize the current best practices and guidelines for accessing the necessary data, understanding its structure, and using the data in practice.

Our aim is to provide a clear and comprehensive guide for effectively using iNaturalist data in research (Figure 1). This effort is inspired by similar guides to using eBird data in research (Johnston et al. 2021; Strimas-Mackey et al. 2025). Specifically, we first provide an overview of iNaturalist data and how it is produced and the resulting implications for biodiversity research. We then provide a "deep-dive" into the critical data components and metadata, details on the various ways to access iNaturalist data, general guidelines on steps to take when using the data in analyses, and guidance on citing and attributing the data with a focus on reproducibility. We additionally provide code examples in R and Python for some of the above, as well as provide perspectives on the future of iNaturalist data usage. While our focus is on biodiversity research, we believe these guidelines will be pertinent to many different research disciplines (e.g., social sciences, computer science) and anticipate that these guidelines will facilitate the further use of iNaturalist data for scientific research.

Foundational concepts for using iNaturalist data

iNaturalist is a global platform with a stated mission to "*connect people to nature and advance biodiversity science and conservation*". Users can upload photos or audio recordings to the iNaturalist platform via a smartphone or the website as evidence of an organism at a certain location and time. The community then helps (Callaghan et al. 2022; Campbell et al. 2023)

identify the observation to the finest possible taxonomic resolution. Although most identifiers are experienced (Campbell et al. 2023), there are no requirements to become an identifier and no extra weight is given to experts.

A key piece of information in iNaturalist data is the "quality grade" of a given record: "Research Grade" (hereafter RG), "Needs ID", or "Casual". Broadly, RG status is attained when an observation meets the site's metadata quality criteria, and has two or more suggested identifications, more than two-thirds of which agree at a species level (i.e., 2/2, 3/3, 3/4, etc.). Importantly, RG status does not only apply to species-level observations, as records identified to a taxonomic level finer than family can also become RG if further progress in identification is deemed not possible. Note that RG does not guarantee a correct identification and researchers should take care to verify observations themselves as appropriate. A recent observation accuracy experiment conducted by iNaturalist found that 1% of RG observations were incorrectly identified and 8% had uncertainty in identification (iNaturalist 2025). By contrast, "Casual" observations are those which lack key metadata; e.g., location or date, are cultivated, captive or otherwise not "wild" and are marked as such via community annotation. See glossary for more details on these key "observation quality grades".

The iNaturalist infrastructure provides mechanisms for organizing these different types of observations into structured collections. Most notably, users can create "Projects" (see Glossary), which other users may join or contribute to. Projects allow observations to be grouped, curated, and enriched around a shared research question, geographic region, taxonomic focus, or community initiative. They also allow for predetermined filters (e.g., of quality grade). Projects

allow for observations to be systematically organized, filtered, and augmented with additional metadata (e.g., via observation fields), enabling coordinated data collection and curation within an otherwise open and dynamic dataset. Observations can be included in an unlimited number of projects.

Another important component of the data submitted to iNaturalist is the licensing. Three types of content can be licensed separately on iNaturalist: (1) observation data (including metadata such as date, geographic location, species identification, etc.), (2) images, and (3) audio recordings (iNatHelp 2024). By default, when creating a new account, all three of these groups are set to CC BY-NC (Creative Commons: Attribution-NonCommercial), meaning that the content can be used with attribution for non-commercial purposes. Other licensing options that users can choose are CC0 (No Copyright), CC BY (Attribution), CC BY-NC-SA (Attribution-NonCommercial-ShareAlike), CC BY-NC-ND (Attribution-NonCommercial-NoDerivs), CC BY-ND (Attribution-NoDerivs), CC BY-SA (Attribution-ShareAlike), and no license (all rights reserved). Licenses are set independently for observations, images, and sounds, and can differ within a single observation (Figure 2). These licensing settings govern whether and to what extent researchers may legally reuse iNaturalist content. Users of iNaturalist data should take note that the license associated with an observation occurrence may differ from the license of the associated piece of evidence; for example, a space/time record of an observation may be licensed for use in research while the photo is not. If no license is set or information about the license is missing, then the data is considered “all rights reserved.”

GBIF is a major, global repository (Heberling et al. 2021) through which researchers can access biodiversity data, including iNaturalist data. iNaturalist generates a weekly export of observations eligible for inclusion in GBIF. Exported observations include only RG observations that are licensed as CC0, CC BY, or CC BY-NC (default option). Newly qualifying RG records are added, and changes to previously exported observations (i.e., updated identifications) are synced during each export cycle. This GBIF export process means that the iNaturalist data most accessible to the scientific community represents a curated, open-licensed subset of all observations submitted to the platform.

A final important concept to understand about iNaturalist data is that it is dynamic. Identifications, taxonomic definitions, metadata, and even licensing settings can change over time as community members contribute new information or users update their content. As a result, datasets retrieved for scientific research on different dates may differ, and reproducible workflows should explicitly document download dates, filters, and observation identifiers. The following sections detail how to interpret the iNaturalist data structure, access the iNaturalist data, and implement analytical workflows that are robust to the platform's inherently dynamic nature.

Critical data components and metadata for working with iNaturalist data

Taxonomy

iNaturalist's taxonomy is highly dynamic, with changes—including the addition, synonymization, and splitting of taxa—implemented daily by site curators to reflect changes in both primary (the scientific literature) and secondary (taxonomic database) sources. Broadly,

curators attempt to align iNaturalist taxonomy to global taxonomic authorities for taxa where these exist, such as Plants of the World Online (Royal Botanic Gardens, Kew 2026) for vascular plants or The World Spider Catalog (Natural History Museum Bern 2026) for spiders; a list of these external authorities is available for curators on the site. Although following designated external authorities as closely as possible is strongly encouraged, curators can 'deviate' from these authorities where appropriate to follow differing regional treatments. Regional taxonomic authorities and the primary literature are referenced for taxa where global taxonomic authorities do not exist, such as fungi or many terrestrial invertebrate groups.

iNaturalist does not maintain or publish historical versions of its taxonomic tree, so if it is to be referenced alongside an iNaturalist data download, both should be obtained at the same time. The latest version of the tree can be downloaded via the iNaturalist Taxonomy DarwinCore Archive, which is generated monthly. It is important to note that, because iNaturalist is an observational database rather than a nomenclatural database, it is not intended to represent an exhaustive source of all names for all taxa. The iNaturalist taxonomic tree differs from the one used by GBIF. When iNaturalist shares data to GBIF, GBIF uses its own taxon match service to align taxa with the GBIF Backbone Taxonomy. When name mismatches occur, GBIF may use fuzzy matching or match to a coarser taxon. The GBIF Backbone Taxonomy is updated regularly and is primarily derived from the Catalogue of Life and other authoritative sources (Bánki et al. 2026). GBIF archives previous versions of the GBIF Backbone Taxonomy, which can be downloaded at any time from GBIF Hosted Datasets.

Geoprivacy, Obscuration, and Data Accuracy

By default, observations uploaded to iNaturalist have their geoprivacy set to "open," meaning that the user permits the data to be publicly shared, and other users can freely see the coordinates and date/time of these records. However, some iNaturalist observations have their geoprivacy changed to "obscured", which can occur in two different ways. First, users may choose to manually obscure any of their records (referred to as "geoprivacy") at any time for privacy or other personal reasons. Second, some observations are automatically obscured by iNaturalist (referred to as "taxon geoprivacy"). This obscuration is implemented on an entire taxon basis (usually implemented at the level of species or infraspecies), either globally or within specified regions (e.g., a country or state), and is intended to help combat threats faced by sensitive-location taxa, that could be exacerbated by public data on the animals' occurrence such as habitat disturbance and poaching. Automatic obscuration can be applied to or removed from taxa by site curators when needed and is usually done with reference to sources such as the IUCN Red List, NatureServe, SANBI, or other national or regional institutions that assess the status of species (see list of obscured and private species on iNaturalist here).

When an observation is obscured, iNaturalist hides the original latitude and longitude, increases the public positional accuracy of the coordinates to the diagonal of a 0.2° x 0.2° grid cell (~500 km² at the equator, increasing in size towards the poles), and replaces the publicly displayed latitude and longitude with randomly assigned coordinates within that grid cell (Supplemental Text S1). On iNaturalist, when you are viewing an obscured observation, there will be a blue rectangle indicating the region of obscuration and a point which indicates the randomized location. When viewing the map of observations on the Explore page of iNaturalist, transparent

points indicate obscured coordinates. Obscured records are made available in GBIF, but the coordinates are the randomized location in the 0.2° x 0.2° grid cell. With this observation, the accuracy value will encompass the entire obscuration grid, unless the user submitted accuracy is larger, in which case this value will be used. Within GBIF, the Darwin Core field "informationWithheld" is populated with text explaining that the coordinate uncertainty has been increased. On iNaturalist, the date stamp is also obscured to only show month and year for the observation and any identifications that have been added to it; however, the exact observation date and time can be accessed via the iNaturalist API or through GBIF. "Private" observations, a third type of geoprivacy, share no geographic information publicly and these observations are not shared with GBIF. Users may choose to trust a project, when joining a project, granting project curators access to coordinates that are otherwise obscured or private. Users may additionally choose to trust an individual user with this information (iNatHelp 2023). Network nodes (localized iNaturalist experience; iNaturalist 2026a) also get access to true coordinates of observations made by users who have chosen to affiliate with them. If granted access to these hidden coordinates, take special care to avoid sharing this sensitive data in publications or data releases.

Users may also upload observations with intentionally coarse positional accuracy for the coordinates (Supplemental Text S1). This may occur because users do not remember the precise location at which they made the observation, or as an alternate method to obscure the location. For each unobscured observation on iNaturalist, a shaded opaque circle surrounds the point on the map to indicate the positional accuracy of the coordinates. The accuracy can also be viewed on the iNaturalist website in the "Details" tab of an observation's geographic information, with

the numeric value corresponding to the radius of the circle. This accuracy information is shared with GBIF and can be accessed via the iNaturalist API.

Data available on iNaturalist

The most commonly used data from iNaturalist are date and time of an observation, consensus taxonomic identification (hereafter Community Taxon), and geographic location (Mason et al. 2025). Other data available on iNaturalist includes images, audio, full identification history, species counts, annotations, observation fields, tags, and more. In this section, we will provide an overview of these data. Further details on accessing the data are provided in the next section.

Spatiotemporal data from iNaturalist are used by researchers for a myriad of reasons, including to analyze species ranges and distributions (Cull 2022; Wolfe et al. 2025), assess biodiversity trends (Beninde et al. 2023), compile species lists (Wenk et al. 2024), and more. The date and time of an observation are recorded in the time zone where the observation was recorded. While the date field is required for an observation to be deemed RG, the time field is not; therefore, some observations do not include a time stamp. Additionally, GBIF data (via the Darwin Core Archive) or the data from the iNaturalist's Export Tool provides the full ancestry of the community taxon, including kingdom, phylum, class, order, family, and genus. Geographic coordinates are reported in World Geodetic System 1984 (WGS84), with decimal latitude and longitude provided in separate fields, along with an estimate of location uncertainty in meters. Species lists, summarized species count data, and range maps are also available, and can be a useful alternative to downloading a full dataset when fine-scale detail is not required. These include summarized species counts for specified parameters and summarized counts by time

(i.e., week of the year, month, year). Furthermore, iNaturalist provides over 100,000 range maps that have been modeled from iNaturalist data (see more here; iNaturalist 2026b).

Media submitted to iNaturalist are also a rich data source that is increasingly used by researchers (Mason et al. 2025). To extract useful data from images and audio files, such as behavior or morphology, researchers may manually review media and record relevant data, or they may use advanced computational techniques such as supervised classification or computer vision models (Perez-Udell et al. 2023). Audio data is less commonly used in the literature, but it has been used to study anuran breeding phenology (Forti et al. 2022) and to test the ability of convolutional neural networks to accurately identify species based on their calls (Wu et al. 2022).

Additional data types, such as identification history and user activity trends, are available for special use cases. While most researchers will use the Community Taxon, the full identification history may provide more fine-scale information about identification trends. This data may be useful to determine identification effort towards specific projects, the top identifiers of a specific taxon, or identification biases. Identification history has been used to assess trends in iNaturalist identification (Campbell et al. 2023), which allowed specific recommendations to be made about iNaturalist data quality. Additionally, summarized user activity trends—including total observations, number of species reported, and number of identifications—can be obtained from iNaturalist. Such data can be useful for assessing engagement trends, identifying top contributors for a specific taxon, or potentially inferring observation effort. For example, researchers have used the iNaturalist API to obtain the dates on which users were active and the number of

observations made on those dates to study observer behavior and iNaturalist data quality (Dimson and Gillespie 2023).

Additional data fields that can be added to iNaturalist observations include annotations, observation fields, and tags (see Glossary). These fields are optional and can be contributed by the original observer or, in the case of annotations and observation fields, by the broader community—unless the observer has restricted who can add fields to their observations. Annotations are a limited set of fields with prescribed values that are decided by the iNaturalist administrators, with options including, but not limited to, “alive or dead”, “evidence of presence”, “life stage”, “sex”, “flowers and fruits”, and “leaves”. Some annotations are only available for certain taxa; for example, the flowers and fruits and leaves fields are only available for plants where those attributes are applicable. While most records are not annotated, the number of annotated records is increasing each year and can provide valuable insights (Figure 3). Annotations may be useful for studying phenology (e.g., timing of flowering) or providing needed human labels to train machine learning models (Dinnage et al, 2025), studying differing life stages (e.g., timing of hatchling emergence), or filtering the data (e.g., removing dead animal observations prior to examining phenology).

Observation fields are generated by the iNaturalist community and are often associated with specific projects that seek to gather additional metadata. These observation fields can include free text entry or drop-down options and can be standardized within a project, if the project administrator chooses to do so. Examples of observation fields are species interaction data (“predated by”, “visited flower of”), behavior observations (“mating”, “nesting”, “pollinating”),

and genetic mutations ("albino", "melanistic"). Because observation fields are user-generated, there may be multiple fields that record essentially the same data (e.g., "habitat", "habitat type", "general habitat", etc.). Tags are user-generated keywords that can be added to any observation by the observer. They allow observations to be searched on iNaturalist, but like observation fields, they are not standardized. Tags are less widely used than observation fields since they are only available via the website interface and can only be added by the observer. Annotations can be accessed from GBIF via columns such as sex, lifeStage, reproductiveCondition, vitality, and establishmentMeans. Observation fields and tags, however, are only available through direct export from iNaturalist or the iNaturalist API. API.

Workflows, and their tradeoffs, for accessing iNaturalist data

There are several options for researchers to access iNaturalist data, each offering their own advantages, limitations, and implications for reproducibility. Here, we summarize the primary routes for accessing observation data, as well as other forms of data and metadata. We also highlight key recommendations and decision points data users should be aware of.

Observation data

There are multiple methods of obtaining observation data (Figure 4). The most straightforward and reproducible method for obtaining iNaturalist observation data is through GBIF (https://www.gbif.org/), which aggregates all RG iNaturalist observations licensed under CC0, CC BY, or CC BY-NC (iNaturalist Research-grade Observations on GBIF). GBIF exports are structured as occurrence records and include metadata limited to standard Darwin Core fields (GBIF 2026a). Data can be retrieved via targeted queries to GBIF

(https://www.gbif.org/occurrence) through either the web-based interface or the GBIF API (GBIF 2026b). Available datasets from GBIF include simple (spatiotemporal data), Darwin Core Archive (all data that can be exported from GBIF), species list (summarized list of species), and cube (data cube based on taxonomic, temporal, or spatial dimension; see Species occurrence cubes in GBIF for more details; GBIF 2026c). If all iNaturalist data are needed, users can download that directly from the iNaturalist Research-grade Observations page on GBIF. Although, note that certain metadata fields (such as identification history) are not included in GBIF exports. Nevertheless, obtaining iNaturalist data via GBIF is most appropriate for studies that only require open-licensed RG observations and standard Darwin Core data fields.

A key advantage of downloading from GBIF is the assignment of digital object identifiers (DOIs) to individual downloaded datasets, generating persistent, citable records that facilitate transparent and reproducible research. The key idea is that the dataset DOI is reported in publications or other media using that dataset and each record in the dataset can be referenced back to the DOI. In this way, DOIs enable tracking of scientific use for individual observations and for iNaturalist-derived datasets more broadly, thereby highlighting the downstream value of these data while ensuring that both contributors and the iNaturalist platform receive appropriate credit through established data-citation pathways. The resulting Darwin Core data download from GBIF will be retained for at least six months, and cited downloads prioritized for retention. Thus, fully reproducible data downloads through GBIF are expected but not guaranteed to be retained when used in research. Also note that fixed DOIs are assigned only to specific queries and downloads from GBIF; these are permanent identifiers for a snapshot of data resulting from a particular query. GBIF also uses DOIs to track published datasets, and thus there is also an

identifier to the whole iNaturalist dataset that also is fixed, although the dataset itself changes during regular, weekly publication updates. If GBIF doesn't detect published literature that cites a download DOI within six months of export, the dataset records may be deleted. However, the metadata about the export (the DOI, the full query used, the date and time, the number of records, etc.) will be retained into perpetuity.

At the time of this publication, the GBIF download is the only method available that allows for direct tracking, reproduction, and attribution of credit to individual observations. All other methods of downloading observations (iNaturalist's Export Tool, the API, and iNaturalist Open Data) require additional steps upon publication to maintain the chain of credit, as well as reproducibility (see Attribution of Data section, below). When a small number of Research Grade observations are required, it may be tempting to avoid downloading the data or to use the iNaturalist's Export Tool or the API, but we still recommend downloading from GBIF if those observations are hosted on the site and the necessary data are available. For a very small number of observations, direct citation of specific observations is acceptable in most circumstances. However, data users should be cognizant of the status of iNaturalist observations as "living". Observations are not archived or versioned, and so features such as licensing and taxonomic identification may change over time for some observations. Observations may also be deleted by the observer at any time. If using data directly from iNaturalist, users should be sure to report both the specific date of access and retain any key information used in subsequent analyses, and publish the final dataset used in an archive such as Zenodo.

When data not available on GBIF are required—such as observations that are not RG, observation fields, observations from collection projects, or full identification history—data can be obtained directly from iNaturalist's Export Tool (https://www.inaturalist.org/observations/export). This website-based platform allows for multiple complex search filters (e.g., by taxon rank, place, date, encompassing projects, geoprivacy). It also has the capability to export additional metadata, including observation fields, tags, and observer metrics such as total number of observations, number of species recorded, and number of identifications. This approach can also be used to access true coordinates of obscured observations that other users have trusted you with, which will show up under the private_latitude and private_longitude columns of the exported data (See Supplemental Text S2). iNaturalist Export Tool is relatively user-friendly but is only able to return a maximum of 200,000 observations per request and does not generate a DOI or version snapshot.

If large datasets are needed and cannot be obtained through GBIF, the recommended approach is to create multiple export queries, each returning fewer than 200,000 results (Supplemental Text S3). The simplest way to generate batch calls is to filter by shorter time periods using the start and end date filters. For the first query, apply your desired filters in iNaturalist; at the top of the page, a URL will be generated that reflects your current query. This URL is a call to the iNaturalist API, discussed below. To streamline the process, copy this URL and adjust the date range (or any other batch filter) as needed, then paste it back into the URL query box to initiate the next batch. Keep in mind that iNaturalist only processes one export request at a time, and processing time depends on system load and dataset size. Wait for each export to finish before submitting the next request. Be aware that this approach is not practical for extremely large

datasets. For example, exporting all iNaturalist data would require nearly 1,500 separate queries as of 2025. In such cases, we recommend either: (1) consider using the API to obtain data summaries (see the "Other Forms of Data" section below); (2) consider making inferences on a subset of iNaturalist data; or (3) tailor the research objective towards data available from GBIF.

For specialty filters not available in the iNaturalist Export Tool (e.g., records updated since, without taxon ID, or unobserved by user) or when using annotations on observations not available in GBIF, the iNaturalist API must be used. This provides the most granular control over both search filters and returned fields. For example, using the updated_since parameter, the API can be used to retrieve records updated since a particular date, allowing a researcher to re-download a dataset while excluding observations unchanged since the previous download. The API also can search by observation or photo license, for example, only selecting observations where at least one photo is openly-licensed using the photo_licensed parameter. These examples barely scratch the surface of what is available. For a full list of parameters, refer to the iNaturalist API documentation (iNaturalist 2026c; see Supplemental Text S4 on how to modify request URLs for the API). Additionally, API calls by default return data in JSON format, which allows complex fields—such as media, annotations, and identifications—to be represented as nested data structures containing all relevant items and their associated metadata. Keep in mind that the API is rate-limited (a maximum of 60 requests per minute is recommended) and is not intended to support data scraping. For observation data, the API returns a maximum of 200 observations per request, so batch queries are usually necessary (see Supplemental Text S3 for more information).

Despite the added complexity, including the need to learn the iNaturalist API syntax and then unpack nested JSON outputs, the API-based workflows enable advanced data filters, provide additional data, and can be used to gather summarized data. We have provided some additional supplementary text detailing how to access data via the API, including guidance on batching, rate-limiting compliance, and handling nested JSON structures (Supplementary Text S1). Additionally, we include a set of annotated R and Python scripts that demonstrate common, reproducible workflows for accessing iNaturalist data, including retrieving observations from a single API query, batching or looping multiple API queries, obtaining summarized species count, and downloading associated media, including images or audio (Supplemental Text S5; https://github.com/Global-Ecology-Research-Group/Example-code-for-accessing-iNaturalist-data).

Data summaries

Summarized species lists (i.e., checklists of species observed in a particular location), species counts, and spatiotemporal data can also be obtained from multiple sources. Species lists can be downloaded directly from GBIF using the "Species List" dataset option, which provides a list of all species within the specified filters. This dataset includes the number of occurrences for each species, higher taxonomic information (i.e., kingdom, phylum, class, order, family, and genus), and the IUCN Red List Category when available. The iNaturalist API can also be used to obtain species lists or species counts for a specified region, time period, project, or other parameters. This is particularly useful when information not available through GBIF is needed, such as observations of all quality grades (i.e., RG, Needs ID, Casual), or when retrieving species lists for iNaturalist projects or locations that exist only within the iNaturalist platform. Species counts can be accessed through the API endpoint observations/species_counts, where users specify the

parameters such as the place, time frame, or taxonomic group of interest (Supplemental Text S5). This endpoint reports the number of "leaf taxa," the default method that iNaturalist uses to count species, in which only the distinct terminal tips of the taxonomic tree are counted to avoid double-counting at different taxonomic ranks. Leaf taxa counts include distinct taxa that have not been identified to the species level—for example, observations identified only to genus when no observations within that genus have been identified to species. As a result, leaf taxa counts are technically counts of taxa, although iNaturalist uses the term *species counts* because it is more familiar to novice users.

The iNaturalist API can also return summarized temporal data through the observations/histogram endpoint, which provides observation counts binned by a specified interval. Available intervals include month of the year (default), year, month, day, week, hour, and week of the year. These data are useful for visualizing temporal patterns in observations for specified taxa, places, projects, and other filters.

Summarized activity trend data—including total observations, number of species reported, number of identifications for each iNaturalist user, and annotation summary data—are also available from the iNaturalist API. While this information is present in the datasets exported by the iNaturalist's Export Tool, it is alongside all observations matching the search criteria, which may be more data than required to answer certain research questions. If only activity trends data are needed, the iNaturalist API can be used via the observations/observers and observations/identifiers endpoints. These endpoints provide user information, total observation counts, number of species recorded, activity count, and more (Supplemental Text S5).

Additionally, the observations/popular_field_values return annotation summary data, such as the phenology of a species over the course of a year.

Media

Image and sound files are another valuable source of data provided by iNaturalist, which allows multiple images or audio clips to be associated with a single iNaturalist record. Media can be accessed via the file URL, which can be found in the Darwin Core Archive dataset on GBIF, in the data from the iNaturalist's Export Tool, API responses, or the iNaturalist Open Data Program. For data from GBIF or exported using iNaturalist's Export Tool, only the URL for the first photo or audio recording is provided; whereas the data exported from the iNaturalist API contains URLs for all media associated with an observation record (See Supplemental Text S5). While GBIF only contains open-licensed media, the iNaturalist API can be used to download media of any license, including open licenses not in GBIF (CC-BY-ND, CC-BY-SA, CC-BY-NC-ND, or CC-BY-NC-SA) as well as all rights reserved media. Special attention should be paid to licenses when using this data. In all cases, photos and sounds must be used in compliance with the license terms, and appropriate attribution must be given (See Creative Commons 2026 for details).

Images are stored in the iNaturalist Open Dataset (iNaturalist 2026d). This dataset uses the Simple Storage Server (S3) from Amazon Web Services (AWS) to provide downloadable images without usage caps. Sound files and all rights reserved images are hosted on iNaturalist servers with a download limit of 5 GB of media per hour or 24 GB of media per day. For AWS hosted images, the URL format is https://inaturalist-open-

data.s3.amazonaws.com/photos/[photo_id]/medium.[file_extension], where photo_id is a unique identifier for each photo that can be linked back to a corresponding observation data record and the file extension is usually jpg or jpeg. iNaturalist-hosted image URLs are of the format https://static.inaturalist.org/photos/[photo_id]/medium.[file_extension] and sound URLs are https://static.inaturalist.org/sounds/[sound_id].[file_extension]. The photo URL can be altered to download different image sizes. In the example above, the photo size is set to "medium" which corresponds to an image size of up to 500 x 500 pixels. Other available size options include "square" (75 x 75 pixels), "thumb" (up to 100 x 100 pixels), "small" (up to 240 x 240 pixels), "large" (up to 1024 x 1024 pixels), and "original" (up to 2048 x 2048 pixels), although the maximum size never exceeds the original size of the image (e.g., a 400 x 400 pixel image uploaded to iNaturalist will download at that size for the medium, large, and original size options). To save downloading and processing speed and to reduce file storage space, images should be downloaded in the smallest size needed for analysis. Images and sound recordings may be downloaded manually by accessing the URL in a web browser or in an automated process using standard HTTP request methods (e.g., via the base R function download.file() or via CURL requests in Python; Supplemental Text S6 and S7). Alternatively, if the photos are to be used for computer vision training, iNaturalist provides a downloadable dataset of 675,000 licensed photos of 5,089 species (iNaturalist Challenge at FGVC: Van Horn et al. 2017) and a dataset of 3.3 million licensed iNaturalist photos of 10,000 species (iNaturalist 2021 Challenge at the FGVC8 workshop). However, these datasets were compiled in June 2017 and March 2021, respectively, and are not updated. Researchers seeking to access subsets of observation-associated media for more specific purposes can access them through scripted queries to iNaturalist's API.

Other datasets

Range maps modeled from iNaturalist data are available under the iNaturalist Open Range Map Dataset (iNaturalist 2026b), which is updated approximately monthly. The data can be downloaded in bulk by species or taxonomic group. To bulk download range maps, navigate to the Open Range Map Dataset page, then choose the taxonomic group of interest. For some taxonomic groups, there are multiple files to download (i.e., birds, insects, plants). These files are formatted as Geopackages and each contains up to 5,000 range maps. Individual range maps can be downloaded from the Taxon Pages. Individual range maps are not available for all species. To check if one is available for a species of interest, navigate to that species' about page. Under the "Map" tab, click the "Overlays" icon. If you see an option for "Expected Nearby Map," then click the link next to that option. This will take you to the geomodel predictions page for that species (e.g., iNaturalist 2026e). From here, you can click the "Download Expected Nearby Map as GeoJSON" to get the range map as a GeoJSON file. iNaturalist also provides access to a small range of basic heat maps that can be used to visualize observation density (Supplemental Text S8).

Data on other iNaturalist processes, such as identifications, can be obtained using scripted queries of the API. One method is to download all observation data using the base URL https://api.inaturalist.org/v2/observations. The json returned from this endpoint contains all identification information, including identifier information and whether each identification was categorized as leading (an identification provides more precise taxonomic information within the community taxon), improving (the first identification of an observation that the community agrees with), supporting (an identification that agrees with the community consensus), or

maverick (an identification that contradicts community agreement). If observations that were identified by a specified user or list of users are desired, then the ident_user_id parameter may be used. Alternatively, if a count of the number of identifications by a user for specified parameters (e.g., specified region, project, taxon, etc.) is desired, then the observations/identifiers endpoint can be used. This provides a json with user information, their identification count, and other activity parameters such as number of observations, number of species recorded, number of journal posts, and other information (See Supplemental Text S5).

Best practices for analyzing iNaturalist data

As illustrated above, iNaturalist is a large unstructured biodiversity dataset, created by observers who choose when, where, and what to record (Bowler et al. 2022). These data are powerful for documenting presence and spatiotemporal patterns, yet because they are unstructured, they are also heterogeneous across space, time, taxa, and observers, and much of that heterogeneity is unreported (Geurts et al. 2023). Scientific use requires understanding the data-generating process (described above), matching analytical tools to the type of information the platform can provide, and applying careful filtering, modeling, and interpretation of the inferences gleaned. In this section we outline six overarching best practices, or topics of consideration, which range from technical decisions (e.g., filtering, accuracy thresholds) to conceptual decisions (e.g., scales of inference) to maximize the scientific use of iNaturalist data.

Evaluate identification accuracy as necessary

iNaturalist species IDs are not perfectly accurate. Even for easy-to-identify species, a small fraction of RG observations may be misidentified, especially those based on evidence of

503 presence other than the organism itself (e.g., scat, footprints). The importance of explicitly
504 considering misidentifications should scale with the inferences being sought. For example, if
505 iNaturalist data are being used for the detection of a threatened taxon or documenting novel
506 range expansions, increased scrutiny should be placed on ensuring the identification of the
507 necessary observations are accurate. But in contrast, if a species distribution model is being fitted
508 with tens of thousands of observations, it is unlikely a single misidentification could cause
509 significant problems in the inference. Because iNaturalist's identifications are dynamic and
510 updated continuously, analysts should explicitly consider the potential for identification
511 uncertainties and ensure that filtering choices (e.g., species-level vs. genus-level identifications)
512 align with the biological question.
513
514 *Do not treat non-detections as absences*
515 iNaturalist records are an extremely convenient source of information on species' presence.
516 However, iNaturalist data are not as appropriate to establish the absence of a species. In areas
517 with many iNaturalist records, a lack of observations of a common and readily identified species
518 may reliably correspond to the absence of that species, but in areas of low overall iNaturalist
519 effort, even common species may go unreported. Even the areas most saturated with iNaturalist
520 data do not document all species, and many that are low density, elusive or especially diverse
521 may simply not be detected. This is a dangerous but common mistake; in a review of iNaturalist
522 data used in environmental impact reports, the authors identified 28 cases where the lack of
523 observations of a certain species was used to establish the absence of that species (Callaghan et
524 al. 2025). We recommend against the use of iNaturalist data to provide evidence of species
525 absence without the use of statistical analyses or supplemental professional sampling.

Apply question-specific filtering and quality control

Filtering must match the research objective. When assessing or analyzing location data of iNaturalist observations, careful attention must be paid to whether obscuration has been applied; otherwise false conclusions may be drawn based on the randomized coordinates rather than the true coordinates of each record. For some questions, only RG observations may be appropriate; for others, Needs ID records may contain valuable information, especially in data-limited taxa. In some cases, non-RG observations do not have significantly worse taxonomic accuracy than RG observations (Hochmair et al. 2020). Similarly, geographic accuracy, date completeness, time stamps, annotations, and media requirements should be chosen intentionally. Filtering decisions should be explicitly justified in the context of the biological question, rather than applied as generic defaults. Choices about which fields to require, which taxonomic ranks to include, or how to treat incomplete metadata directly shape the ecological inferences that can be drawn.

Match data treatment to the scale of inference

Across methodologies and taxa, there is some consensus that analysis of iNaturalist data at very fine spatial scales is especially challenging. Firstly, the geolocations associated with many iNaturalist observations have a level of uncertainty that is not conducive to very fine scale applications (~10% of all RG observations have a coordinate uncertainty of more than 500 m). Secondly, spatial variation in survey effort (where observers are not looking) becomes extreme at fine scales. Consequently, special care should be taken if modeling iNaturalist data as a function of spatial processes with a scale finer than ~1km; for many applications, an even coarser grid will be more appropriate. Further, obscured records require special attention when considering scale of analyses, as obscuration replaces true coordinates with randomized coordinates within a

0.2° × 0.2° grid cell. Researchers should determine whether these randomized coordinates are compatible with the resolution of their study, and when they are not, transparently report how these records were treated or excluded. Finally, if users filter observations by a specified level of uncertainty, they must also decide how to treat "NA" uncertainty values, which indicate missing information and not necessarily data with high uncertainty.

The appropriate approach for dealing with record-level spatial uncertainty depends on the research objective. For analyses at coarse spatial scales, such as summaries of species occurrences across large regions (e.g., a state, province, county, etc.), coarse-grain data may still be useful at higher levels of uncertainty. However, for fine-scale analyses — for example, determining habitat associations or occurrences in small urban parks — obscured locations and locations with high uncertainty require careful treatment. In many cases, especially at fine spatial resolutions, obscured coordinates should not be used as they distort the basic information and analysis results (Koo et al. 2025). Typical approaches include filtering observations by a required level of location certainty but in some cases it may be reasonable to instead add buffers equal to the uncertainty radius around points and assess whether they overlap the area of interest. Analysts should be familiar with how positional uncertainty is determined in iNaturalist (see Supplemental Text S1).

Recognize and account for the heterogeneity of sampling

Uses of iNaturalist data must contend with the fact that observation patterns arise from the behavior, preferences, and constraints of human observers. iNaturalist sampling is unstructured: observers post at their own discretion and typically do not report any metadata describing their

search effort or motivations. Consequently, iNaturalist data arise from a sampling process that is uneven in space, time, and taxonomy. Thankfully, researchers have done a great deal of work characterizing this hidden process. At the simplest level, iNaturalist sampling effort is clustered in times and places of high human activity; e.g., near human population centers and along roads and trails. Most observations are diurnal and observer activity peaks in the spring and summer (though there is substantial geographic variation in seasonal patterns). Effort also varies by species. Observers tend to preferentially report larger, rarer, and more colorful species (Callaghan et al. 2021, Stoudt et al. 2022). Individual observers also specialize in particular taxa and repeatedly observe them, meaning that search effort may be spatially clustered (Goldstein and Stoudt 2025, Di Cecco et al. 2021). This structured heterogeneity matters because it shapes detection probabilities, can bias richness estimates, and complicates comparisons across sites or taxa. Any use of iNaturalist data should consider this uneven sampling effort and ensure that the scientific question is robust to it or that these patterns are accounted for statistically.

A variety of specific modeling approaches to account for heterogeneity in iNaturalist data have been demonstrated in the literature. Many researchers retrieve and visualize iNaturalist data simply to establish places where the species occurs with certainty, then use critical thinking and interpretation of auxiliary data to weigh the roles of sampling effort and species absence in causing spatial gaps in species detections (e.g. Hobson et al. 2017, Jachowski et al. 2021). This approach is reasonable, especially for hard-to-detect species where iNaturalist data are often the best available information on a species' range and for applications where statistical inference of spatial effects is not of interest. Another common approach is to thin iNaturalist data to standardize effort before modeling occurrences as presence-only data under the assumption that

effort is constant (e.g. Dimson et al. 2023, Hill et al. 2024). For example, the analyst may retain only one iNaturalist observation per 10 km spatial grid. This approach may be appropriate for some cases, but it breaks down when sampling effort varies dramatically at the scale of spatial control. Finally, some approaches generate non-detection data to represent heterogeneous sampling (see "Estimating non-detections" below).

Make clear inferences in the context of how iNaturalist operates

iNaturalist sampling is structured by human behavior, not by standardized ecological survey design. As a result, any ecological inference must consider whether observed spatial or temporal patterns could plausibly reflect variation in observer activity rather than variation in biodiversity itself. Many observer tendencies are consistent *across* broad gradients—people often observe similar kinds of organisms in similar ways in urban, suburban, and rural settings—but the *intensity* of observation effort can vary dramatically along these same gradients. This means that some questions (e.g., coarse-scale range limits or broad biogeographic patterns) are robust to uneven effort, while others (e.g., local abundance or fine-scale habitat preferences) are far more sensitive to it. Researchers should therefore interpret results conservatively when appropriately, articulate how sampling processes may interact with their hypotheses, and transparently report the assumptions and uncertainties underlying their inferences.

Attribution and reproducibility in published research using iNaturalist data

One of the many strengths of iNaturalist data is that they are "living", meaning that almost any aspect, including their existence, could change between iterative downloads. This dynamism allows for the improvement of taxonomic identifications, addition of data quality annotations, and correction of errors, but it can also result in meaningful changes to existing iNaturalist data

at any time. This might mean that observations are deleted; or their licensing, taxonomic identification, geographic associations, or other attributes changed. It also means that analyses are not always directly reproducible. To account for this, we suggest that at a minimum, researchers explicitly document how and when the data were accessed.

As discussed above in detail, GBIF-mediated downloads generate versioned, citable datasets that preserve the state of iNaturalist records at the time of download for at least six months and often longer (see Supplementary Text S8 for further details). In contrast, data obtained through the iNaturalist API or export tool require users to manage their own versioning, including retaining downloaded data locally and documenting all filters, endpoints, and access dates, which should be reported and end products archived. These key data management aspects are essential for scientific transparency and compliance with FAIR (Findable, Accessible, Interoperable, and Reusable) principles, which are increasingly recognized for their importance by journals, funders, and researchers in ecology and evolution research (Manzano and Julier 2021; Koivisto and Mäntylä 2024). While FAIR principles should guide data, metadata and other digital objects (e.g. code) created by researchers, community science datasets also require respect for contributor decisions about licensing and CARE principles for contributor-generated data (Jennings et al. 2023).

Below, we provide a concise checklist of information that should be included when publishing analyses using iNaturalist data. Following these steps will maximize reproducibility, ensure proper attribution to iNaturalist contributors, i.e., observers and identifiers, and support the long-

term integrity of biodiversity data workflows (see Supplemental Text S9 for reporting templates).

- **Include the pathway to data access**
  - Report GBIF download DOI(s), including the exact query used.
  - Report iNaturalist API endpoints or Export Tool URLs used, including all parameters.
  - State whether multiple data sources were combined (e.g., GBIF + API).
- **Specify the exact filters applied**
  - Specify quality grades included (RG, Needs ID, Casual).
  - Specify taxonomic, geographic, temporal, annotation, and observation field filters.
  - Report any user-level filters applied (e.g., exclusion lists, project-based queries).
- **Document download and access dates**
  - Report the date(s) of data retrieval.
  - Specify any "created on" query used to define a terminal inclusion date prior to download.
- **Report licensing and usage conditions**
  - Summarize the licenses attributed to observations and associated media.
  - Describe how licensing restrictions were enforced (e.g., exclusion of "all rights reserved" content).
  - Confirm compliance with all applicable license terms.
- **State taxonomic versioning details**

- Report the date of taxonomic tree or backbone retrieval, if used.
- Describe how taxonomic changes (splits, lumps, renames) were handled.
- Specify whether iNaturalist or GBIF taxonomic backbones were used.

- **Include geoprivacy and coordinate resolution handling**
  - Describe how obscured and/or private observations were treated.
  - Specify thresholds used for coordinate uncertainty filtering.
  - Report any buffers or uncertainty-based exclusion rules applied.
- **Archive data in a long term repository**
  - Specify where processed datasets are archived (e.g., Zenodo, GitHub, Dryad).
  - Archive all observation IDs used in analyses where permissible, noting that trusted coordinates cannot be re-shared and that archived datasets may therefore be incomplete.

Emerging research frontiers for iNaturalist data

iNaturalist data have seen a tenfold increase in use in scientific literature between 2017 and 2022 (Mason et al. 2025), illustrating the vast potential of these data for use in biodiversity research and conservation. Nevertheless, as illustrated above, these data are complicated, heterogeneous along multiple dimensions, and therefore require careful treatment in analyses. The challenge moving forward is not whether iNaturalist data *can* be used, but *how best to use them* and maximize the value of these data for biodiversity and conservation research while avoiding erroneous interpretations. Here, we highlight four different forward-looking uses of iNaturalist data which we feel represent the next wave of potential research questions (Figure 5).

Secondary data

While iNaturalist data usage has primarily been focused on the spatiotemporal coordinates of observations, an increasing number of papers are beginning to extract ecological information from images (Mason et al. 2025). Such 'secondary data' (Pernat et al. 2024a), defined as additional information unintentionally captured in species records, are becoming an important and expanding research frontier. Photos from iNaturalist have been used to investigate plant-pollinator interactions (e.g., Pernat et al. 2024b), obtain color data of flowers (Luong et al. 2023), quantify morphological traits such as plumage patterns in birds (Vrettos et al. 2021) or color polymorphisms in salamanders (Hantak et al. 2022), study dietary patterns in birds (Panter et al. 2024), identify mating behavior (Ament and Brown 2025), and assess habitat composition (Bolt et al. 2022), as just a few examples. Another example within the frontier of secondary data is the field of phenology, where image timestamps allow flowering times and leaf-outs (Dinnage et al. 2025) as well as animal diel activity patterns (Idec et al. 2025) to be quantified at large geographic and taxonomic scales. This growing field of secondary data will continue to shape how iNaturalist data are used in the future. As large-scale text-to-image retrieval and multimodal machine learning systems continue to advance (Vendrow et al. 2024), we expect secondary-data applications to play a central role in the next wave of biodiversity research that uses iNaturalist.

Data integration

As researchers extract increasingly diverse ecological information from iNaturalist records, a natural next step is to combine these rich, heterogeneous data with other biodiversity datasets. An increasingly important research frontier is incorporating iNaturalist data with standardized data (e.g., structured or semi-structured survey data or specimen data from natural history

collections) to better inform ecological and biogeographic studies (e.g., Yoder et al. 2024), taxonomic works (e.g., Smith et al. 2022), or conservation and management initiatives (e.g., Ackerfield et al. 2024). In many cases, data integration maximizes the temporal and spatial breadth of the data and improves our understanding of biodiversity from local to global scales. Although the number and diversity of research papers using integrated datasets has rapidly increased over the last decade (Mason et al. 2025), there is still broad scope and exciting opportunities to further integrate datasets.

Understanding the best approaches and methods for integrating iNaturalist data and traditional data sources, and when or where these approaches are applicable, is critically important. This can be done at both the scale of entire datasets and through more tailored approaches involving individual records or curated partial datasets. A key frontier is understanding how integrative approaches can mitigate the idiosyncratic biases present in each data stream. By combining datasets with different strengths and weaknesses, researchers may be able to reduce bias, increase completeness, and better capture true ecological and biogeographic patterns (see Calvente et al. 2023 for a compelling recent example). Developing curated frameworks for combining iNaturalist data with structured surveys, museum collections, and emerging sensor-based datasets will be essential for fully realizing the scientific potential of biodiversity data integration and unlocking the full potential of iNaturalist data.

Estimating non-detections

Even with integrated datasets, many analyses require non-detection information, which has fueled a new frontier focused on estimating pseudo-absences from iNaturalist data. Because

iNaturalist is inherently a presence-only dataset, the analyses and inferences are limited, and it means estimating detectability is more difficult than presence-absence datasets. A promising frontier of current and future research is the development of approaches which derive and/or estimate non-detection information from unstructured iNaturalist data. One direction is the construction of participant-level sampling events, where observations within a place-time window are grouped into a 'visit' or 'checklist'. Rapacciuolo et al. (2021) proposed a framework for deriving participant events (akin to checklists in eBird) which could be used to demarcate everything a participant could have documented in a sampling event. For example, when participants blacklight to attract moths, inferences could be made about the species that were not uploaded during that backlighting event. Another strategy leverages trait-based or ecological similarity to infer plausible pseudo-absences (e.g., Koen et al. 2025; Sibiya et al. 2025). Species with comparable detectability, body size, microhabitat use, or phenology can be used to estimate when a focal species might reasonably have been observed but was not, thereby providing estimates of non-detections. Best practices for deriving non-detections from unstructured data remain an open challenge (Davis et al. 2023), and future research should focus on clarifying when different strategies are appropriate and how they influence downstream inference.

Trend detection

Estimating biodiversity trends, either at a community level, or a species-specific level, remains an important goal for biodiversity monitoring and research in the future. iNaturalist was only launched in 2008, and with increasing time, there is an increasing amount of data to potentially estimate trends. A key frontier for iNaturalist research, therefore, is determining if, when, and how these data can be used to infer occurrence or abundance trends. A fundamental research

question is whether the frequency of iNaturalist records in a region approximates the relative abundance of the species in that region. If a stable, interpretable relationship exists then trend detection using iNaturalist data may become statistically feasible. For instance, if a species is consistently observed hundreds of times per year and then suddenly disappears from the platform, is this a true biological signal, a change in observer behavior, or a consequence of sampling heterogeneity? Disentangling these potential relationships, how they correlate with one another, and how to spatially and temporally identify trends, are critical open questions that will inform about being able to estimate trends using iNaturalist data.

Emerging research is starting to address these questions. For example, Kaiser et al. (2025) used a synthetic-control framework to tease apart causal patterns in iNaturalist reporting, showing how a disturbance event influenced bee detections while appropriately accounting for uneven spatial sampling across regions. Additionally, researchers aiming to make use of iNaturalist data to characterize variation in species' relative abundance, and conduct inference on the drivers of said variation, have introduced a number of approaches that attempt to address the problem of spatial variation in effort through model-based inference. Some approaches assume that the reporting rate of a species on iNaturalist is a function of its relative abundance, which is reasonable in many cases but breaks down if the abundance of all target species strongly covaries or if the target species makes up a large percentage of all reports in an area of interest.

To fully disentangle biological trends from sampling, research needs to continue to (1) quantify and understand observer effort and reporting behavior, (2) integrate iNaturalist data with structured survey data, and (3) develop and refine statistical approaches suitable for unstructured

779 presence-only time series data for both occupancy and especially relative abundance.
780 Importantly, progress in trend detection will likely depend on developments across several of the
781 other frontiers outlined above. Secondary data can provide ecological covariates that help
782 explain variation in reporting rates; integrated datasets offer structured benchmarks from other
783 survey types for calibration; and pseudo-absence or non-detection inference is essential for
784 understanding detectability. Through a combination of appropriate data treatment, data
785 integration and advancement of statistical methods, we envision a future where iNaturalist data
786 contributes to trend estimation at multiple levels from species-specific to assemblage and
787 community.
788

789 **Conclusions**

790 iNaturalist should not be viewed only as a data source and this best-practices oriented review
791 focuses on it roots as an evidence-oriented community of participants. The scientific value of
792 iNaturalist is built on the enthusiasm of millions of people who choose to share their time,
793 curiosity, and observations. It is therefore essential that researchers approach iNaturalist as a
794 collaborative knowledge-making system that also supports the betterment of biodiversity
795 research. Respecting licenses and contributor intent is the minimum standard for any use.
796 Wherever possible, researchers should also work to give back to the community. Researchers can
797 credit observers and identifiers in publications, offer co-authorship to significant contributors,
798 share project results in accessible forms, acknowledge taxonomic expertise, and participate in the
799 platform by identifying, annotating, or engaging with users (Callaghan et al. 2022). Such acts of
800 reciprocal contribution strengthen trust, improve data quality, sustain long-term participation and
801 grow the community. Biodiversity research continues to shift as core theory meets new technical

802 advances and community engagement. Ensuring this shift is reciprocal is essential to ultimately
803 build the richest possible network of data, people and evidence to support biodiversity research.

Example code is available at https://github.com/Global-Ecology-Research-Group/Example-code-for-accessing-iNaturalist-data and contains scripts to:

1. **Get Observations:** Code to get observations from a single URL request. For example, all butterfly observations in Texas.
2. **Loop to Get Observations:** Code to get observations from multiple URL requests. For example, get all observations from park 1, park 2, park 3, park 4, and park 5 OR get all observations from user A, user B, user C, etc.
3. **Species Counts:** Code to get a table of species counts from one URL request. For example, get a count of observations of all angiosperm species observed in Florida.
4. **Download Images from URL:** Code to download iNaturalist images from URL path. In the example, observation data are gathered from the iNaturalist API. Then parallel processing is used to download images hosted on Amazon Web Services. Images are named according to observation ID and photo number, so they can be linked to the full observation details later.
5. **Example of complex calls:** Code to get annotations of breaking leaf buds from specified user ID.
6. **Get data from observation/observers, observations/identifiers, and observations/popular_field_values endpoints:** Code to get observer, identifier, and annotation summary data.

References

Ackerfield, J., Boom, B. M., Gandy, E., & Paradiso, L. (2024). EcoFloras elucidate insights from biodiversity data: evaluating the strengths and limitations of iNaturalist observations and herbarium specimens. *Preprints.org*. https://doi.org/10.20944/preprints202410.0017.v1

Ament, D. C., & Brown, B. V. (2025). Phoridae (Diptera) mating behaviour: identification of general patterns, exceptions and the use of sexually dimorphic structures based on vast citizen science data. *Journal of Natural History*, *59*(9-12), 739-761.

Bánki, O., Roskov, Y., Döring, M., Ower, G., Hernández Robles, D. R., Plata Corredor, C. A., Stjernegaard Jeppesen, T., Örn, A., Pape, T., Hobern, D., Garnett, S., Little, H., DeWalt, R. E., Miller, J., Orrell, T., & Aalbu, R. (2026). *Catalogue of Life* (2026-01-16 XR). Catalogue of Life Foundation. https://doi.org/10.48580/dgw64

Beninde, J., Delaney, T. W., Gonzalez, G., & Shaffer, H. B. (2023). Harnessing iNaturalist to quantify hotspots of urban biodiversity: The Los Angeles case study. *Frontiers in Ecology and Evolution*, *11*, 983371.

Bolt, M. H., Callaghan, C. T., Poore, A. G., Vergés, A., & Roberts, C. J. (2022). Using the background of fish photographs to quantify habitat composition in marine ecosystems. *Marine Ecology Progress Series*, *688*, 167-172.

Bowler, D. E., Bhandari, N., Repke, L., Beuthner, C., Callaghan, C. T., Eichenberg, D., ... & Bonn, A. (2022). Decision-making of citizen scientists when recording species observations. *Scientific Reports*, *12*(1), 11069.

Callaghan, C. T., Poore, A. G., Mesaglio, T., Moles, A. T., Nakagawa, S., Roberts, C., ... & Cornwell, W. K. (2021). Three frontiers for the future of biodiversity research using citizen science data. *BioScience*, *71*(1), 55-63.

Callaghan, C. T., Mesaglio, T., Ascher, J. S., Brooks, T. M., Cabras, A. A., Chandler, M., ... & Young, A. N. (2022). The benefits of contributing to the citizen science platform iNaturalist as an identifier. *PLoS Biology*, *20*(11), e3001843.

Callaghan, C. T., Winnebald, C., Smith, B., Mason, B. M., & López-Hoffman, L. (2025). Citizen science as a valuable tool for environmental review. *Frontiers in Ecology and the Environment*, *23*(1), e2808.

Calvente, A., da Silva, A. P. A., Edler, D., Carvalho, F. A., Fantinati, M. R., Zizka, A., & Antonelli, A. (2023). Spiny but photogenic: Amateur sightings complement herbarium specimens to reveal the bioregions of cacti. *American Journal of Botany*, *110*(10), e16235.

Campbell, C. J., Barve, V., Belitz, M. W., Doby, J. R., White, E., Seltzer, C., ... & Guralnick, R. (2023). Identifying the identifiers: How iNaturalist facilitates collaborative, research-relevant data generation and why it matters for biodiversity science. *BioScience*, *73*(7), 533-541.

Creative Commons. (2026). *Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)*. Retrieved March 11, 2026, from https://creativecommons.org/licenses/by-nc-sa/4.0/deed.en#ref-appropriate-credit

Cull, B. (2022). Monitoring trends in distribution and seasonality of medically important ticks in North America using online crowdsourced records from iNaturalist. *Insects*, *13*(5), 404.

Davis, C. L., Guralnick, R. P., & Zipkin, E. F. (2023). Challenges and opportunities for using natural history collections to estimate insect population trends. *Journal of Animal Ecology*, *92*(2), 237-249.

Di Cecco, G. J., Barve, V., Belitz, M. W., Stucky, B. J., Guralnick, R. P., & Hurlbert, A. H. (2021). Observing the observers: How participants contribute data to iNaturalist and implications for biodiversity science. *BioScience*, *71*(11), 1179-1188.

Dimson, M., Berio Fortini, L., Tingley, M. W., & Gillespie, T. W. (2023). Citizen science can complement professional invasive plant surveys and improve estimates of suitable habitat. *Diversity and Distributions*, *29*(9), 1141-1156.

Dimson, M., & Gillespie, T. W. (2023). Who, where, when: Observer behavior influences spatial and temporal patterns of iNaturalist participation. *Applied Geography*, *153*, 102916.

Dinnage, R., Grady, E., Neal, N., Deck, J., Denny, E., Walls, R., ... & Li, D. (2025). PhenoVision: A framework for automating and delivering research-ready plant phenology data from field images. *Methods in Ecology and Evolution*, *16*(8), 1763-1780.

Feldman, M. J., Imbeau, L., Marchand, P., Mazerolle, M. J., Darveau, M., & Fenton, N. J. (2021). Trends and gaps in the use of citizen science derived data as input for species distribution models: A quantitative review. *PloS One*, *16*(3), e0234587.

Forti, L. R., Hepp, F., de Souza, J. M., Protazio, A., & Szabo, J. K. (2022). Climate drives anuran breeding phenology in a continental perspective as revealed by citizen-collected data. *Diversity and Distributions*, *28*(10), 2094-2109.

GBIF. (2026a). *Darwin Core*. Global Biodiversity Information Facility. Retrieved March 11, 2026, from https://www.gbif.org/darwin-core

GBIF. (2026b). *Occurrence (OpenAPI v1) — GBIF API technical documentation*. Global Biodiversity Information Facility. Retrieved March 11, 2026, from https://techdocs.gbif.org/en/openapi/v1/occurrence

GBIF. (2026c). *Species occurrence cubes*. Global Biodiversity Information Facility. Retrieved March 11, 2026, from https://www.gbif.org/occurrence-cubes

Geurts, E. M., Reynolds, J. D., & Starzomski, B. M. (2023). Turning observations into biodiversity data: Broadscale spatial biases in community science. *Ecosphere*, *14*(6), e4582.

Goldstein, B. R., & Stoudt, S. (2025). Evidence of novelty and specialization behavior in participatory science reporting. *Oikos*, *2025*(6), e10938.

Hantak, M. M., Guralnick, R. P., Zare, A., & Stucky, B. J. (2022). Computer vision for assessing species color pattern variation from web-based community science images. *Iscience*, *25*(8).

Heberling, J. M., Miller, J. T., Noesgaard, D., Weingart, S. B., & Schigel, D. (2021). Data integration enables global biodiversity synthesis. *Proceedings of the National Academy of Sciences*, *118*(6), e2018093118.

Hill, J., Teal, E., Cross, C. L., Sanchez, Z., Webber, M. M., Oxborough, R. M., & Messenger, L. A. (2024). Using iNaturalist presence data to produce suitability maps for Triatoma protracta, T. rubida and T. recurva in the American Southwest, Texas and northern Mexico, to identify potential transmission zones of Chagas disease. *Scientific Reports*, *14*(1), 26879.

Hobson, E. A., Smith-Vidaurre, G., & Salinas-Melgoza, A. (2017). History of nonnative monk parakeets in Mexico. *PLoS One, 12*(9), e0184771.

Hochmair, H. H., Scheffrahn, R. H., Basille, M., & Boone, M. (2020). Evaluating the data quality of iNaturalist termite records. *Plos One*, *15*(5), e0226534.

Idec, J., Campbell, C. J., Belitz, M., Akshay, V. A., & Guralnick, R. (2025). Using citizen science data to estimate trait and climate drivers of daily activity patterns in temperate butterflies. *PLoS One*, *20*(11), e0335856.

iNatHelp. (2023). *How can I access the hidden coordinates of sensitive species?* Retrieved March 11, 2026, from https://help.inaturalist.org/en/support/solutions/articles/151000170343-how-can-i-access-the-hidden-coordinates-of-sensitive-species-

iNatHelp. (2024). *What are licenses? How can I update the licenses on my content?* https://help.inaturalist.org/en/support/solutions/articles/151000175695-what-are-licenses-how-can-i-update-the-licenses-on-my-content-

iNaturalist. (2025). *Observation accuracy experiment #9: Methods*. https://www.inaturalist.org/observation_accuracy_experiments/9?tab=methods

iNaturalist. (2026a). *iNaturalist Network*. Retrieved March 11, 2026, from https://www.inaturalist.org/sites/network

iNaturalist. (2026b). *iNaturalist Open Range Map Dataset 2.28.* Retrieved March 11, 2026, from https://www.inaturalist.org.

iNaturalist. (2026c). *iNaturalist API v2 documentation*. Retrieved March 11, 2026, from https://api.inaturalist.org/v2/docs/

iNaturalist. (2026d). *iNaturalist open data* [GitHub repository]. GitHub. Retrieved March 11, 2026, from https://github.com/inaturalist/inaturalist-open-data/tree/main

iNaturalist. (2026e). *iNaturalist Geomodel explanation for taxon 26159*. Retrieved March 11, 2026, from https://www.inaturalist.org/geo_model/26159/explain

Jachowski, D., Kays, R., Butler, A., Hoylman, A. M., & Gompper, M. E. (2021). Tracking the decline of weasels in North America. *PLoS One*, *16*(7), e0254387.

Jennings, L., Anderson, T., Martinez, A., Sterling, R., Chavez, D. D., Garba, I., ... & Carroll, S. R. (2023). Applying the 'CARE Principles for Indigenous Data Governance' to ecology and biodiversity research. *Nature Ecology & Evolution*, *7*(10), 1547-1551.

Johnston, A., Hochachka, W. M., Strimas-Mackey, M. E., Ruiz Gutierrez, V., Robinson, O. J., Miller, E. T., ... & Fink, D. (2021). Analytical guidelines to increase the value of community science data: An example using eBird data to estimate species distributions. *Diversity and Distributions*, *27*(7), 1265-1277.

Kaiser, A., Dee, L., & Resasco, J. (2025). Synthetic control methods enable stronger causal inference using participatory science data in cities. *Research Square.* https://doi.org/10.21203/rs.3.rs-6213868/v1

Koen, E. L., Ellington, E. H., Barichivich, W. J., Kochman, H., Enge, K. M., & Walls, S. C. (2025). Persistence of a declining anuran species across its distribution. *PLoS One*, *20*(9), e0332991.

Koivisto, E., & Mäntylä, E. (2024). Are Open Science instructions targeted to ecologists and evolutionary biologists sufficient? A literature review of guidelines and journal data policies. *Ecology and Evolution*, *14*(7), e11698.

Koo, K. S., Lee, K. H., Lee, D., & Jang, Y. (2025). Impact of obscured data on species distribution models. *Conservation Biology*, *39*(5), e70050.

Luong, Y., Gasca-Herrera, A., Misiewicz, T. M., & Carter, B. E. (2023). A pipeline for the rapid collection of color data from photographs. *Applications in Plant Sciences*, *11*(5), e11546.

Manzano, S., & Julier, A. (2021). How FAIR are plant sciences in the twenty-first century? The pressing need for reproducibility in plant ecology and evolution. *Proceedings of the Royal Society B: Biological Sciences*, *288*(1944).

Mason, B. M., Mesaglio, T., Barratt Heitmann, J., Chandler, M., Chowdhury, S., Gorta, S. B. Z., ... & Callaghan, C. T. (2025). iNaturalist accelerates biodiversity research. *BioScience*, *75*(11), 953-965.

Natural History Museum Bern. (2026). *World Spider Catalog* (Version 27). https://wsc.nmbe.ch/

Panter, C. T., Naude, V. N., Barbar, F., & Amar, A. (2024). Continental scale dietary patterns in a New World raptor using web-sourced photographs. *Plos One*, *19*(7), e0304740.

Perez-Udell, R. A., Udell, A. T., & Chang, S. M. (2023). An automated pipeline for supervised classification of petal color from citizen science photographs. *Applications in Plant Sciences*, *11*(1), e11505.

Pernat, N., Canavan, S., Golivets, M., Hillaert, J., Itescu, Y., Jarić, I., ... & Groom, Q. (2024a). Overcoming biodiversity blindness: Secondary data in primary citizen science observations. *Ecological Solutions and Evidence*, *5*(1), e12295.

Pernat, N., Memedemin, D., August, T., Preda, C., Reyserhove, L., Schirmel, J., & Groom, Q. (2024b). Extracting secondary data from citizen science images reveals host flower preferences of the Mexican grass-carrying wasp Isodontia mexicana in its native and introduced ranges. *Ecology and Evolution*, *14*(6), e11537.

Rapacciuolo, G., Young, A., & Johnson, R. (2021). Deriving indicators of biodiversity change from unstructured community-contributed data. *Oikos*, *130*(8), 1225-1239.

Royal Botanic Gardens, Kew. (2026). *Plants of the World Online*. https://powo.science.kew.org/

Sibiya, M. D., Goldstein, B. R., Pacifici, K., Kays, R., Cove, M. V., Jensen, A. J., ... & Baruzzi, C. (2025). Assessing spatial ecological patterns of wild turkey relative abundance across North America. *Wildlife Society Bulletin*, *49*, e1633.

Smith, L. T., Magdalena, C., Przelomska, N. A., Pérez-Escobar, O. A., Melgar-Gómez, D. G., Beck, S., ... & Monro, A. K. (2022). Revised species delimitation in the giant water lily genus Victoria (Nymphaeaceae) confirms a new species and has implications for its conservation. *Frontiers in Plant Science*, *13*, 883151.

Strimas-Mackey, M., Hochachka, W. M., Ruiz-Gutierrez, V., Robinson, O. J., Miller, E. T., Auer, T., Kelling, S., Fink, D., & Johnston, A. (2023). *Best practices for using eBird data* (Version 2.0). Cornell Lab of Ornithology. https://ebird.github.io/ebird-best-practices/

Stoudt, S., Goldstein, B. R., & de Valpine, P. (2022). Identifying engaging bird species and traits with community science observations. *Proceedings of the National Academy of Sciences*, *119*(16), e2110156119.

Van Horn, G., Mac Aodha, O., & Kan, W. (2017). *iNaturalist Challenge at FGVC 2017* [Kaggle competition]. Kaggle. https://www.kaggle.com/competitions/inaturalist-challenge-at-fgvc-2017

Vendrow, E., Pantazis, O., Shepard, A., Brostow, G., Jones, K. E., Mac Aodha, O., ... & Van Horn, G. (2024). INQUIRE: A natural world text-to-image retrieval benchmark. *Advances in Neural Information Processing Systems*, *37*, 126500-126514.

Vrettos, M., Reynolds, C., & Amar, A. (2021). Malar stripe size and prominence in peregrine falcons vary positively with solar radiation: support for the solar glare hypothesis. *Biology Letters*, *17*(6).

1001 Wenk, E., Mesaglio, T., Keith, D., & Cornwell, W. (2024). Curating protected area-level species
1002 lists in an era of diverse and dynamic data sources. *Ecological Informatics*, *84*, 102921.
1003 Wolfe, B. W., Champion, C., Gaston, T., Brown, J., Coleman, M. A., Davis, T. R., ... & Pecl, G.
1004 T. (2025). Continental-scale assessment of climate-driven marine species range
1005 extensions using a decade of citizen science data. *Diversity and Distributions*, *31*(4),
1006 e70022.
1007 Wu, S. H., Chang, H. W., Lin, R. S., & Tuanmu, M. N. (2022). SILIC: A cross database
1008 framework for automatically extracting robust biodiversity information from soundscape
1009 recordings based on object detection and a tiny training dataset. *Ecological*
1010 *Informatics*, *68*, 101534.
1011 Yoder, J. B., Andrade, A. K., DeFalco, L. A., Esque, T. C., Carlson, C. J., Shryock, D. F., ... &
1012 Smith, C. I. (2024). Reconstructing 120 years of climate change impacts on Joshua tree
1013 flowering. *Ecology Letters*, *27*(8), e14478.

Figures

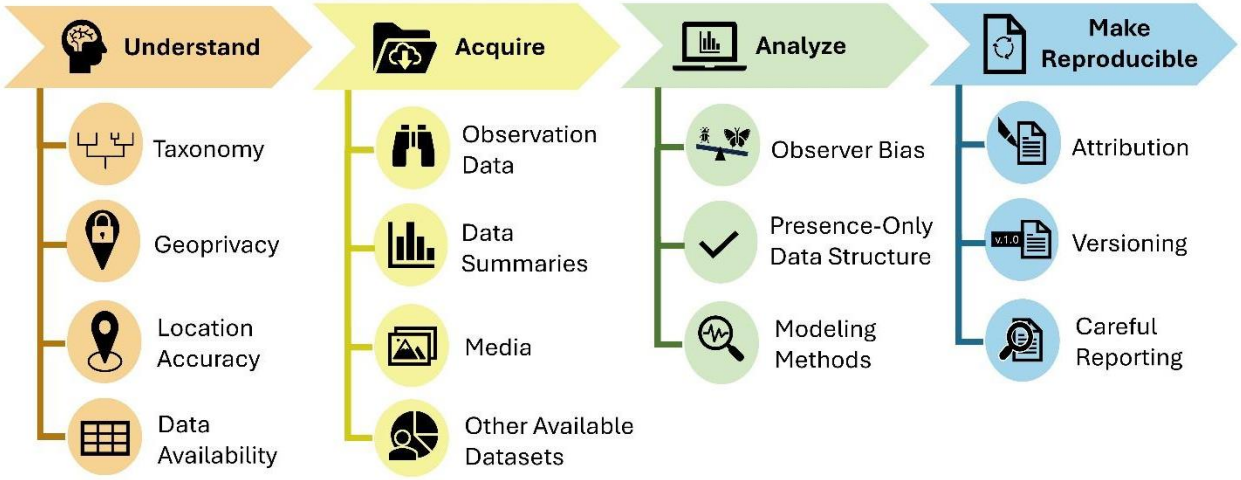


Figure 1. Using iNaturalist data effectively in scientific research requires understanding at multiple levels how to interpret the data, acquire the data, analyze the data, and ensure reproducibility.

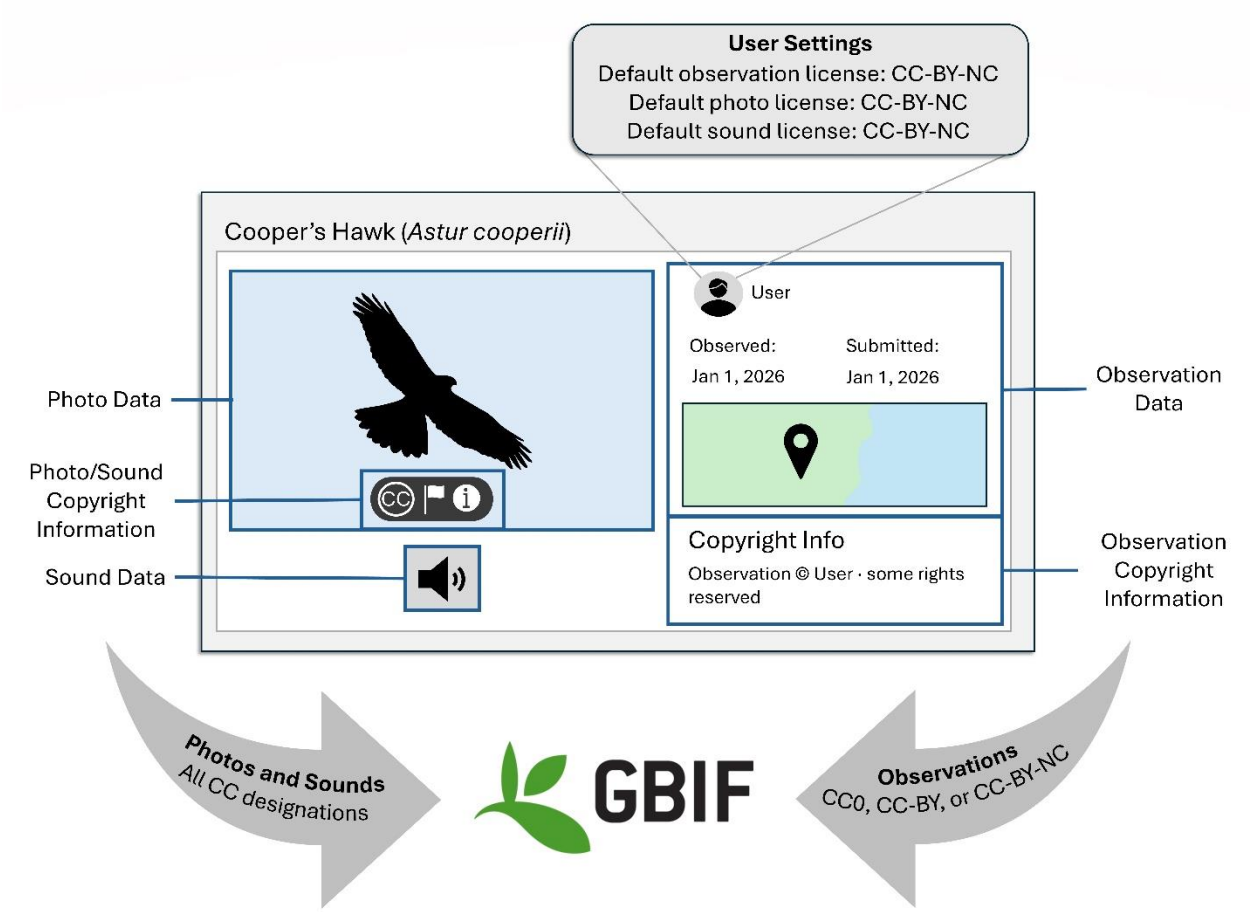


Figure 2. Diagram illustrating the three independently copyrighted data sources associated with iNaturalist records: photo data, sound data, and observation data. The diagram also shows where copyright and licensing information can be found for each data type. Photos and sounds with any Creative Commons (CC) designation are shared with GBIF for scientific use, whereas observations must be designated as CC0, CC-BY, or CC-BY-NC (the default) to be shared with GBIF. Researchers are responsible for ensuring that data are used in accordance with the applicable licenses.

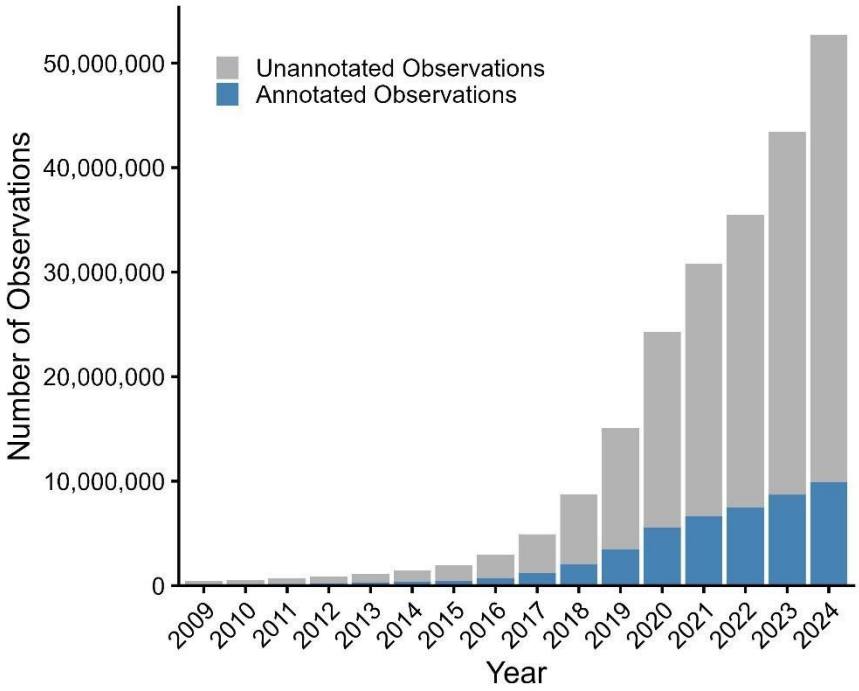
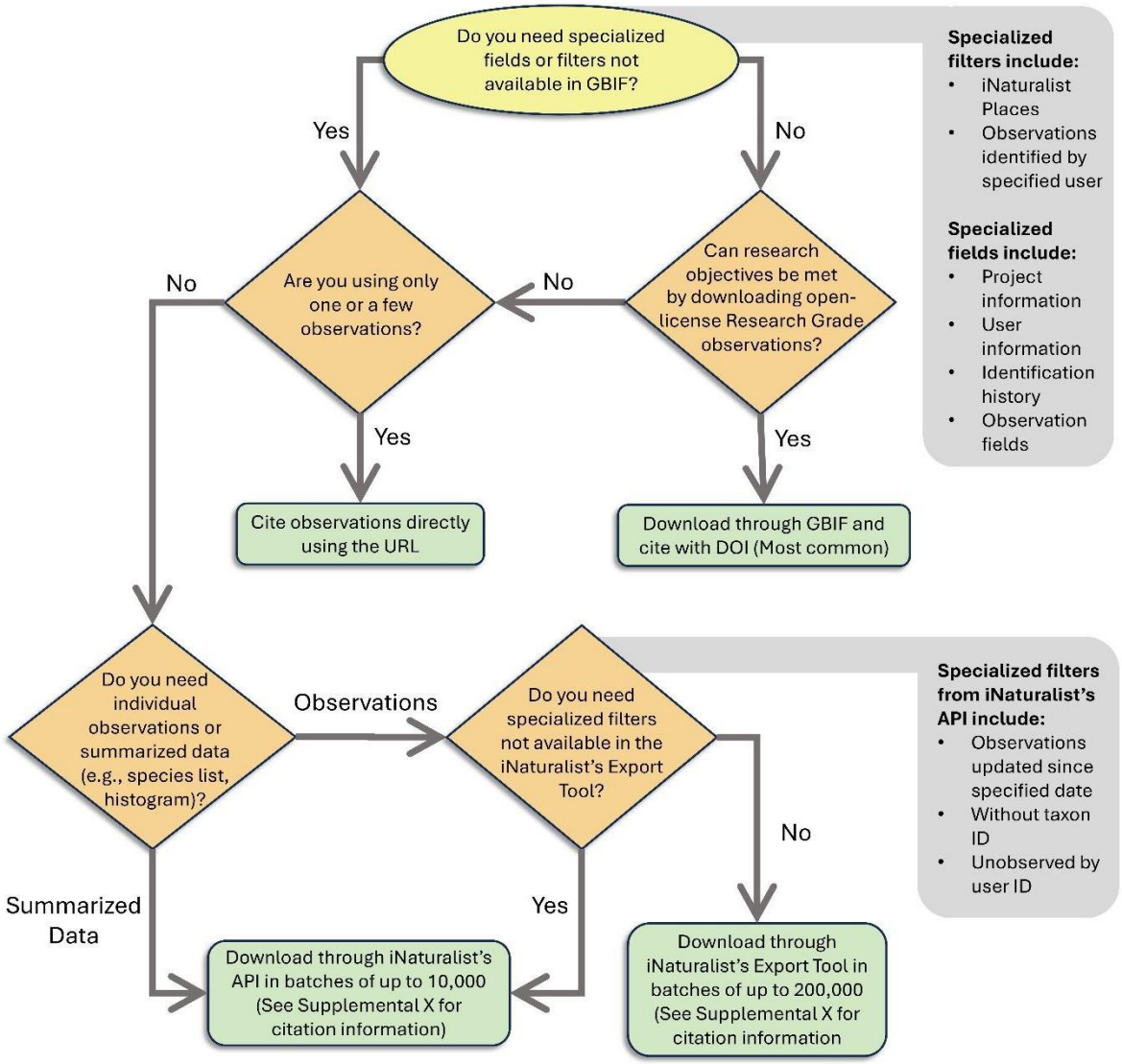


Figure 3. Number of observations with at least one annotation versus unannotated iNaturalist observations by year. Over 47 million observations made between 2009 and 2024 have at least one annotation.



1032

1033 **Figure 4.** Flow chart to determine the best method of downloading iNaturalist observation data.

1034

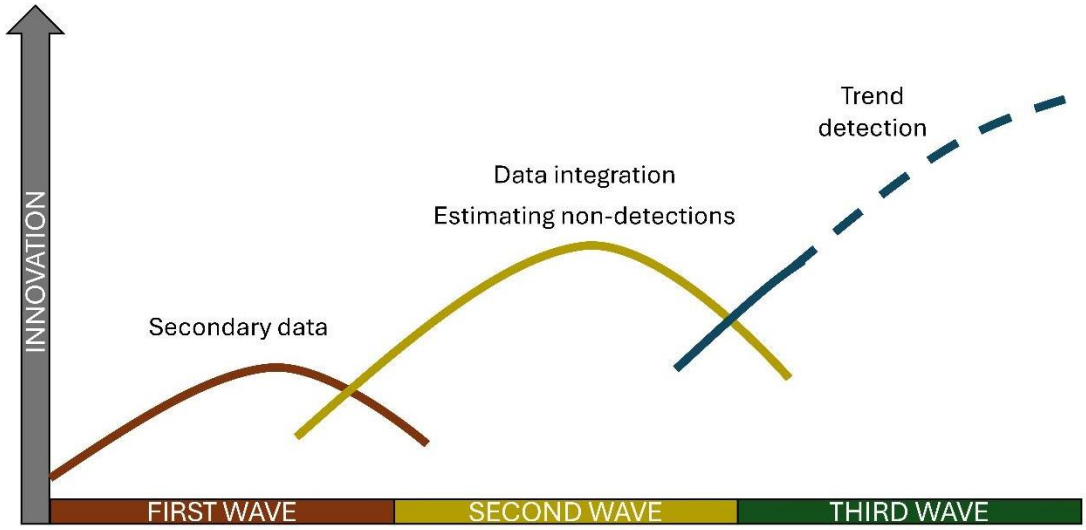


Figure 5. Emerging waves of research frontiers for iNaturalist data. Secondary data refers to additional information unintentionally captured in species records, such as details contained in photographs. Data integration refers to combining iNaturalist data with other data sources, such as professionally collected datasets. Estimating pseudo-absences involves using presence-only observations from iNaturalist to model predicted absences. Trend detection involves using iNaturalist data to monitor population trends over time.

Tables

Table 1. Glossary of key terms used throughout this manuscript. We further refer the reader to this link with further terms relevant for iNaturalist: https://www.inaturalist.org/terminology.

Term	Definition	Notes / Implications for Research Use
Observation	A record submitted to iNaturalist. Records that include media (photo/audio) and associated metadata (date, location) are deemed 'verifiable'.	Observations do not imply systematic sampling; absence of an observation of a species ≠ true absence of a species.
Computer Vision	A machine-learning model that generates initial species identification suggestions from the user's uploaded photos.	While initial identifications from this model can increase the likelihood that an observation reaches Research Grade, they should not be relied upon until they receive community support.
Identification (ID)	A taxonomic label suggested by a user or computer vision for an observation. Multiple IDs can be attached to a single observation.	Each ID is weighted equally, regardless of who suggested it.
Community Taxon	The taxonomic consensus resulting from the community identification process based on a >2/3 majority agreement rule.	IDs can **change over time** as more users review; the *community taxon* reflects consensus at a given point in time.
Research Grade (RG)	A status assigned to observations that meet: (1) verifiable metadata (date, location, media), (2) community taxon finer than family, and (3) ≥2/3 ID agreement.	Only RG observations with appropriate licenses are exported to GBIF.
Needs ID	Observations that are verifiable but have not yet reached community consensus at a fine taxonomic level.	Often valuable in data-limited taxa; may include high-quality records missing expert review or only identified by a single expert.
Casual	Observations lacking key metadata (e.g., no media or location), depicting captive/cultivated organisms, or having low quality metadata (e.g., date is not accurate).	Usually excluded from biodiversity analyses unless the research question requires them.

License	Usage rights chosen by observers for their observation data, images, and audio files (e.g., CC0, CC BY, CC BY-NC).	License type determines how data can be legally downloaded, shared, and exported to GBIF.
Geoprivacy	Setting determining how precisely location data are displayed: *open*, *obscured*, or *private*.	Obscured locations are randomized within a 0.2° × 0.2° cell; private observations hide locations entirely and are **not** exported to GBIF.
Obscured Observation	An observation whose coordinates are intentionally randomized to protect sensitive species or user privacy.	Suitable for coarse-scale analyses; **do not** use for fine-scale habitat inference without buffers or without obtaining true coordinates.
GBIF Export	The subset of iNaturalist observations and metadata shared automatically with GBIF under compatible licensing.	Best for reproducibility because GBIF provides a **DOI** reference for citation.
iNaturalist's Export Tool	A web-based tool on iNaturalist that allows users to search for and download observation data.	This method offers more filtering options and exports more data than GBIF, but queries are limited to 200,000 observations at a time, and DOI references are not provided.
API Access	A programmatic method of requesting data directly from iNaturalist endpoints.	Allows for full control and access to *all* metadata, but requires rate-limit awareness and documentation of query parameters.
iNaturalist Open Data	Dataset of iNaturalist images, updated in real time, and their associated metadata, updated monthly. Hosted on Amazon Web Services.	Photo URLs formatted as s3://inaturalist-open-data/photos or https://inaturalist-open-data.s3.amazonaws.com/photos are hosted on Amazon Web Services and can be downloaded without placing any load on the iNaturalist servers.
Sampling Effort	The latent and uneven distribution of where, when, and what observers choose to record.	Must be modeled or accounted for to avoid misinterpreting sampling bias as ecological signal.

Presence-only Data	Data describing where a species has been observed, without information on where it was searched for but *not* found.	iNaturalist data are inherently presence-only; absences must be inferred statistically, never assumed.
Reporting Rate	The fraction of all observations in a given unit that belong to a particular species.	Can approximate relative abundance if sampling effort is adequately modeled.
Taxonomic Backbone	The taxonomic reference system used by iNaturalist to organize and reconcile species names.	Changes over time; analysts should record the *date* when data were downloaded.
Leaf Taxa	iNaturalist defines species counts based on the distinct terminal tips of its taxonomic tree to the species-level, which prevents double-counting. For more detailed information, see How iNaturalist Counts Taxa.	The iNaturalist website and species count API call return a count of leaf taxa.

1046

Supplementary Text: Guidelines and best practices for the scientific use of global iNaturalist data

Corey T. Callaghan[1,*], CJ Campbell[2,3], Jane Widness[4], Benjamin R. Goldstein[5], Thomas Mesaglio[6], Carrie Seltzer[7], Tony Iwane[7], Robert P. Guralnick[8], Brittany M. Mason[1,*]

Table of Contents

S1: How positional accuracy is determined in iNaturalist 2
S2: How to access hidden coordinates.....4
S3: How to batch query data.....6
S4: Introduction to modifying request URLs for the API.....7
S5: Code examples.....9
S6: Accessing data via the API.....11
S7: List of resources for accessing iNaturalist data via the API.....14
S8: How to obtain iNaturalist heat maps 20
S9: Suggested reporting template for reproducible use of iNaturalist data 22

S1: How positional accuracy is determined in iNaturalist

For any given iNaturalist observation, the positional accuracy of its coordinates may have been applied in one of several ways, including automated and manual approaches.

1) User A takes a photograph with their smartphone, with location services turned on. A positional accuracy value is embedded in the metadata of the photograph, and when the user uploads this photograph as an observation to iNaturalist the value is automatically applied to the coordinates/location. In this scenario, this value will often be relatively small and accurate, especially in areas with reliable satellite access, however, large errors may occur in some situations, such as where GPS signals are blocked or obstructed (e.g., within a cave or amongst many tall buildings) or if there is a poor Wi-Fi or data connection. This may also happen when photos are taken immediately after opening a phone that hasn't been used recently in the current location, such as when someone is hiking down a trail (in this case, the point may end up centered by the nearest cell tower).

2) User B takes a photograph with their camera which does not have geotagging enabled. When they upload the photograph to iNaturalist, it has no embedded location information at all. The user manually enters coordinates by navigating on the map to a relatively precise location where they took the photograph, and then also manually assigns a positional accuracy value based on their judgement. Here, the positional accuracy is influenced by factors such as the reliability of the user's memory or whether they took notes about the photographic location.

3) User C takes a photograph with their camera which does not have geotagging enabled, however, they have a GPS device running during their walk. When they get home, the user downloads the GPS data as a GPX file, and then uses software on their computer to embed

location data in their photograph by matching the timestamps between the GPX file and the photograph. The coordinates are embedded in the photograph, however, a positional accuracy is not. When uploading the photograph to iNaturalist, the user follows a similar process to User B, i.e., manually assigning a positional accuracy value, but in this case based on their judgement of how accurate their GPS device was.

4) User D takes a photograph with their camera which does not have geotagging enabled. When they upload the photograph to iNaturalist, it has no embedded location information at all. The user searches for a location using the Google Maps search interface, such as a city or town name (e.g., 'New York') or the name of a reserve or park (e.g., 'Royal National Park'). Google Maps then automatically assigns a pre-determined set of coordinates representing that locality (usually a central point) and a pre-determined positional accuracy value that encompasses the entire locality. Here, the positional accuracy value is generally determined by the size of the feature searched for, i.e., a small rural town will likely have a smaller fixed positional accuracy, as determined by Google Maps, than an enormous national park or nature reserve.

5) User E takes a photograph with their smartphone, with location services turned on. A positional accuracy value is embedded in the metadata of the photograph, and when the user uploads this photograph as an observation to iNaturalist the value is automatically applied to the coordinates/location. In this scenario, the organism that was photographed is an endangered species threatened by poaching, and so its location data is automatically obscured upon upload. The original positional accuracy is still stored within the record and can be viewed by the observer or other users that they 'trust' with their hidden coordinates, but the public positional accuracy is automatically changed to the diagonal of a 0.2° x 0.2° grid cell.

S2: How to access hidden coordinates

Researchers can access the hidden coordinates of sensitive species when they have been trusted with those coordinates, even if the observations are automatically obscured. This access occurs when a user either trusts a project that you curate or trusts you directly through their account settings. In both cases, you will be able to view unobscured locations of the user's observations directly on iNaturalist.

If you are a researcher who needs access to sensitive species locations, the recommended approach is to create a project and allow users to trust the project with hidden coordinates. You can then invite users to join the project and have them grant permission for researchers to access their hidden coordinates. Keep in mind that you are asking users to share sensitive information and you should follow the Best practices when asking a user for a hidden location.

If you are a project curator and users have trusted the project with their coordinates, you can download those coordinates using the iNaturalist Export Tool while logged into your account. In the exported dataset, the columns private_latitude and private_longitude contain the trusted coordinates.

S3: How to batch query data

This reference explains how to run multiple queries using the iNaturalist Export Tool and the iNaturalist API when the amount of data needed exceeds the request limits. These request limits are 200,000 observations for the iNaturalist Export Tool and 10,000 requests per 24 hours (which may include up to 200 observations per request) for the iNaturalist API. The processes for batch querying each dataset rely on the same fundamental steps, but they are executed differently. We first describe the best practices for batch queries, then outline the procedures for (1) the iNaturalist Export Tool and (2) the iNaturalist API.

Best practices for batch queries

Batch processing data can be time-consuming and places additional strain on servers, so researchers should carefully consider whether it is necessary before running batch queries. If the required data are available through GBIF, they should be downloaded directly from GBIF (see iNaturalist Research-grade Observations). If iNaturalist exported data are needed, all possible filters should be applied before downloading. Researchers should also consider whether summarized datasets can be used instead of full observation records. For instance, if only species counts are needed, the observations/species_counts API endpoint should be used rather than downloading all observation data. Because the iNaturalist Export Tool allows larger batches to be downloaded (200,000 at once compared to the smaller batches required from the iNaturalist API) the iNaturalist Export Tool should be used when possible. Keep in mind that batch querying should not be used to scrape data. In addition to being time-intensive, excessive batch queries place unnecessary strain on the servers. This approach should be reserved for medium-sized datasets and should always follow the API Recommended Practices. If a particular research

question requires large amounts of iNaturalist export data, contact help@inaturalist.org for support.

When querying the dataset, you should maximize the amount of data downloaded per request. When using the iNaturalist Export Tool, downloads should be close to, but not exceed, 200,000 records, while API returns 200 observations per request. Requesting a small number of large datasets places less strain on the server than making many small requests.

Query data on iNaturalist Export Tool

One benefit of the iNaturalist Export Tool is that it provides a preview of the data, including the number of observations returned by the query (Figure S2). To begin, enter your filtering criteria in the "Create a Query" section, then scroll down to the "Preview" section. Here, you will see a sample of the data returned by the query as well as the number of observations. If your query returns more than 200,000 results, a warning will appear stating: "Your export will be too big. Change your filters until there are fewer than 200,000 observations."

There are multiple ways to filter the data into subsets. One of the common methods is to filter by date using the start and end date fields. If start date is entered with no end date, all data after the start date will be returned. Conversely, if an end date is entered without a start date, all the data

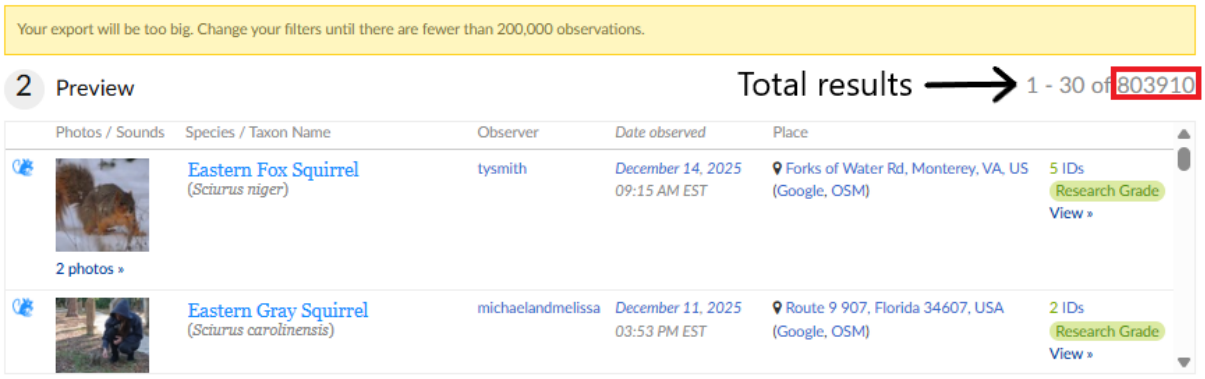


Figure S2. Example of the data preview in the iNaturalist Export Tool. The red box highlights the total number of results returned by the query. Because the total exceeds 200,000, there is a warning stating that the export is too large.

before that date will be returned. Date ranges can be adjusted iteratively until the query returns close to, but fewer than, 200,000 records. This approach is often effective because it allows precise control over the number of records returned.

Another option is to filter by taxon or place, depending on the needed dataset. Taxa can be filtered at any taxonomic level, from broad groups such as kingdoms (i.e., Plants, Animals, Fungi, etc) or at more specific levels such as family. Place filters can be applied at multiple spatial scales, including country, state, city, place ID, and others. However, this strategy is only effective if each grouping has large amounts of data that are less than 200,000. If that is not the case, then the date method is likely the best option.

Filters can be applied using the fields provided by iNaturalist or by editing the query string shown at the top of the "Create a Query" section (Figure S3). When this string is edited, the

corresponding fields below are automatically populated. This approach may be more preferable when running multiple queries with a similar structure. In such cases, copy and editing the string for subsequent queries can save time. Saving these parameter strings also helps document how the data were batch queried for future reference.

Regardless of the method used, it is good practice to first define the full set of data needed and note the total number of observations. Subsequent queries should then be designed so that their parameters collectively account for all observations without overlap.

When using the iNaturalist Export Tool, keep in mind that only one request can be submitted to the iNaturalist data server at a time. The time that requests take to finish depends on available server resources. When the export is complete, the "Finished" column will populate and a download option will appear on the left of the request. At this point, the data should be

downloaded and saved with a unique filename indicating that it is part of a batch query (eg. iNat_Export_Batch_1). Once the request is finished, the next query may be submitted.

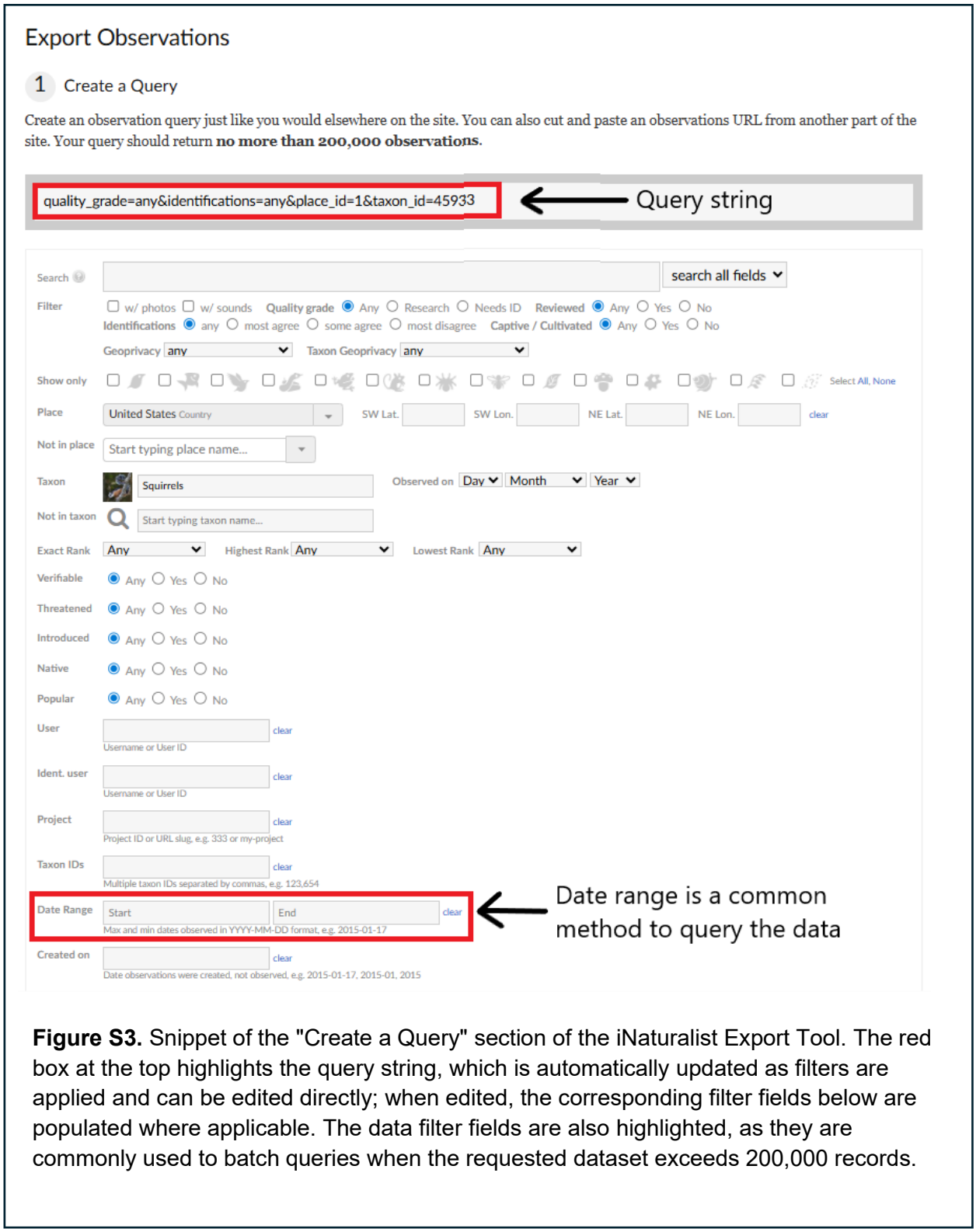


Figure S3. Snippet of the "Create a Query" section of the iNaturalist Export Tool. The red box at the top highlights the query string, which is automatically updated as filters are applied and can be edited directly; when edited, the corresponding filter fields below are populated where applicable. The data filter fields are also highlighted, as they are commonly used to batch queries when the requested dataset exceeds 200,000 records.

Query data using the iNaturalist API

The iNaturalist API call can return up to 200 observations per request, so querying this data may be pertinent. Similar queries as the iNaturalist Export Tool may be used, but the iNaturalist API offers additional parameters that may be used to query the data.

When using the iNaturalist API, you can extract the content of the request to get the total results. This can be done through the iNaturalist API webpage by entering filtering parameters then trying out the request. Under "Response Body," at the top of the text you will see total_results. This tells you how many results the query returns. Alternatively, you can do this through code by retrieving content of API request and examining the total_results data field. For example, see code snippet below:

```
{r}
library(httr)
library(jsonlite)

# API Request URL
api_request <- "https://api.inaturalist.org/v2/observations?place_id=21&taxon_id
=4715&order=desc&order_by=created_at"

# send HTTP GET request to the iNaturalist API
resp <- GET(api_request)

# extract the body of the HTTP response as plain text and convert to R object
data_parsed <- fromJSON(
  content(resp, as = "text", encoding = "UTF-8")
)

# View the total results
print(data_parsed$total_results)
```

If there are more than 200 results, then multiple requests will need to be made to the API to pull all the data. This can be done using the `per_page` parameter for up to 50 pages. This method works for up to 10,000 observations if requesting 200 observations per page. If more than 50

pages of data are needed, then separate queries will need to be made using the `id_above` parameter (see Pagination section of API Recommend Practices for more information).

When running batch queries keep in mind the iNaturalist query rate which is 1 request per second or 10,000 API requests per day. Throttles should be added into code to ensure that usage is under these limits.

S4: Introduction to modifying request URLs for the API

To acquire data using the iNaturalist API, a customized URL must be constructed. The base URL for the recommended API version is:

https://api.inaturalist.org/v2

From there, select an endpoint to retrieve data. The most commonly used endpoint is observations, but other endpoints include places, projects, taxa, observation_tiles, and more. Many endpoints also offer additional resources or paths to access specific types of data. For example, within the observations endpoint, you can access resources such as histogram, identifiers, observers, species_counts, and others. For example, to retrieve species counts specifically, the URL would be:

https://api.inaturalist.org/v2/observations/species_counts

To see a list of all endpoints and available parameters, visit the iNaturalist API website. The site conveniently allows users to enter parameters and automatically generates the corresponding request URL. It also provides a view of the response body and headers, which can help determine whether the desired data is returned before downloading it.

When parameters are applied, a "?" is added to the end of the base URL, and all filtering options are listed afterward. Multiple parameters are separated using "&". For example:

https://api.inaturalist.org/v2/observations?native=true&taxon_id=47157&acc_below=1000&order=desc&order_by=created_at

How to acquire ID values

Many of the filtering fields are straightforward. However, some require numeric ID values, such as taxon_id, place_id, project_id, site_id, user_id, user_login, and others. If only one or a few of these ID's are needed, the easiest method to acquire these values is by using the https://www.inaturalist.org/ website.

For taxon_id, navigate to the iNaturalist Explore Page and search for the taxon of interest (Figure S1). Once the filter is applied, notice that the URL has changed to show the taxon_id. This same methodology can be applied for place_id, project_id.

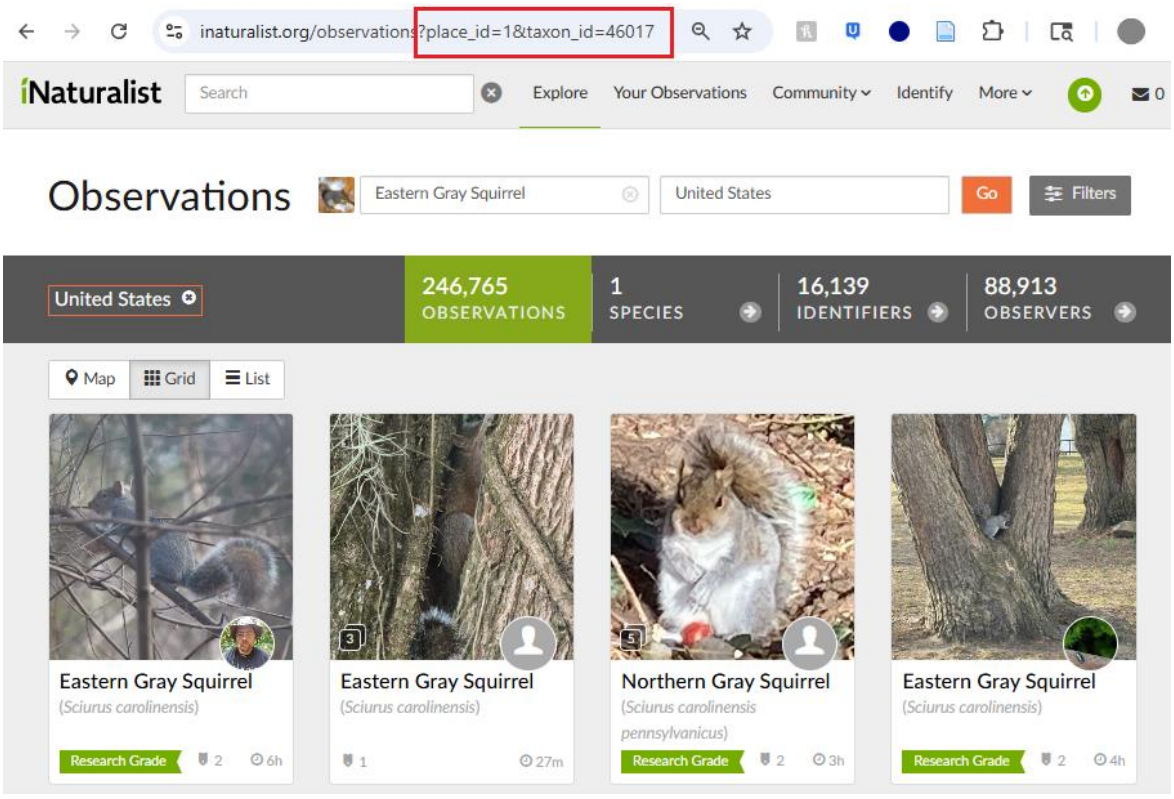


Figure S1. Using the iNaturalist Explore Page to obtain taxon_id and place_id. In this example, a search for eastern grey squirrels in the United States was performed, and the URL updated to include the corresponding taxon_id and place_id. This indicates that the taxon_id of eastern grey squirrels is 46017, and the place_id of the United States is 1.

To get user_login, navigate to the users profile page and see their username at the end of the URL. The user_id can be obtained by adding .json to the end of this URL (e.g. https://www.inaturalist.org/people/[username].json). This will return the JSON representation of the user's profile. The id field in the JSON object corresponds to the user's ID. If multiple user ID's are needed, then they can be obtained from iNaturalist exported data.

How to acquire and handle many ID values

If you need many ID values, downloading summarized datasets can save time. For example, if you need the taxon_id for all species in a place, project, from a specific user, or other filtering parameter, the observations/species_counts endpoint returns one row per species, including its taxon_id. For places, the places endpoint offers two useful options: places/autocomplete to search for places beginning with a given text string, and places/nearby to find places within a user-defined bounding box.

When setting up a request URL that includes multiple ID values, they should be added as a comma-separated list without spaces. For example:

https://api.inaturalist.org/v2/observations?taxon_id=47157,47222,51558,127588,321450

Alternatively, the commas can be encoded as "%2C", which will produce the same result.

Resources

The iNaturalist API documentation

For a more comprehensive overview on how to use iNaturalist's search URLs see this webpage.

S5: Code examples

We provide code examples on GitHub, which can be found at this link: https://github.com/Global-Ecology-Research-Group/Example-code-for-accessing-iNaturalist-data. This repository contains the following scripts.

1. **Get Observations:** Code to get observations from a URL request. For example, all butterfly observations in Texas.
2. **Species Counts:** Code to get a table of species counts from one URL request. For example, get a count of observations of all angiosperm species observed in Florida.
3. **Download Images from URL:** Code to download iNaturalist images from URL path. In the example, observation data are gathered from the iNaturalist API. Then parallel processing is used to download images hosted on Amazon Web Services. Images are named according to observation ID and photo number, so they can be linked to the full observation details later.
4. **Annotation Data:** Code to get annotation data from an API request.
5. **Identifier Data:** Code to get information on identifiers from an API request.
6. **Observer Data:** Code to get information on observers from an API request.
7. **Popular Field Values:** Code to get relevant controlled term values and monthly histogram of observations.

S6: Accessing data via the API

Acquiring iNaturalist data via the API is the more technically involved method of downloading the data, but it can be used to apply special filters to the data or acquire data not available via GBIF or the iNaturalist export tool. iNaturalist currently has three API versions: v0 (written in Ruby on Rails), v1 (in Node), and v2 (also in Node). The v0 API is considered legacy and should only be used for the few authentication endpoints that are not available in v1 or v2. The version recommended for most purposes is v2, with v1 being phased out (see https://api.inaturalist.org/v2/docs/). In general, v2 gives smaller, more customized responses. Please refer to iNaturalist's own documentation in the links provided in Supplemental Text S7 for the most up to date recommendations.

When downloading data via the API, iNaturalist throttles API usage to a max of 100 requests per minute, but they ask that users keep it to 60 requests per minute or lower. The maximum number of calls per day should not exceed 10,000 requests. By setting the per_page parameter to its maximum (usually 200 or 500), queries can be consolidated into fewer calls. In terms of media, downloads should be limited to 5GB of data per hour or 24 GB per day for files hosted directly on iNaturalist (there is no limit for images hosted in the iNaturalist Open Dataset). Given this, the API can be used to acquire small to medium amounts of data, but it is not meant to be used for bulk download.

There are several methods to download data from the iNaturalist API, but the most common approaches involve using coding languages such as Python or R. Both of these programming languages provide unofficial libraries that follow the iNaturalist API Recommended Practices and can be used to access the iNaturalist API. In Python, the `pyinaturalist` library offers functions for many commonly used API endpoints including observations, identifications,

annotations, observation fields, identifications, media, projects, and users. It also supports access to summarized data such as observation histograms, species counts, life lists, and more. In R, the `spocc` library may be used to retrieve observation data from GBIF or the iNaturalist API. There is another library in R called `rinat`, but this library relies on the v0 API and is therefore not recommended for general use.

While using existing libraries is preferred because they follow API Recommended Practices, there may be cases where data must be retrieved directly from the API because no prebuilt functions are available. This can be done using Python and R. In Python, the library `requests` can be used to send HTTP requests to the API. When making multiple calls to the API in python, the `time.sleep()` function from the built-in `time` module should be used to avoid exceeding the API usage limits In R, the `httr` and `jsonlite` libraries are commonly used. The `httr` package contains the `GET()` function which sends a request to the API. The `jsonlite` package parses the JSON data, the format used by the iNaturalist API, into an R object. Keep in mind that API responses are nested data structures. If you want to preserve this nested structure, you should save the data in .RDS format using `saveRDS()`. When making multiple calls to the API in R, the `Sys.sleep()` function from base R should be used to avoid exceeding API usage limits.

In this section, we gave a brief overview of using the iNaturalist API to download data. Users planning to work with the API should consult available resources (see S2: List of Resources for Accessing iNaturalist Data via the API below). With this article, we also provide code examples to directly retrieve data from the API using R and Python (see S3: Code Examples below). If

challenges arise, iNatForum can be searched for similar issues or used to post new questions for members of the community to answer.

S7: List of resources for accessing iNaturalist data via the API

There are many available resources for accessing data via the API. Here we list some of the most relevant resources, categorized by their purpose.

API references

API recommended practices: This webpage documents the best practices for using the iNaturalist API including authentication, pagination, bulk requests, query rate, and building software using the API. Anyone who plans to use the API to access data should read this page and respect the outlined practices.

iNaturalist v2 API: Use this webpage to view all available endpoints and parameters in version 2 of the iNaturalist API. It also allows you to enter parameters directly, and will generate the corresponding request URL along with a sample of the response and headers.

How to use iNaturalist's search URLs: Use this webpage to learn more about modifying request URLs.

Developers page: This webpage contains additional information specific to developers.

R Code

R for beginners: Using R coding to access data via the API requires an understanding of R. This resource provides a good introduction to R, but to become more proficient, you should seek out additional online resources or relevant workshops/courses.

spocc website: This website provides detailed information about the spocc package which can be used to download occurrence data from iNaturalist.

httr package information: A PDF file detailing the 'httr' package used to send requests to the API.

jsonlite package information: A PDF file detailing the 'jsonlite' package used to parse JSON data exported from the API to an R friendly format.

Python Code

User guide to pyinaturalist: A comprehensive guide to using the pyinaturalist package to download iNaturalist data.

Help Page

iNatForum: Any questions related to iNaturalist—including request URLs or coding questions—can be posted on the iNatForum. Start by searching the forum to see if your question has already been asked and answered. If not, you can create a new post to request help from the iNaturalist community.

S8: How to obtain iNaturalist heat maps

iNaturalist provides access to a small range of basic heat maps that can be used to visualize observation density. Entering the following URL will generate a heat map:

https://www.inaturalist.org/observations/map?[insert query]&color=heatmap

In place of the string '[insert query]', users can include queries to visualize species based on filters such as place, project, user, year or taxon. For example, a user may want to visualize all iNaturalist observations of peregrine falcons. Navigating to the About page for this species on iNaturalist gives the URL https://www.inaturalist.org/taxa/4647-Falco-peregrinus. In this URL, '4647' is the unique identifying number for peregrine falcons. To generate a heat map, a user would therefore use the following URL:

www.inaturalist.org/observations/map?taxon_id=4647&color=heatmap

The heatmap can be zoomed in on, or moved around. Some example queries include:

taxon_id=

place_id=

user_id=

project_id=

year=

Users can also combine multiple queries, separated by the & symbol, in a single URL. For example, the URL

https://www.inaturalist.org/observations/map?taxon_id=4647&place_id=97391&year=2024&color=heatmap

will show a heatmap of observations of peregrine falcons (taxon_id=4647) from Europe (place_id=97391) from the year 2024 (year=2024).

S9: Suggested reporting template for reproducible use of iNaturalist data

To support transparency, attribution, and reproducibility when using iNaturalist data in scientific analyses, we recommend that researchers explicitly document how the dataset used in their study was obtained, filtered, and processed. The following template provides a suggested structure that can be adapted for use in the Methods or Supplementary Materials of a publication.

Example reporting template for GBIF Data Downloads

iNaturalist observation data were obtained on **[DATE]** via **GBIF** using the following DOI: **[GBIF DOI]**. The dataset included observations that met the following criteria: Research Grade, belonging to the taxonomic scope **[taxon names or taxon IDs]**, occurring within the spatial extent **[geographic boundaries, place IDs, bounding box, etc.]**, and recorded during the time period **[date range]**.

Example reporting template for data obtained via the iNaturalist API or iNaturalist Export Tool

iNaturalist observation data were obtained on **[DATE]** via **[iNaturalist API / iNaturalist Export Tool / combination of sources]**. The **[iNaturalist API / iNaturalist Export Tool]** was queried using the following parameters: **[list of parameters or provide request / query URL]**.

The dataset included observations that met the following criteria: **quality grade [Research Grade / Needs ID / Casual / combination]**, belonging to the taxonomic scope **[taxon names or taxon IDs]**, occurring within the spatial extent **[geographic boundaries, place IDs, bounding box, etc.]**, and recorded during the time period **[date range]**. Additional filters were applied as

needed to match the research objective, including **[e.g., annotations, observation fields, project membership, or user-level filters]**.

Example reporting template to describe data filtering methods

Licensing filters were applied to ensure compliance with contributor licensing. Observations included licenses **[CC0 / CC-BY / CC-BY-NC / etc.]**, and associated media were used only where licenses permitted reuse. All media and observation records were attributed according to the requirements of the corresponding Creative Commons licenses. Geographic uncertainty was handled as follows: observations with coordinate uncertainty greater than **[threshold in meters]** were **[excluded / retained / treated using buffers]**. Observations with obscured or private coordinates were **[excluded / included with randomized coordinates / accessed via trusted coordinates where permitted]**. Taxonomic interpretation followed **[iNaturalist taxonomy / GBIF Backbone Taxonomy / other authority]**, accessed on **[DATE]**. Taxonomic updates occurring after this date were not incorporated into the analysis. The processed dataset used for analysis, including observation ids and derived variables, has been archived at **[repository, e.g., Zenodo, Dryad, GitHub]**, with the archive DOI **[DOI]**. Scripts used to retrieve and process the data are provided in **[repository link]**.