*Biodiversity Informatics*, 20, 2026, pp. 1-4

# DOWNLOADING IMAGES FROM GBIF: LICENSES, CITATION, AND LINK ROT

Mark Pitblado[1] and Quentin Cronk[1,2*]

[1]*Beaty Biodiversity Museum, University of British Columbia, Vancouver, BC, Canada*
[2]*Department of Botany, University of British Columbia, Vancouver, BC, Canada*

*Abstract.* Downloading images of preserved specimens in bulk is becoming increasingly important for many research projects, especially those connected with machine learning and image analysis. A useful source of images is the standard biodiversity aggregator, the Global Biodiversity Information Facility (GBIF). Here we identify four major issues connected to GBIF image downloads, distinct from those associated with text downloads. These are (1) license considerations, (2) citation issues, (3) restricting to specific providers for project reasons or cybersecurity concerns, and, finally, (4) attempting to use links that are no longer functioning (often referred to as "link rot" or "data rot"). We suggest an incremental approach to downloading and suggest techniques for improved image download. We provide an implementation of our suggestions in Python (gbif-image-downloader).

*Key words.*—GBIF, image download, machine learning, museums, specimen images, link rot, data rot, copyright

## Introduction

The Global Biodiversity Informatics Facility (GBIF) aggregates over 3 billion biodiversity records from around the world and makes the data freely available to all. This aggregation has enabled large-scale scientific study of biodiversity at the planet scale, unlocking methodologies that previously were not possible. Individual countries, museums, and platforms and aggregators such as iNaturalist contribute to GBIF as publishers, largely self-enforcing issues of data quality and accuracy.

Importantly, GBIF does not host or store multimedia data such as images, sounds, or videos. Instead, GBIF relies on publishers to serve files from their own infrastructure. Through the publication process, publishers link to locations where they store the multimedia files, either through a hyperlink in the associatedMedia column of their records or via the Audiovisual Media Description Darwin Core extension. As these multimedia files are much larger than the text data for records, this model permits substantial cost savings for GBIF, while allowing for multimedia to be shown and made available to end users through GBIF websites and tools.

In addition to conducting traditional analysis on tabular data, researchers are increasingly interested in leveraging information held in the form of images to derive insights from specimens (Körschens et al. 2024). Many image-based AI models improve when trained on large amounts of data, thus making GBIF an ideal scientific resource, given its status as a worldwide data aggregator. However, in view of the model described above, in which publishers independently host image content, downloading image data from GBIF can be less straightforward than downloading tabular data that are hosted directly by GBIF.

The differences in using image data compared to textual data can broadly be defined by four categories: (1) license considerations, (2) citation, (3) restrictions to specific providers (for project reasons or cybersecurity concerns), and (4) attempting to use links that are no longer functioning; the latter problem is often referred to as "link rot" or "data rot" (Briney, 2024). Link rot in biodiversity informatics is not a problem unique to images (Elliot et al. 2020), but can be harder to detect systematically in images than in textual data, owing to the data being hosted across more sites.

* Corresponding author: quentin.cronk@ubc.ca

*License considerations*

In 2014, the GBIF Governing Board established a policy of only permitting CC0, CC BY, and CC BY-NC licenses for occurrence datasets. However, it is unclear whether the license of a dataset applies to the images to which it links. In theory, for publishers using the Audiovisual Media Description Darwin Core Extension, terms are available by which to communicate usage rights for images (e.g., rights, UsageTerms, and Owner). In practice, however, the Simple Media Extension is what is used by most users, and that extension does not provide any means by which to add these terms. Instances in which the data are licensed but no license exists for an image may also exist.

*Citation*

Citation is foundational to science, both because it allows for claims made by current authors to build on centuries of previous work, and because it creates a crucial metric of contribution to the broader scientific community that is important for funding and career advancement. GBIF provides a robust mechanism for tabular data by minting a DOI at the point of download for a dataset, allowing researchers to include a DOI specifically for the information used in their publication (as opposed to the entirety of GBIF-mediated data). Still, though the GBIF DOIs ensure that tabular data can be obtained exactly as presented to the original researcher, the same cannot be said about the images, as they are linked to an external source that may change or be removed.

*Source restrictions*

If a data user wishes to use images for research, they often need to download these images to their machine through a scalable and automated method. This step can be achieved through a few lines of code that take a URL as input and download the image. As the user downloading images through this methodology does not preview the image before downloading it to their machine, they must place trust in the entity hosting the image to have protections in place against malicious distribution, and to be a good-faith actor in the scientific community. While usually not a problem when interacting with a single publisher, the difficulty increases drastically when downloads are done on a worldwide scale over many publishers. Although instances are rare, image files can be and have been used to spread malware (Carter & Randolph 2015).

*Link rot*

Lastly, although GBIF has mechanisms in place to preserve data if the original source becomes inactive, the same benefits do not extend to images. Since GBIF does not store image files, it is entirely up to the data publisher to maintain the links to their images over time. Publishers that are facing cost pressure may stop serving their images before resorting to taking down other services, as serving image files is expensive, from both a bandwidth and a storage perspective. Alternatively, if publishers change the mechanism by which images are delivered, either by switching to a different provider or changing where the images are hosted, URLs will break if a dedicated and stable structure for these links was not solidified beforehand.

## Methods to Download Images

Conceptually, two ways to download images are available: (1) the requester can request a dataset for download (through either the web interface or the API), or (2) images can be requested incrementally until a certain quota is met using the GBIF occurrence search API. Each method comes with benefits and drawbacks. An example program is rgbif (Chamberlain et al., 2024), which provides an R interface with GBIF. It can potentially implement both methods, but does not have the ability to handle link rot.

*Method 1.* If a dataset is downloaded, a DOI is minted for the dataset, and it is available as a stable text file, but there is a small amount of time required for GBIF to prepare the dataset and send a notification to an email address that the dataset is ready. Links to images in the text file could be reviewed before initiating download requests. However, there is no way to know what percentage of those links will be valid before making the requests or running the links through a link checker. If the number of valid links does not meet the minimum number of images that the researcher needs, the researcher will have to reinitiate the process by broadening the filter, requesting a new dataset, and making the image requests again. This results in duplication of computational work, unnecessary minting of DOIs, and an increase in wait-time for the researcher.

*Method 2.* If images are requested incrementally through the GBIF occurrence search API, the program will continue requesting images until it has obtained at least the minimum number of images desired by the researcher. Done optimally, this approach minimizes the number of requests made, and does not leave the researcher with a DOI that references data not used. Requesting images incrementally does not allow for checking the links ahead of time; rather, the researcher will likely never see the links to the images before the requests are made.

This paper describes a means of implementing both methods—first Method 2 and then Method 1—to solve the link rot issue while maintaining citability. A solution is offered here in Python (gbif-image-downloader), but

**Table 1**: HTTP Status codes for three sample requests by scientific name. Each request was made for 150 images. The total number of requests made will exceed the number images, due to invalid requests not resulting in an image, and requests being sent in batches.

| Scientific Name | Valid | Invalid | Link rot percentage |
|---|---|---|---|
| *Vulpes vulpes* | 159 – (HTTP 200) | 15 – (HTTP 401 unauthorized) | 9% |
| *Artemisia verlotiorum* | 159 – (HTTP 200) | 1 – (HTTP 429 too many requests) [temporary error]<br>90 – (HTTP 403 Forbidden)<br>134 – (No HTTP code) | 59% |
| *Bubo virginianus* | 154 – (HTTP 200) | 8 – (HTTP 401 unauthorized)<br>72 – (No HTTP code) | 34% |

other languages (including R) would be equally suitable. The application has many potential uses, including training neural networks for taxonomic identification tools, or image analysis to reveal geographic patterns in morphology. Beside taxon name, other search terms available in the GBIF API may be used (see the README for details and limitations); it has been tested with "year" and "basisOfRecord." To allow multiple searches or projects to be conducted at once, the program asks the user for a prefix, then creates the output directory for them. For details of error handling, potential users should consult the README document.

## Techniques for Improved Incremental Image Download

Given the above considerations, we established a methodology that would allow researchers to download images from GBIF in a user-friendly way via scientific name, provide attribution to the original publishers of images, mitigate cybersecurity risk of downloading images from such a wide variety of publishers, and check for the validity of links before consuming bandwidth when making a full request.

1. Request images incrementally via batches, with each batch proportional to the number of images requested. If "strict" mode is enabled and the researcher has specified only an allowlist of publishers (say, only get images from Kew, the Smithsonian, and Beaty), then make incremental requests only from those publishers.
2. Along with saving each image, save two additional pieces of information: the GBIF occurrence ID, and the license of the image (scan through license columns in both the data and in the extension file if the publisher is using it).
3. After the minimum number of images requested has been downloaded, request a GBIF download using the assembled occurrence IDs. This provides a DOI, but instead of doing it beforehand, we assemble it after the fact. This provides the benefits of incremental download, while preserving the benefits of a download request.

## Observations From Use

Above we provide data on request responses using the tool described here. Although based on a small sample size, we found that (1) a surprisingly high number of cases (close to 100%) returned specified licenses for the image, (2) the failure rate varied drastically based on how concentrated the images were between publishers. This variation likely emerges because link rot is not distributed randomly: if a particular publisher changes or removes their image system, *all* links from that publisher simultaneously break. (see Table 1).

## RESOURCE AVAILABILITY

The code used in this paper is available on GitHub (https://github.com/mark-pitblado/gbif-image-downloader)

## ACKNOWLEDGMENTS

Work in the laboratory of QC is funded by the Natural Science and Engineering Research Council (NSERC, Canada) Discovery Grant program (RGPIN-2025-04763).

## COMPETING INTERESTS

The authors have declared that no competing interests exist.

## LITERATURE CITED

Briney, K. A. (2024). Measuring data rot: An analysis of the continued availability of shared data from a single university. *PLoS One, 19(6)*, e0304781. https://doi.org/10.1371/journal.pone.0304781

Chamberlain, S., Barve, V., Mcglinn, D., Oldoni, D., Desmet, P., Geffert, L., and Ram, K. (2024). rgbif: Interface to the Global Biodiversity Information Facility

API. R package version 3.8.1, https://CRAN.R-project.org/package=rgbif

Carter, E., and Randolph, N. (2015, February 25). Malicious PNGs: What you see is not all you get! https://blog.talosintelligence.com/malicious-pngs-what-you-see-is-not-all/

Elliott, M. J., Poelen, J. H., and Fortes, J. A. B. (2020). Toward reliable biodiversity dataset references. *Ecological Informatics*, *59*, 101132. https://doi.org/10.1016/j.ecoinf.2020.101132

Körschens, M., Bucher, S. F., Bodesheim, P., Ulrich, J., Denzler, J., and Römermann, C. (2024). Determining the community composition of herbaceous species from images using convolutional neural networks. *Ecological Informatics*, *80*, 102516. https://doi.org/10.1016/j.ecoinf.2024.102516