Abstract for Digital Data in Biodiversity Research 2023
https://www.idigbio.org/content/digital-data-2023-leveraging-digital-data-service-conservation-ecology-systematics

title:

Signed Biodiversity Data Packages: A Method to Cite, Verify, Mobilize, and Future Proof, Large Image Corpora.

authors:

Jorrit H. Poelen, Ronin Institute, Cheadle Center for Biodiversity and Ecological Restoration UC Santa Barbara https://orcid.org/0000-0003-3138-4118

Jason Best, Botanical Research Institute of Texas, https://orcid.org/0000-0002-7414-5523

abstract:

Access to Natural History Collections helps researchers to better understand the natural world. Millions of digital images of herbarium specimens are openly available via the Internet. However, using these images in a data-intensive research project raises basic questions like: "How do I efficiently access, and verify, hundreds of thousands of images?", and, "How do I cite a version of a large image corpus?"

Here, we present a method to cite, verify and mobilize such image corpora across different locations and medium types. We demonstrate our method with >100k images made available through the Botanical Research Institute of Texas using available tools (e.g., rsync, Preston) and technologies (e.g., internet, postal service). Our results show that our packaging method allows the US Postal Service to transfer a packaged corpus at about 3 images/s, whereas retrieving individual images via HTTP achieved a transfer rate of about 0.2 images/s.

Our results support that signed digital packaging of image corpora enables distributed storage using readily available transfer and storage methods. In addition, our method is future proof because they can be used with any digital media, including those that are not yet available.