BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 3: e37080
doi: 10.3897/biss.3.37080

OPEN ACCESS

Conference Abstract

Zenodo, an Archive and Publishing Repository: A tale of two herbarium specimen pilot projects

Mathias Dillen[‡], Quentin Groom[‡], Donat Agosti[§], Lars Holm Nielsen[|]

‡ Meise Botanic Garden, Meise, Belgium
§ www.plazi.org, Bern, Switzerland
| CERN, Meyrin, Switzerland

Corresponding author: Mathias Dillen (mathias.dillen@plantentuinmeise.be)

Received: 11 Jun 2019 | Published: 18 Jun 2019

Citation: Dillen M, Groom Q, Agosti D, Nielsen L (2019) Zenodo, an Archive and Publishing Repository: A tale of two herbarium specimen pilot projects. Biodiversity Information Science and Standards 3: e37080. https://doi.org/10.3897/biss.3.37080

Abstract

Zenodo (https://zenodo.org) is an open-access repository operated by CERN (European Organization for Nuclear Research), which provides researchers with an easy and stable platform to archive and publish their data and other output, such as software tools, manuals and project reports. In the context of the ICEDIG (Innovation and Consolidation for Large scale Digitisation of Natural Heritage) project, Zenodo was investigated for its usability as a platform where digitized images of collection specimens could be archived and published. In a production digitization pipeline, we foresee the automated archiving of daily image production. If Zenodo could be used for this purpose, such a process would also immediately mean that data and images are published FAIR-ly (Findable, Accessible, Interoperable and Reusable) within hours of their creation.

To evaluate performance of the system, we first used a test dataset of 1800 herbarium specimen images, which was uploaded using Zenodo's API (Application Programming Interface) (Dillen et al. 2019). This dataset includes lossless TIFF images, label-segmented overlays and JSON-LD (JavaScript Object Notation for Linked Data) metadata using DwC (Darwin Core) terminology, constituting over 208 gigabytes of data. In addition, for all individual digital specimens the data about the specimen (in DwC) as well as metadata about its deposition on Zenodo (in Zenodo's internal data model) were available in multiple machine-readable formats. All data in DwC were provided as linked data with their DwC identifiers (e.g. http://rs.tdwg.org/dwc/terms/basisOfRecord). All individual specimens received minted DOIs (Digital Object Identifiers). A second upload of 280,000 herbarium JPEG images from a single institution (ca. 1 terabyte of data) with limited metadata (but using the same approach) was launched as well.

In this presentation, the workflow for proper usage of the API will be described as well as some performance metrics, flexibilities and functionalities of the platform. Some issues and potential developments to tackle them will be discussed. Currently, the rate of ingestion into Zenodo seems only fast enough for small scale digitization pipelines. However, a modest improvement in transfer rate would make this a realistic proposition for large volume usage.

Keywords

linked data, long-term preservation, workflow, Darwin Core

Presenting author

Mathias Dillen

Presented at

Biodiversity_Next 2019

Funding program

Horizon 2020 Framework Programme of the European Union

Hosting institution

ICEDIG – “Innovation and consolidation for large scale digitisation of natural heritage” H2020-INFRADEV-2016-2017 – Grant Agreement No. 777483

References

- Dillen M, Groom Q, Chagnoux S, Güntsch A, Hardisty A, Haston E, Livermore L, Runnel V, Schulman L, Willemse L, Wu Z, Phillips S (2019) A benchmark dataset of herbarium specimen images with label data. Biodiversity Data Journal 7: e31817. https://doi.org/10.3897/bdj.7.e31817