Biodiversity Information Science and Standards 3: e37081
doi: 10.3897/biss.3.37081

Conference Abstract

Next Steps in Data Capture from Specimen Labels and Data Integration: Lessons learnt from the ICEDIG pilots

Mathias Dillen[‡], Quentin Groom[‡], Sarah Phillips[§], Irena Spasic[|]

‡ Meise Botanic Garden, Meise, Belgium
§ Royal Botanic Gardens Kew, Surrey, United Kingdom
| Cardiff University, Cardiff, United Kingdom

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

Received: 11 Jun 2019 | Published: 18 Jun 2019

Citation: Dillen M, Groom Q, Phillips S, Spasic I (2019) Next Steps in Data Capture from Specimen Labels and Data Integration: Lessons learnt from the ICEDIG pilots. Biodiversity Information Science and Standards 3: e37081. https://doi.org/10.3897/biss.3.37081

Abstract

The rapid development and refinement of digital technologies in the last two decades has spearheaded a wave of digitization in natural history collections. This has generated a massive number of digitized images and many more are expected with the planned European Distributed Systems of Scientific Collections (DiSSCo) infrastructure. Many of these images will contain labels with data written on them, typed on them or interpretable from them, but capturing these data remains a challenge. Automated as well as manual methods are being investigated and have yielded mixed results. In addition, previously captured data or data to be captured this way will need to be interoperable in order to make digital access and enrichment most effective. Finally, institutions holding the physical specimens will need to remain capable of efficiently curating the digital, potentially annotated, counterparts. This will require compatibility with the diverse data models of local Collection Management Systems (CMS).

In the context of the ICEDIG (Innovation and consolidation for large scale digitisation of natural heritage) project, a benchmark dataset of herbarium specimens was assembled from nine contributing institutions (Dillen et al. 2019). This dataset was used to evaluate automated methods of text recognition, such as OCR (Optical Character Recognition) and

HTR (Handwritten Text Recognition), and post-capture classification, such as language identification or NER (Named Entity Recognition). A pipeline from scan to scientifically useful data was drafted and guidelines for selecting appropriate software solutions were provided.

The benchmark dataset was also processed through multiple crowdsourcing platforms, after which the quality and interoperability of the resulting transcriptions was analyzed. The aptitude of local Collection Management Systems to curate these digitized specimens efficiently was investigated, as well as the fitness of data standards in use to ensure and maintain proper interoperability. In addition, available surveys on CMS use and satisfaction were summarized and in-depth assessments of the CMS in use at the ICEDIG partner institutes were performed. A summary of results and recommendations will be presented.

Keywords

automation, machine learning, data capture, interoperability, collection management system

Presenting author

Mathias Dillen

Presented at

Biodiversity_Next 2019

Funding program

Horizon 2020 Framework Programme of the European Union

Grant title

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