Towards a COST MOBILISE Guideline for Long Term Preservation and Archiving of Data Constructs from Scientific Collections Facilities

Dagmar Triebel[‡], Dragan Ivanovic[§], Gila Kahila Bar-Gal[|], Sven Bingert[¶], Tanja Weibulat[‡,#]

‡ Staatliche Naturwissenschaftliche Sammlungen Bayerns, SNSB IT Center, Munich, Germany
§ University of Novi Sad, Novi Sad, Serbia
| The Hebrew University of Jerusalem, Rehovot, Israel
¶ Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen, Göttingen, Germany
# German Federation for Biological Data (GFBio e.V.), Bremen, Germany

Corresponding author: Dagmar Triebel (triebel@snsb.de), Dragan Ivanovic (dragan.ivanovic@uns.ac.rs)

Abstract

COST (European Cooperation in Science and Technology) is a funding organisation for research and innovation networks. One of the objectives of the COSTAction called “Mobilising Data, Policies and Experts in Scientific Collections“ (MOBILISE) is to work on documents for expert training with broad involvement of professionals from the participating European countries. The guideline presented here in its general concept will address principles, strategies and standards for long term preservation and archiving of data constructs (data packages, data products) as addressed by and under control of the scientific collections community. The document is being developed as part of the MOBILISE Action targeted towards primarily scientific staff at natural scientific collection facilities, as well as management bodies of collections like museums, herbaria and information technology personnel less familiar with data archiving principles and routines.

The challenges of big data storage and (distributed, cloud-based) storage solutions as well as that of data mirroring, backing up, synchronisation and publication in productive data environments are well addressed by documents, guidelines and online platforms, e.g., in the DISSCo knowledge base (see Hardisty et al. (2020)) and as part of concepts of the European Open Science Cloud (EOSC). Archival processes and the resulting data constructs, however, are often left outside of the considerations. This is a large gap because archival issues are not only simple technical ones as addressed by the term “bit preservation” but also envisage a number of logical, functional, normative, administrative and semantic issues as addressed by the term “functional long-term archiving”.

The main target digital object types addressed by this COST MOBILISE Guideline are data constructs called Digital or Digital Extended Specimens and data products with the persistent identifier assignment lying under the authority of scientific collections facilities.

Such digital objects are specified according to the Digital Object Architecture (DOA , see Wittenburg et al. 2018) and similar abstract models introduced by Harjes et al. (2020) and Lannom et al. (2020). The scientific collection-specific types are defined following evolving concepts in the context of the Consortium of European Taxonomic Facilities (CETAF), the research infrastructure DiSSCo (Distributed System of Scientific Collections), and the Biodiversity Information Standards (TDWG). Archival processes are described following the OAIS (Open Archival Information System) reference model. The archived objects should be reusable in the sense of the FAIR (Findable, Accessible, Interoperable, and Reusable) guiding principles. Organisations like national (digital) archives, computing or professional (domain-specific) data centers as well as libraries might offer specific archiving services and act as partner organisations of scientific collections facilities.

The guideline consists of key messages that have been defined. They address the collection community, especially the staff and leadership of taxonomic facilities. Aspects of several groups of stakeholders are discussed as well as cost models. The guideline does not recommend specific solutions for archiving software and workflows. Supplementary information is delivered via a wiki-based platform for the COST MOBILISE Archiving Working Group WG4.

Keywords

data standards, data packages, data products, digital archives, digital objects, DiSSCo, DOA, EOSC, functional long term archiving, OAIS

Presenting author

Dragan Ivanovic

Presented at

TDWG 2021

Funding program

COST Action CA 17106

Conflicts of interest

References

- Hardisty A, Saarenmaa H, Casino A, Dillen M, Gödderz K, Groom Q, Hardy H, Koureas D, Nieva de la Hidalga A, Paul D, Runnel V, Vermeersch X, van Walsum M, Willemse L (2020) Conceptual design blueprint for the DiSSCo digitization infrastructure - DELIVERABLE D8.1. Research Ideas and Outcomes 6: e54280. https://doi.org/10.3897/rio.6.e54280
- Harjes J, Link A, Weibulat T, Triebel D, Rambold G (2020) FAIR digital objects in environmental and life sciences should comprise workflow operation design data and method information for repeatability of study setups and reproducibility of results. *Database* 2020: 1-20. https://doi.org/10.1093/database/baaa059
- Lannom L, Koureas D, Hardisty AR (2020) FAIR data and services in biodiversity science and geoscience. *Data Intelligence* 2 (1-2): 122-130. https://doi.org/10.1162/dint_a_00034
- Wittenburg P, Strawn G, Mons B, Boninho L, Schultes E (2018) Digital objects as drivers towards convergence in data infrastructures. *Technical paper* https://doi.org/10.23728/b2share.b605d85809ca45679b110719b6c6cb11