BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 3: e35268
doi: 10.3897/biss.3.35268

OPEN ACCESS

Conference Abstract

People of Collections: Facilitators of Interoperability?

Chloé Besombes‡, Simon Chagnoux‡, Gildas Illien‡

‡ Muséum national d'histoire naturelle, Paris, France

Corresponding author: Chloé Besombes (chloe.besombes@mnhn.fr)

Received: 08 Apr 2019 | Published: 18 Jun 2019

Citation: Besombes C, Chagnoux S, Illien G (2019) People of Collections: Facilitators of Interoperability? Biodiversity Information Science and Standards 3: e35268. https://doi.org/10.3897/biss.3.35268

Abstract

In March 2019, the Muséum national d'histoire naturelle, Paris (MNHN) launched the datapoc.mnhn.fr project, funded by the French research infrastructures CollEX-Persée and E-recolnat. This proof of concept was imagined and is supported by a group of partners coming from different communities working at the Muséum (specimen collection curators, librarians, researchers, data scientists, publishers). The initial motivation of this team for getting together was to imagine a way to link the massive data produced and preserved in the heterogeneous institutional collection databases and repositories of the Muséum in order to improve global access and visibility for the benefit of end-users as well as data curation processes. After a year of sharing and deliberating, the group concluded that focusing on people's names and identification, could be a promising way to explore interoperability and alignment solutions in order to match data hosted in the different systems.

The project has thus two main goals: first, to improve biodiversity and taxonomic data quality for the qualification of personal identities, publications and scientific names by resolving frequent ambiguities and issues in people's names assignment ; second, to develop and assess machine-driven linking strategies between specimen and authorship metadata and resources derived from various institutional datasilos of interest to the research community.

In order to test this idea and to experiment innovative data computing and visualization technologies, all parties involved in the project agreed to develop a proof of concept focused on a dataset of 500 names of major MNHN naturalists from its foundation until nowadays. This proof of concept will consist in building a structured authority file for people's names, which could be shared by all services producing and using biodiversity data at MNHN, as well as reusable as open data by external stakeholders and international partners. This structured file will strengthen data and databases production and maintenance workflows, but could also help improving the quality of end-user experience by allowing individuals or machines to match, link or otherwise compute and analyse data that is still difficult to handle because of the diversity of IT applications and limited standardisation practises. It is key to the project that this structured file should somehow comply with international interoperability and semantic web standards so to facilitate global access and data exchanges with similar institutions around the world. Linked datasets and related resources derived from this work will be displayed on a public website designed for researchers as well as for the public via diverse applications and formats (API, RDF). The project will be run from April 2019 to April 2020 by the core team of partners who initiated it, with the support of a private IT and data computing service called Logilab.

Some of the challenges of this project include finding an efficient way for building the structured file and then succeed in aligning and disambiguising names already present existing databases. A way to approach this issue is to confront and consolidate MNHN biodiversity datasets with external repositories by using people identifiers systems like ISNI, VIAF, IdREF, which are already familiar to libraries, archives and other cultural institutions. How can those various people identifiers systems be profitable to parse MNHN "people of collections" and help disambiguise them? Is there a particular people identifier system which will prove to be most relevant for all types of collections? Which parsing method will give the best results, and how could it scale up and possibly be reused by other institutions or even future European taxonomic infrastructures? Those are some of the questions the MNHN team is eager to deal with and to share and discuss at the Biodiversity Next Symposium.

Keywords

interoperability, people identifiers, people names, linked data, shared standards

Presenting author

Gildas Illien, Chloé Besombes