From Shells in House Cabinets to Structured Data for Research: The mobilization of frozen biodiversity data in Italy

Arianna Giannini‡, Marco Oliverio‡

‡ Department of Biology and Biotechnology “Charles Darwin”, La Sapienza University, Rome, Italy

Corresponding author: Arianna Giannini (arianna.giannini@uniroma1.it)

Abstract

In recent decades, technological development has accelerated exponentially, and with it the volume of data that can be accumulated and processed (Runting et al. 2020). The big data revolution has enabled great steps forward in natural sciences, allowing the study of global changes at different scales (Nelson and Ellis 2018). Today, biodiversity research has focused more on data quantity than quality, leading to a shift in the collecting methods of primary biodiversity data from specimen-based to observation-based. Some authors argued that the increasing disconnection of occurrence data from actual specimens has some aspects of suboptimality that cannot be ignored, despite also having many benefits (Troudet et al. 2018). In this context, Natural History Collections (NHCs) contain data of potential high quality when specimens are collected and identified by experts; however, most NHCs' data are not databased, records must be digitized to become usable by researchers and other stakeholders, and not all owners have the tools to do so (Fig. 1). In Italy—as in other countries—many specimens of invertebrates are stored in private collections, the majority not databased, and even when they are digitized, they rarely follow international standards, such as Darwin Core - DwC (Darwin Core Task Group 2009). We call this type of data *frozen*. The production of an accessible nationwide database derived from the digitization of these records could significantly support research and national conservation strategies. This project aims to support the databasing of private collections in Italy and collect their records in one structured geo- and chrono-referenced database of biodiversity data in line with international standards. We have chosen marine molluscs as a pilot taxon, based on three criteria: 1) existence of an updated checklist of the Italian fauna (Renda et al. 2022); 2) existence of an updated taxonomic reference to serve as a thesaurus for the database, namely MolluscaBase (MolluscaBase eds. 2022) and the World Register of Marine Species - WoRMS (WoRMS Editorial Board 2022); 3) management and conservation relevance of the taxon, based on classic criteria for selecting indicator taxa (e.g., Pearson 1994). For data collection, we built an empty template Excel spreadsheet, for ease of use by the terminal operator. The template file contains 21 fields, summarized in Fig. 2, and it is accompanied by other support files (Fig.

3). As of 01 Jul 2022, we had contacted only a small number of specialists, collecting >9500 records. While data are collected from different collections, records will be reorganized into a single database according to the DwC standard. Each record will then be georeferenced following Zermoglio et al. (2020)'s protocol and it will be traceable through a system of Persistent Identifiers. By this project, we aim to foster the mobilization of frozen biodiversity data through a process of digitization and integration of different sources. We expect to produce a database containing a large number of records in a few years, making it available for research and biodiversity management.

Keywords

database, natural history collections, big data, marine molluscs, Mollusca

Presenting author

Arianna Giannini

Presented at

TDWG 2022

Conflicts of interest

References

- Darwin Core Task Group (2009) Darwin Core. Biodiversity Information Standards (TDWG). URL: http://www.tdwg.org/standards/450
- MolluscaBase eds. (2022) MolluscaBase. https://www.molluscabase.org. Accessed on: 2022-6-30.
- Nelson G, Ellis S (2018) The history and impact of digitization and digital data mobilization on biodiversity research. Philosophical Transactions of the Royal Society B 374 (20170391): 1-9. https://doi.org/10.1098/rstb.2017.0391
- Pearson DL (1994) Selecting indicator taxa for the quantitative assessment of biodiversity. Philisophical Transactions of the Royal Society of London, Series B: Biological Sciences 345 (1311): 75-79. https://doi.org/10.1098/rstb.1994.0088
- Renda W, Amati B, Bogi C, et al. (2022) The new Checklist of the Italian Fauna: marine Mollusca. Biogeographia – The Journal of Integrative Biogeography 37 (1): 1-86. https://doi.org/10.21426/B637156028
- Runting R, Phinn S, Xie Z, et al. (2020) Opportunities for big data in conservation and sustainability. Nature Communications 11 (2003): 1-4. https://doi.org/10.1038/s41467-020-15870-0

- Troudet J, Vignes-Lebbe R, Grandcolas P, et al. (2018) The Increasing Disconnection of Primary Biodiversity Data from Specimens: How Does It Happen and How to Handle It? Systematic Biology 67 (6): 1110-1119. https://doi.org/10.1093/sysbio/syy044
- WoRMS Editorial Board (2022) World Register of Marine Species. https://www.marinespecies.org. Accessed on: 2022-6-30.
- Zermoglio PF, Chapman AD, Wieczorek JR, et al. (2020) Georeferencing Quick Reference Guide. GBIF Secretariat, Copenhagen. https://doi.org/10.35035/e09p-h128

SOURCE	PUBLIC COLLECTIONS	PRIVATE COLLECTIONS	STRUCTURED CITIZEN SCIENCE PROJECTS	OTHER OBSERVATION DATA
TYPE	SPECIMEN-BASED	SPECIMEN-BASED	OBSERVATION-BASED	OBSERVATION-BASED
QUALITY				
Taxonomic quality	Expert-based	Expert-based	Difficult to assess	Difficult to assess
Georeference quality	Expert-based	Expert-based	Difficult to assess	Difficult to assess
ACCESSIBILITY				
Access	Usually open	Private	Usually open	Usually open
Databasing	Funds-dependent	Sometimes	Yes	No
International standards are met	Sometimes	No	Sometimes	No
Imaging	Funds-dependent	No	Not possible	Not possible
ANCILLARY DATA				
Multimedia files	Can be acquired later	Can be acquired later	Must be acquired at the moment	Must be acquired at the moment
DNA	Can be acquired later	Can be acquired later	Not possible	Not possible
USABILITY				
Ecological studies	Yes	Yes	Yes	Only if data is collected
Taxonomic studies	Yes	Yes	Not possible	Not possible
Phylogenetic analyses	Yes	Yes	Not possible	Not possible

Figure 1.

Features of four different sources of occurrence data: 1) public collections, 2) private collections, 3) structured citizen science projects (i.e., projects where occurrences are combined into a single database), and 4) other observation data (e.g., scattered data from online sources).

DWC CLASS	*CATEGORY*	*DESCRIPTION*
Record-level	Ownership	The name or the acronym of the institution or person having custody of the specimen.
Record-level	CollectionCode	The name, acronym or code identifying the collection.
Record-level	Record	The nature of the record. It corresponds to the DwC term *basisOfRecord*. Most of the records are *PreservedSpecimen*.
Occurrence	CatalogNumber	A unique identifier of the specimen within the collection.
Occurrence	RecordedBy	The name of who recorded the original occurrence and collected the specimen (= *legit*).
Occurrence	IndividualCount	The number of individuals of the same species present at the time of the occurrence.
Occurrence	IndividualCountShells	The number of empty shells of the same species present at the time of the occurrence.
Event	Date	The date of the collecting event. It corresponds to the DwC term *EventDate*.
Identification	IdentifiedBy	The name of who assigned the taxon to the specimen (= *determinavit*).
Taxon	ScientificName	The full scientific name at species or subspecies taxonomic rank, according to WoRMS.
Taxon	TaxonRank	The taxonomic rank of the scientific name.
Location	MinDepth	The depth of the occurrence or, in case of ranges, the minimum depth.
Location	MaxDepth	The maximum depth of the occurrence.
Location	Locality	A locality name or a specific description of the place.
Location	Latitude	Latitude of the occurrence.
Location	Longitude	Longitude of the occurrence.
Location	CoordinateFormat	Format of the coordinates.
Location	Datum	Geodetic datum, ellipsoid or spatial reference system of the coordinates.
Location	CoordinateSource	A list of sources or devices used to obtain the coordinates.
Location	SourceAccuracy	The accuracy in meters of the source used to obtain the coordinates, as a component of the global uncertainty.

Figure 2.

The 21 fields (= categories) of the template file, which contain the information requested from specialists. Each category is associated with a DwC class for reference. Note that fields do not always match the DwC terms, since the file is only used to collect data.

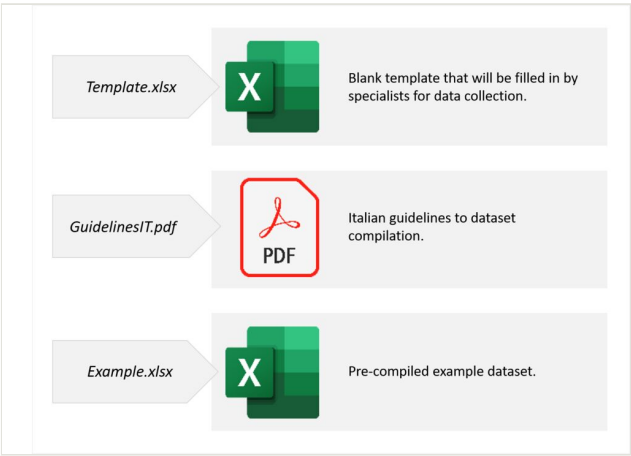

Figure 3.

The three files used for data collection.