Biodiv. Res. Conserv. 65: 35-48, 2022

BRC
www.brc.amu.edu.pl

DOI 10.2478/biorc-2022-0001

Submitted 01.02.2022, Accepted 24.03.2022

The Adam Mickiewicz University Nature Collections IT system (AMUNATCOLL): metadata structure, database and operational procedures

Marcin Lawenda[1]*, Justyna Wiland-Szymańska[2], Maciej M. Nowak[2,3], Damian Jędrasiak[1] & Bogdan Jackowiak[2,4]

[1]Poznań Supercomputing and Networking Center Jana Pawła II 10, 61-139 Poznań, Poland; ORCID: ML https://orcid.org/0000-0003-4844-3655; DJ https://orcid.org/0000-0002-4812-2862

[2]Department of Systematic and Environmental Botany, Faculty of Biology of Adam Mickiewicz University in Poznań, Uniwersytetu Poznańskiego 6, 61-614 Poznań, Poland; ORCID: JWSz https://orcid.org/0000-0002-1916-7201; MMN https://orcid.org/0000-0003-0005-196X**;** BJ https://orcid.org/0000-0003-1684-7380

[3]Biological Spatial Information Laboratory, Faculty of Biology of Adam Mickiewicz University in Poznań, Uniwersytetu Poznańskiego 6, 61-614 Poznań, Poland

[4]Biodiversity Digitization Laboratory, Faculty of Biology of Adam Mickiewicz University in Poznań, Uniwersytetu Poznańskiego 6, 61-614 Poznań, Poland

* corresponding author (e-mail: lawenda@man.poznan.pl)

Abstract. This paper describes the procedures and operational aspects related to the proper storage and handling of taxonomic, biogeographic and ecological data of biological specimens digitised under the AMUNATCOLL project. In the introductory phase of this process, the definition of the metadata is carried out, which is the formal handler of the structure, based upon the analysis of existing standards. The set of parameters derived from the standard is extended by data that is important according to the point of view of the specificity and functionality of the developed system. Subsequently, the database, as a key element in many IT systems, must be set up for data storage along with the suitable structure that reinforces efficiency. The process of preparing and casting a large amount of data requires automated procedures with dedicated tools attached. These approaches address a variety of processes starting from data preparation, where occasionally conversion must occur, aggregation and finally validation, which guarantees that data apply defined rules. Above all, dedicated operational procedures must be defined and applied to enable proper handling of the entire process.

Key words: biodiversity data standards, ABCD standard, Darwin Core, database structure, conversion, validation, IT system, import, iconography, backups

1. Introduction

Biological institutes and museums of natural history around the world curate large collections, including type specimens. Their uniqueness, as well as the issues of physical access to and the protection of artefacts constitute an argument for their digitisation, even though there is remote accessibility for a wider group of beneficiaries interested in biodiversity. The material that is formed out of these resources feeds digital databases. With available and correct data at their disposal, nature conservation managers, policy-makers, scientists and local communities can consciously make decisions for managing natural resources (Ballard *et al.* 2017). Digitisation revives the Natural History Collections (NHCs) documenting the biota of the Earth, which contributes to a better understanding and protection of biodiversity (Butchart *et al.* 2012; Baldwin & Fouch 2018). However, the process of digitising biological data is a great challenge considering the size of the collection, the complexity of the descriptions, the visual representation, and the appealing demonstration of the data to the recipient (Kays *et al.* 2020; Whitlock 2011).

Currently, many projects that face the digitisation process of the collections meet the analogous challenges related to the aspects of proper specimen

© 2022, Lawenda et al. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivs License (http://creativecommons.org/licenses/by-nc-nd/3.0/)

description with metadata and the data storing rules that assure further efficient handling and facilitation of the overall process flow, which is reinforced by application solutions (Beaman *et al.* 2004; Torres *et al.* 2006). The dynamic progress in biodiversity science is related to the development of the language and the standards for the description of specimens, samples and other forms of analogue documentation. This plays a particularly important role in the beginning of the digitisation process when the determination of the information scope becomes the su bject of this process (Silva da *et al.* 2014). Moreover, this design phase heavily affects the later functionality and usability of the entire system (Feest *et al.* 2011). To reduce data discovery barriers, the collected information should be organised to reflect the most important data groups and characteristics of entities properly depicting the digitised objects (Walls *et al.* 2014). Parameters and types of data are important information categories that considerably influence effective data management. Wrongly defined metadata structures hinder the retrieval and exploration of data, resulting in insufficient support for scientists during their work (Kacprzak *et al.* 2018; Löffler *et al.* 2021). Due to the complexity of the process of defining the metadata structure, a recommended way to avoid common mistakes is to follow the path given the world biodiversity metadata standards. There are two main standards for biodiversity informatics that are well-acknowledged and used by the largest networks: Darwin Core (DwC 2021) and Access to Biological Collections Data (ABCD 2021). Following its specification, it should be noted that the most important aspects of the description of the specimen's characteristics, along with their grouping, have already been outlined in a way that corresponds to the needs of the majority of collections. Basically, they cover areas related to the taxonomic description of specimens, their specification, spatial attributes, description of related multimedia files or references to the sources of information. The use of such a standard also facilitates the subsequent interoperability of systems of similar purpose, which considerably contributes to increasing their possibilities in the area of data mining and analysis (Hardisty *et al.* 2019). The amount of data necessary to process during a typical digitisation process requires automatisation by using specially developed procedures and tools. They cover the rules for processing information in the form of hard copy records, which are transferred to a predefined format and a subsequent verification of the prepared data. These tools are coupled with database systems that have the capability to incorporate this data and manage it effectively (Zalani *et al.* 2019).

The purpose of this paper is to present the key components of the AMUNATCOLL IT system for biodiversity data organisation, storage and processing. The metadata and the database structure, as well as applications and procedures are used in the data flow. The work has been conducted within the AMUNATCOLL project, which is co-financed by the European Union from the European Regional Development Fund under the Operational Program Digital Poland (ANC 2022; Jackowiak *et al.* 2022). The overall goal is to make information on the natural collections at the Faculty of Biology, Adam Mickiewicz University (FBAMU 2022) openly available to all interested parties. The specific project aim is to develop a standardised set of metadata that enables the registration of unique NHC values to be compatible with international standards. On this basis, the design and development of a digital database with the structure corresponding to the metadata was carried out in the implementation phase. In addition, the portal (ANCPortal 2022) and the mobile application (ANDR 2022)were developed and provided access for data exploration in a way that was convenient for groups of stakeholders with different requirements.

2. Metodological assumptions

The basis for developing the methods and the mechanisms of the data storage system is the proper understanding of the needs of target groups and the technological solutions that ensure the required functionality and compliance with the applicable solutions. In this study, they relate to metadata, database and supporting operational procedures.

It is of utmost importance to recognise and use the standards for handling biodiversity data under IT system implementation. The standards allow us to follow common rules in the description of specimens, and due to their specificity, affect the database design and layout of the information management system. Moreover, they facilitate common and repeated use of data and enable following similar rules in various applications and processing methods, improving usability. Finally, the standards define the protocols for the data exchange process between compatible entities casting information accordingly.

2.1. Metadata

The metadata specification developed within the AMUNATCOLL project is based on the international standards of Darwin Core and ABCD and simultaneously considers the project-specific assumptions (Jackowiak *et al.* 2022).

Darwin Core (DwC), which is an extension of the Dublin Core standard (DCMI 2022) oriented on biodiversity. It is mainly used for providing a stable reference in distribution data on biological diversity

by describing natural history collections hosted at museums and scientific institutions in a standardised way. It facilitates storing and exchanging information about specific taxon, its geographic occurrence in nature (observation) or in collection (e.g., herbarium voucher, bottled specimen samples). The standard has a glossary of parameters specifying discovery, acquisition, location, time of obtention, supporting evidence (scan or picture), reference to bibliography, commentary, etc. The semantic definitions of specimens defined in DwC facilitate their better use in a variety of contexts. In terms of definitions, Darwin Core does not specify data types and constraints; it relies on recommendations to property values that can be limited, e.g., by a specific vocabulary. The most recent version of the Darwin Core Standard was released in 2009 and consists of 169 defined terms (DwCS 2009).

ABCD is a schema that was designed for describing taxon occurrence and data exchange in a highly structured way supporting the diversity of databases and compatibility with other data standards. ABCD is considered an evolution of the Darwin Core, having a more refined structure and containing approximately 1200 elements. Processing of the ABCD information is facilitated by compliance with eXtensible Mark-up Language (XML 2016), which is maintained by a variety of applications. The current major version, widely adopted and endorsed by Biodiversity Information Standards (BIS 2021), is 2.06. ABCD is composed of three main concepts: mandatory, highly recommended and commonly used. The mandatory concept outlines parameters that are compulsory for creating valid ABCD documents. It is related to the dataset description and pertains to aspects, such as contact with the people responsible for the content and the technical issues and the metadata general description, while also specifying the institution from which the data comes from. The highly recommended concept describes a common entry point to the data. The data contains information about licence, language, units with details on identification, locality and URI. Commonly used concepts encompass more data on specimen descriptions, e.g., the kind of unit, the collector number, sex, age, link to multimedia, and log of editing.

Both the ABCD and Darwin Core are compatible. Thus, the corresponding mapping for data exchange purposes can be arranged. They are also widely utilised in the operational work of biodiversity networks (Güntsch *et al.* 2007), such as the Global Biodiversity Information Facility (GBIF 2021) and BioCASe (2021). Another important aspect of the abovementioned standards is the adherence to the FAIR principles (FAIR 2022). Operations, such as finding, accessing, interoperating and reusing data are facilitated by openness, widespread use and a set of clear rules.

2.2. Database

The definition and the design of the database is the consecutive aspect after the metadata structure. Despite standardisation at the metadata level, the structure of a database does not fully correspond to the representation of its full dataset. In individual biodiversity systems, it reflects the specific needs and requirements underlying their definition. The need to adapt the scope and the structure of the stored data to the offered functionality in a way that ensures the highest efficiency frequently translates to relying directly on popular database management systems (DBMS), such as MySQL (2022), PostgreSQL (2021), Microsoft SQL Server (MSSQL 2022), and Oracle Database (Oracle 2022). The serving of multimedia files is facilitated by the Multimedia Database Management System (MMDBMS 2022) that is built upon the DBMS and offers additional functionality, such as integration, data independence, concurrency control, persistence, privacy integrity control, recovery and query control. A special category of MMDBMS is the Digital Library (DL 2022) exemplified by the dLibra system (dLibra 2021) used for development in the AMUNATCOLL project.

2.3. Operational procedures

A similar approach adjusting to the requirements is applied for the implementation of the operational procedures. These are used for facilitating data preparation, validation, casting to the database tables and securing multimedia files. Moreover, they can be designed in the form of regulations (e.g., explaining the who, what and when is implemented in the digitisation procedure) and the accompanying set of tools for conducting this process separately or in a workflow by its automation. The implementation method depends on the dataset specificity (e.g., the form of source materials being digitised), as well as the decisions and preferences of the developers themselves. Therefore, specific programming languages, libraries and frameworks cannot be named here just by indicating the only correct solution.

3. Metadata specification

The ultimate purpose of the metadata definition task is to prepare the structure that links to the need for formal characterisation of the unique resources of the natural collections gathered at the Faculty of Biology of Adam Mickiewicz University in Poznań (Jackowiak *et al*. 2022). Considering the high development complexity of the database dealt with, the proposed formalisation enables the unambiguous elaboration of essential information in line with international standards. Nevertheless, despite the will to comply with the standard, to meet specific requirements addressed by the project and the system specification itself, a regular set of features

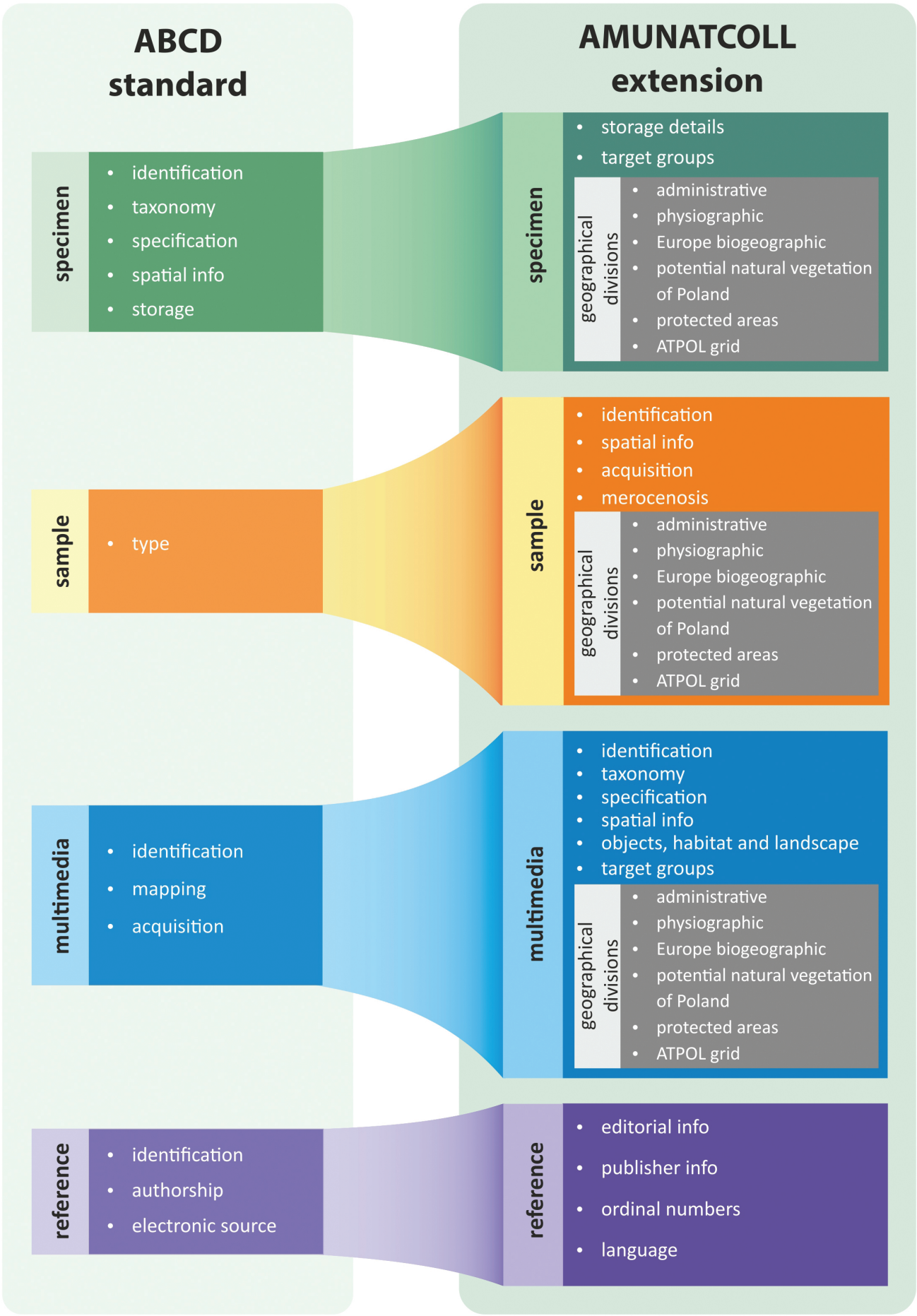


Fig. 1. Overview of the AMUNATCOLL metadata structure

offered by ABCD were extended by a number of non-standard fields, such as those related to protected areas. Moreover, the final shape of the proposed structure was uttered by the necessities of a wide and diverse group of recipients, e.g., state administration units and the local government, state services and officials, teachers, students, research workers, doctoral students and nongovernmental organisations. This resulted in a definition of dedicated fields where specific knowledge conforming to those areas was kept. For the system definition, the designed metadata scheme ensures proper linking to the base record with respect to the illustrations and bibliography.

The structure of the metadata is expressed in the four main metadata sections: specimen, sample, multimedia and reference (Fig. 1). In total, it constitutes 212 fields. Detailed specification of the fields that are the basis for their creation goes beyond the scope of this publication. Only a synthetic overview of the extensions that were implemented in the AMUNATCOLL database in relation to the ABCD standard are more thoroughly discussed.

A crucial facet in the process of the metadata structure definition was the generation of the identification number (ID) of the described object. The number unequivocally identifies a single specimen, an individual in the observation or a sample. Multimedia files, such as scans or photographs, are identified twofold. Herbarium scans of specimens were marked by the same number as the specimen itself. It enabled joining a picture with the corresponding item under presentation in the portal. The different approach was used for the photos that did not represent the specimens stored in the NHCs (e.g., landscape, habitats), for which a unique ID was generated.

In the next step, criteria necessary to quantify the taxonomic information were identified. The table of taxonomic specimens containing the attributes necessary to describe plants, fungi and animals was translated into an extensive structure. This approach resulted in defining individual taxonomic attributes allocated to separate database columns. Another identified challenge was the alignment of the metadata structure for taxonomic synonyms for species with multiple names in particular. A special case considered the taxonomic names appearing on the herbarium sheets are no longer valid today. This was solved by implementing a synonym list containing these names.

The specimen section provides supplementary information detailing the storage of specimens up to the indication of a specific rack and shelf. This makes it easier to find the specimen in case of a need to reach a physical object. In dealing with specimens originating from protected areas (e.g., national and landscape parks, nature reserves, Natura 2000 area), such information is recorded in specially designated fields. Information related to or belonging to specific geographical divisions (administrative, physiographic, European biographic and potential natural vegetation of Poland) constitute a separate group of information increasing the data value in the context of the functioning groups of organisms and their interdependencies. In addition, information on the Distribution Atlas of Vascular Plants in Poland (ATPOL) grid system (Zając 1978) as a regional spatial division system of Poland and suitability for target groups is supplemented.

The sample data section is entirely an additional group of information to ABCD on the specification of the unexplored and unmarked material (e.g., soil sample, timber, nest) from which the selected specimens are derived. Once marked, they are transferred to the "specimen" group. The attributes of the samples are characterised by the specificity that requires the definition of a separate form where parameters, such as merocenosis (considered organisms colonizing the merotope) or groups of animals marked from a given sample can be specified. Moreover, the data include identification (required to create a new record in the database), spatial references, protected areas, the ATPOL grid system, information about gathering, merocenosis data and belonging to geographical divisions, which are similar to specimens.

Spatial data on specimens represent another important group of metadata. Based on the original description of the collection or the observation place, data on geographic coordinates were elaborated. However, a text description of the location is also stored. Occasionally, the original location description differs in the level of detail and readability; therefore, the data was enriched by corresponding comments and information of geotagging quality (Nowak *et al.* 2021). The first contained information on the selection of geographic coordinates for a location where the original description was geographically ambiguous, illegible or incomplete. The second one was utilised at the portal level to facilitate the selection of records for joint specimen analysis at the same spatial accuracy. The metadata group holds information on continent and country, as well as on the altitude above sea level. Another section of the spatial metadata enclosed particulars on an affiliation of records to the protected nature areas, administrative division, regional geographic and biogeographic divisions, or a location in relation to the UTM (2022) and ATPOL grid fields (Zając 1978). Considering this type of metadata allows for more spatially precise searches of records in the database and provides a more accurate survey of the specimen distribution depending on the characteristics of the terrain.

In a different way from the ABCD standard, the multimedia data in the AMUNATCOLL database

constitutes a valuable source of information about landscapes, habitats and objects present there. As such, these objects are subject to an identification process similar to that conducted for specimens. This means preparation of a full set of data, including identification, taxonomy of objects located there, their detailed description, spatial references, protected areas, ATPOL grid system, description of the habitat, target groups and geographical divisions.

Data on references, understood as information about publications where specimens from the AMUNATCOLL database are described, have been extended with a set of parameters describing them in the context of a specific publication type, including a journal, monography or book chapter. This enables the addition of extra details about volume number, issue, pages, publisher, publication location, and book title for a chapter. Bibliographic data are kept in a different table. Each item receives its individual number, which is then used to create a link to an NHC object referred to in a publication.

Additional value was obtained by defining the "Comments" field in several metadata sections. It permitted the inclusion of additional information that, for formal reasons, could not be entered into other attributes of a corresponding section. For example, in the case of multimedia not associated with a specific specimen entity, the field enables us to describe the picture content (animate and inanimate objects, surroundings, environmental conditions) in a flexible way, which is essential for understanding the content.

The metadata specification also includes information, such as the allowed range of values for a given attribute and the connections between the fields that define the conditionality of their use. It facilitates the process of data preparation and the input into the database by several dozen people, which usually leads to numerous mistakes. The most common mistakes were related to the wrong date format, the use of illegal abbreviations and the use of values that at the moment could not be used in connection with the values of other fields.

Finally, it is worth mentioning that in addition to specimen descriptive fields, supplementary fields are also introduced in the database, which facilitates the process of collection management, introducing extra functionality to the portal.

4. Database structure

The database is a key element of the AMUNATCOLL IT system, where all metadata are stored; therefore, much time and attention were devoted to its efficient implementation. It has been designed to store all 212 metadata fields describing specimen specifics and the organisational, technical and support fields to ensure an effective way to access and manage them.

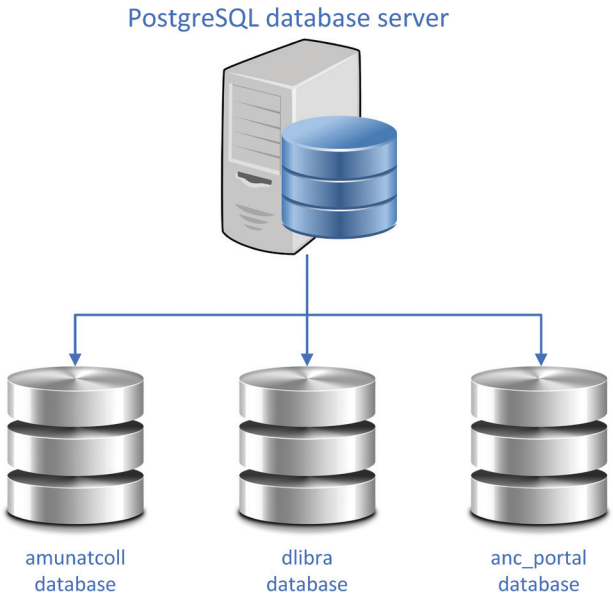


Fig. 2. The logical structure of the AMUNATCOLL database

The implementation was based on the PostgreSQL server (PostgreSQL 2021). It is a free, open-source relational database management system that emphasises extensibility and SQL compliance. It has the functionality required by the specifics of the project, such as ACID (Atomicity, Consistency, Isolation, Durability) properties, automatically updated views, foreign key triggers and stored procedures. PostgreSQL is designed to handle heavy workloads such as data warehouses or web services with multiple concurrent users. The database storage structure is divided into three logical databases (Fig. 2). Each database contains data related to a specific operational part of the project. The "*amunatcoll*" database is associated with the storage of specimen information, the "*dlibra*" database houses the multimedia data, and the "*anc_portal*" database is for the portal operational tasks.

The "*amunatcoll*" database corresponds to storing data related to specimens and contains tables with information about specimen taxonomic names and their synonyms, samples, bibliographic data about publications, kinds of specimens, collections and subcollections of specimens, statistics, and specimen import history. The "*dlibra*" database is the database of the dLibra digital library, and its task is to store data on imported multimedia objects. From all the tables available in the library, the project uses the data contained in the tables storing information about metadata related to multimedia files and metadata of publication scans. In this database, the attributes and their values are stored in two columns common to all attributes. Contrary to the "*amunatcoll*" database, the attributes

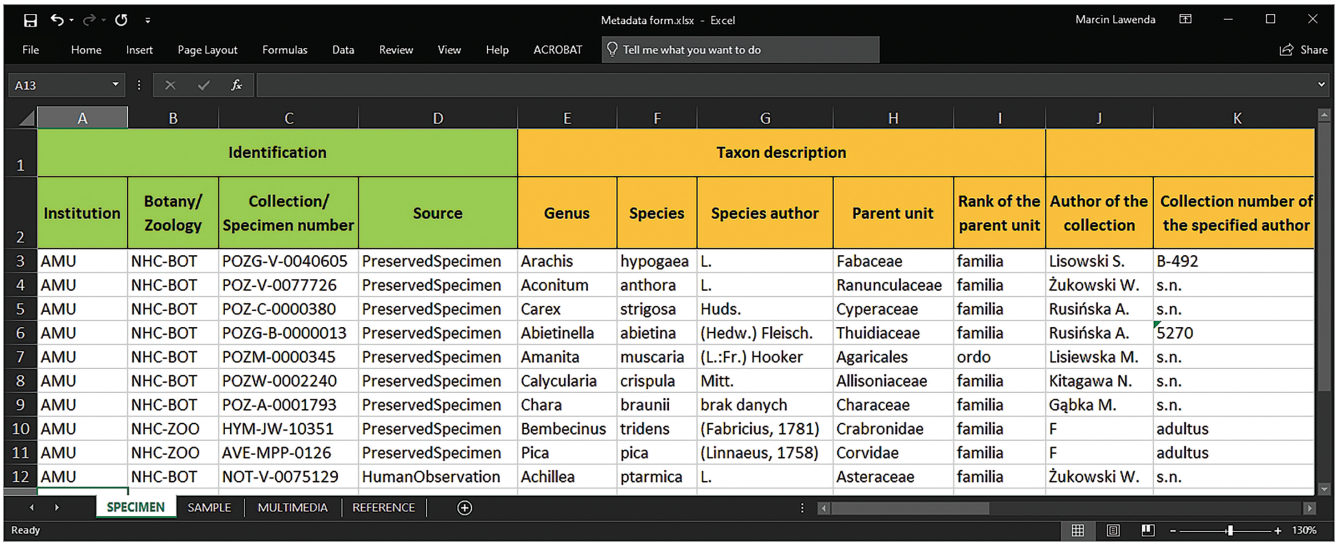
do not have dedicated columns to store corresponding parameters in a separate way. The "*anc_portal*" database is accountable for storing information related to the portal and its functions. It contains tables storing information about user data, user permissions, resources created by users from the mobile application (projects, observations and their files), other resources created by users (albums, filters, base maps), teams and their members, and visiting statistics.

5. The data transfer procedures

To improve the process of data preparation and import, several applications and solutions have been developed to make it viable. They help check the correctness of the information and speed up the procedures by performing mass operations on the data.

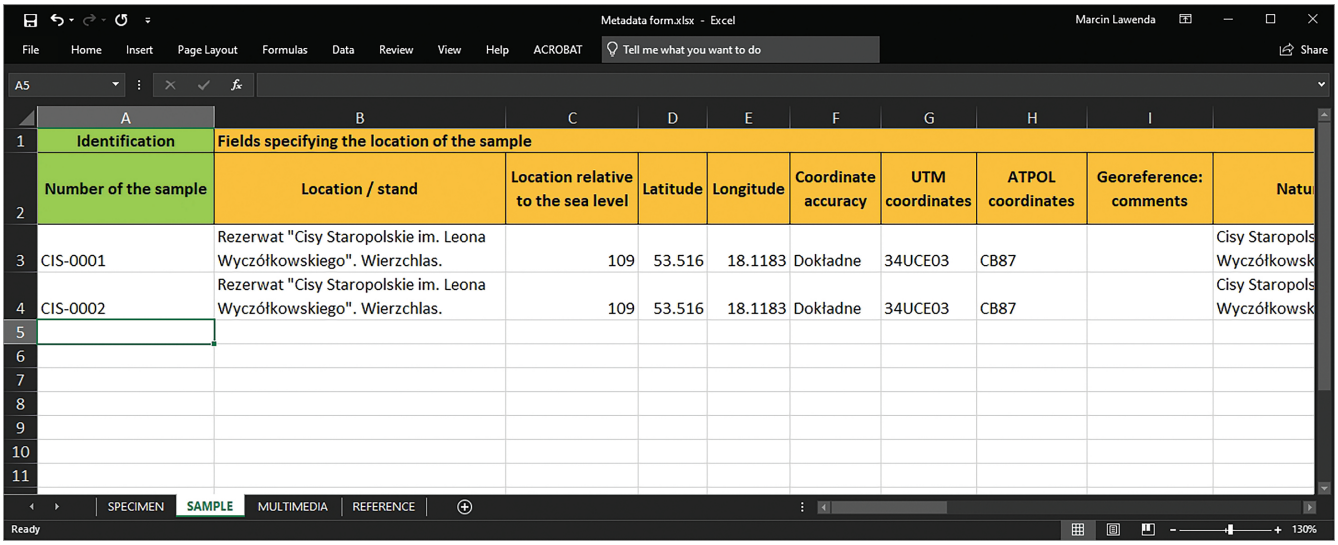
5.1. Data preparation

The programs and applications described below use a metadata specification-based form as their primary input format. To maintain a common format and avoid inconsistencies between the data provided by different contributors, an Excel spreadsheet form was designed (Figs. 3-6). It contains all the necessary attributes according to the metadata specification, divided into sheets including specimens, samples, iconography and bibliographic entries. The persons responsible for preparing the form enter the data on the labels of the digitised objects into specific fields. It should be noted that the fields are divided into obligatory and optional. In addition, there are defined dependencies between the selected fields, which require the editor to adhere to certain rules. All principles are described in the internal guidance book provided to interested parties.



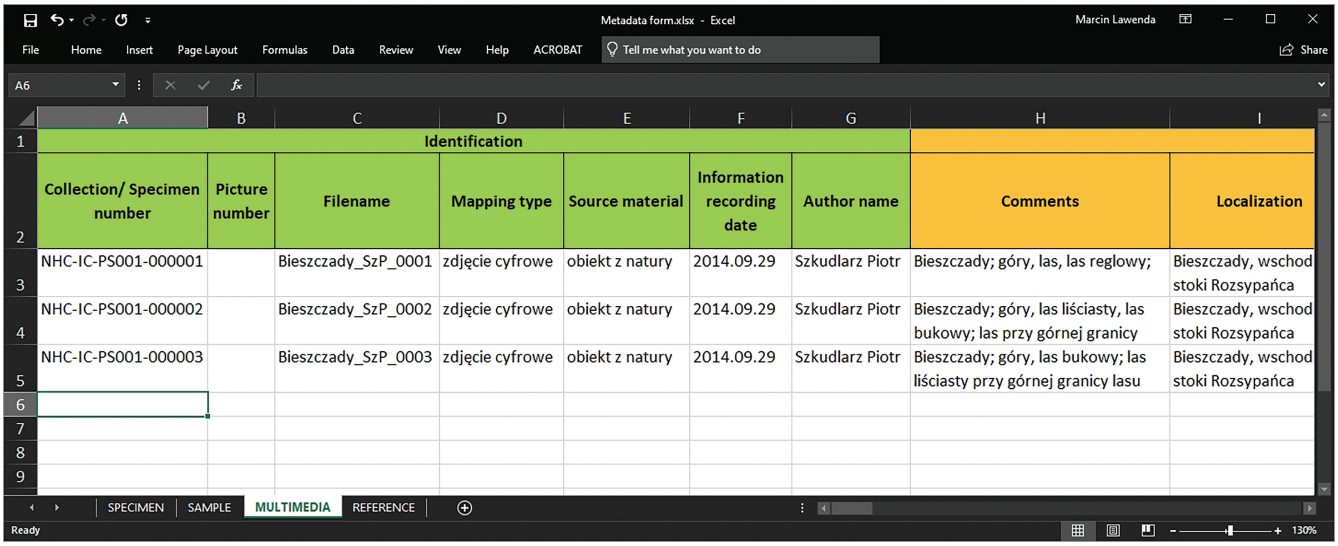
	A	B	C	D	E	F	G	H	I	J	K
1	Identification				Taxon description						
2	Institution	Botany/ Zoology	Collection/ Specimen number	Source	Genus	Species	Species author	Parent unit	Rank of the parent unit	Author of the collection	Collection number of the specified author
3	AMU	NHC-BOT	POZG-V-0040605	PreservedSpecimen	Arachis	hypogaea	L.	Fabaceae	familia	Lisowski S.	B-492
4	AMU	NHC-BOT	POZ-V-0077726	PreservedSpecimen	Aconitum	anthora	L.	Ranunculaceae	familia	Żukowski W.	s.n.
5	AMU	NHC-BOT	POZ-C-0000380	PreservedSpecimen	Carex	strigosa	Huds.	Cyperaceae	familia	Rusińska A.	s.n.
6	AMU	NHC-BOT	POZG-B-0000013	PreservedSpecimen	Abietinella	abietina	(Hedw.) Fleisch.	Thuidiaceae	familia	Rusińska A.	5270
7	AMU	NHC-BOT	POZM-0000345	PreservedSpecimen	Amanita	muscaria	(L.:Fr.) Hooker	Agaricales	ordo	Lisiewska M.	s.n.
8	AMU	NHC-BOT	POZW-0002240	PreservedSpecimen	Calycularia	crispula	Mitt.	Allisoniaceae	familia	Kitagawa N.	s.n.
9	AMU	NHC-BOT	POZ-A-0001793	PreservedSpecimen	Chara	braunii	brak danych	Characeae	familia	Gąbka M.	s.n.
10	AMU	NHC-ZOO	HYM-JW-10351	PreservedSpecimen	Bembecinus	tridens	(Fabricius, 1781)	Crabronidae	familia	F	adultus
11	AMU	NHC-ZOO	AVE-MPP-0126	PreservedSpecimen	Pica	pica	(Linnaeus, 1758)	Corvidae	familia	F	adultus
12	AMU	NHC-BOT	NOT-V-0075129	HumanObservation	Achillea	ptarmica	L.	Asteraceae	familia	Żukowski W.	s.n.

Fig. 3. The record form for the specimen description



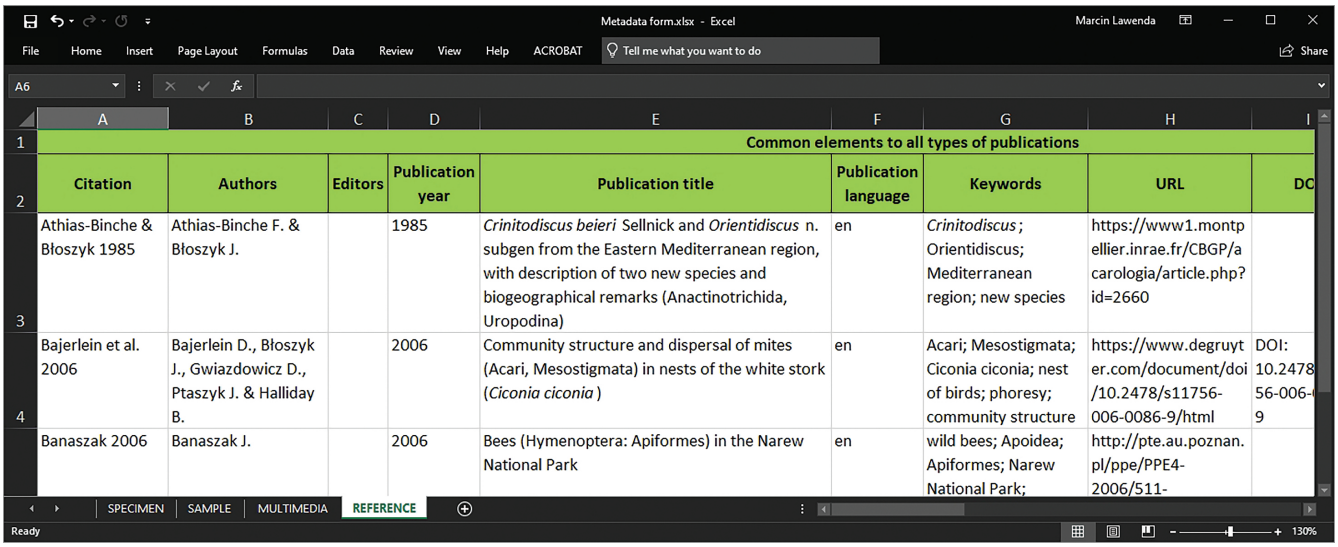
	A	B	C	D	E	F	G	H	I
1	Identification	Fields specifying the location of the sample							
2	Number of the sample	Location / stand	Location relative to the sea level	Latitude	Longitude	Coordinate accuracy	UTM coordinates	ATPOL coordinates	Georeference: comments
3	CIS-0001	Rezerwat "Cisy Staropolskie im. Leona Wyczółkowskiego". Wierzchlas.	109	53.516	18.1183	Dokładne	34UCE03	CB87	
4	CIS-0002	Rezerwat "Cisy Staropolskie im. Leona Wyczółkowskiego". Wierzchlas.	109	53.516	18.1183	Dokładne	34UCE03	CB87	
5									
6									
7									
8									
9									
10									
11									

Fig. 4. The record form for the sample description



	A	B	C	D	E	F	G	H	I
1	Identification								
2	Collection/ Specimen number	Picture number	Filename	Mapping type	Source material	Information recording date	Author name	Comments	Localization
3	NHC-IC-PS001-000001		Bieszczady_SzP_0001	zdjęcie cyfrowe	obiekt z natury	2014.09.29	Szkudlarz Piotr	Bieszczady; góry, las, las reglowy;	Bieszczady, wschodnie stoki Rozsypańca
4	NHC-IC-PS001-000002		Bieszczady_SzP_0002	zdjęcie cyfrowe	obiekt z natury	2014.09.29	Szkudlarz Piotr	Bieszczady; góry, las liściasty, las bukowy; las przy górnej granicy	Bieszczady, wschodnie stoki Rozsypańca
5	NHC-IC-PS001-000003		Bieszczady_SzP_0003	zdjęcie cyfrowe	obiekt z natury	2014.09.29	Szkudlarz Piotr	Bieszczady; góry, las bukowy; las liściasty przy górnej granicy lasu	Bieszczady, wschodnie stoki Rozsypańca
6									
7									
8									
9									

Fig. 5. The record form for the multimedia description



	A	B	C	D	E	F	G	H	I
1	Common elements to all types of publications								
2	Citation	Authors	Editors	Publication year	Publication title	Publication language	Keywords	URL	DO
3	Athias-Binche & Błoszyk 1985	Athias-Binche F. & Błoszyk J.		1985	*Crinitodiscus beieri* Sellnick and *Orientidiscus* n. subgen from the Eastern Mediterranean region, with description of two new species and biogeographical remarks (Anactinotrichida, Uropodina)	en	*Crinitodiscus*; Orientidiscus; Mediterranean region; new species	https://www1.montpellier.inrae.fr/CBGP/acarologia/article.php?id=2660	
4	Bajerlein et al. 2006	Bajerlein D., Błoszyk J., Gwiazdowicz D., Ptaszyk J. & Halliday B.		2006	Community structure and dispersal of mites (Acari, Mesostigmata) in nests of the white stork (*Ciconia ciconia*)	en	Acari; Mesostigmata; Ciconia ciconia; nest of birds; phoresy; community structure	https://www.degruyter.com/document/doi/10.2478/s11756-006-0086-9/html	DOI: 10.2478/s11756-006-0086-9
	Banaszak 2006	Banaszak J.		2006	Bees (Hymenoptera: Apiformes) in the Narew National Park	en	wild bees; Apoidea; Apiformes; Narew National Park;	http://pte.au.poznan.pl/ppe/PPE4-2006/511-	

Fig. 6. The record form for the reference description

5.2. Converter

The purpose of another tool named "Converter" is facilitating the process of converting existing specimen descriptions prepared beforehand (often many years before) to a new format agreed upon per the project design work. The converter, in the form of a web application, automatically changes the format of existing files according to the developed scheme. Input files are converted into files conforming to the metadata specification using the specified configuration (rule set), allowing for later optional editing of the input data. Configurations are divided into two types:

Standard – the conversion creates missing columns and form sheets, sorts them in the order identical to the specification, and changes the cell value formatting to the text to ensure data correctness in the import process. Custom – in this case, apart from the operations performed in the standard mode, additional operations (custom defined upon user request) are conducted on the input data to facilitate the preparation of data in the older format for import into the database, which could be changing the date format, combining or splitting columns, etc.

In case of errors during the conversion or when detecting any irregularities in the file, the converter displays a detailed report. Below, the main page of the tool is presented (Fig. 7), where the user specifies the file for the conversion and the configuration scheme.

5.3. Validator

Validator is a tool for checking (validating) Excel files, which is available in the form of a web application. Its main purpose is to analyse the correctness of the data delivered to the form, considering the context of the sheet (specimen, sample, iconography and bibliography) and field. Moreover, it verifies the presence of

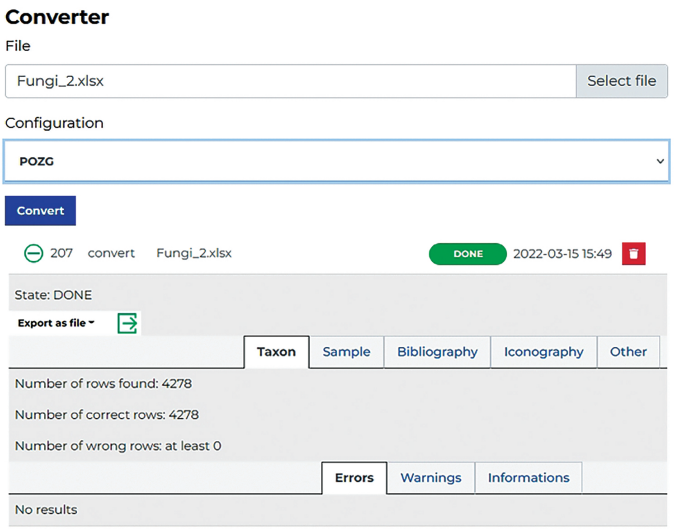


Fig. 7. The interface of the web application converter

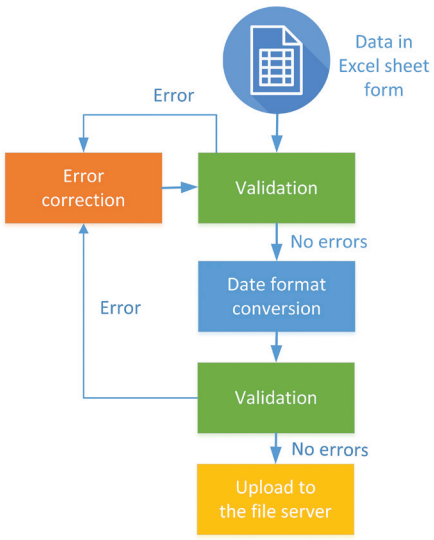


Fig. 8. Procedure for preparing and checking the file for import

mandatory fields, duplicates, and provides in-between relations with specific fields. The tool returns a report detailing errors by occurrence and description, which is then broken down into each detected sheet. The tool along with the converter is an indispensable part of the import procedure presented in the figure below (Fig. 8). It must be noted that the validator is also utilised during the automatic import process to revalidate the data, which prevents information that does not meet the assumptions appear in the database. The interface page of the tool is shown below (Fig. 9).

5.4. Aggregator

The program "Aggregator" is used to combine XLSX files that are consistent with the AMUNATCOLL specification. It is used in the digitisation process, mainly during the stage of describing records by a georeferencing team or a translation team. Due to the specifics of the data preparation process, it results in many relatively small files with information about specimens. Information added at a later stage is often repeated for many records from the same collection. It is evident that it is easier to work with large datasets by properly sorting and filtering them. The aggregation program combines many smaller files into a one bigger file or set of files with a given maximum size whenever it is exceeded. The complimentary operation is performing the basic file format verification (e.g., name of headers, duplicated records). Since "Aggregator" is a command line application, the communication with it takes place via the command line and the defining parameters. The first argument just after the program name is the location of the directory, where the input files are placed. The second argument is optional and specifies the maximum number of records included in one output file (the default is 10000). The result of the program is a set of XLSX files with the name as the name of the directory being processed, with the word "_combined" as the suffix, the type ('SPECIMEN', 'SAMPLE' or 'MULTIMEDIA') and the sequence number of the file.

5.5. Reporter

The application "Reporter" is used to summarise the number of records in XLSX files that are consistent with the AMUNATCOLL specification and it is used by project coordinators for the purpose of the ongoing monitoring of the progress of digitisation work.

Similar to the "Aggregator" application, "Reporter" is a command line solution. The argument to run the application is the name of the directory where the input files are located. The result of the program operation

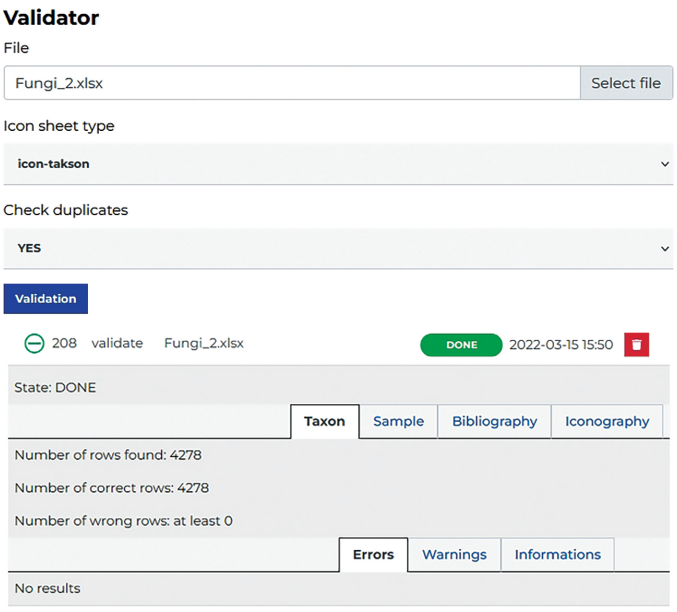


Fig. 9. The validator web application form

is the information about the number of unique records in the categories 'SPECIMEN', 'SAMPLE' or 'MULTIMEDIA.' It also displays information about errors found in the files, such as an incorrect header or duplicate records.

6. The descriptive data import process

The overall purpose of the import process is to incorporate data assembled by the digitalisation team into the database structure following consistency defined by metadata specification. The benefit of this process is obtaining easier access to digitised resources by making them available for external systems, e.g., portals, mobile applications or cooperating external databases. The process of importing data to the database is carried out in three steps. The first is the data preparation by a person with specific privileges called the *data administrator*. Second, files are checked against compliance with the documentation (this time in an automated process), and the data is placed in the database. Finally, the files containing specimen data are sent to the definite directory on the cloud disk (data buffer).

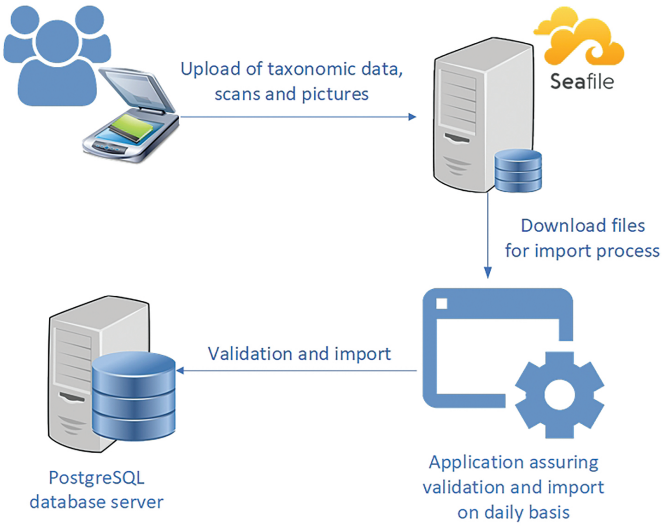


Fig. 10. Diagram showing the import process of the descriptive data, scans and pictures

The data import procedure starts automatically at a programmed time of the day, provided that changes in the input files compared to the previous import are detected under the scan process (Fig. 10). Newly provided records are validated, and in case of success (data conforming to the specification), they are placed in the database. After completion processing of all files, the statistical data related to the imported records are recalculated (updated), which allows for the prompt observation of new statistics on the portal. A report summarising the entire process is sent by an e-mail. It also contains attachments describing the details of errors found in the input data and a list of imported files with record statistics.

The data buffer import procedure described above was used successfully in almost the entire duration of the project. In the final phase, after users' comments, a much more convenient way of importing with the use of the portal interface was developed. It is described in more detail in the publication dealing with the interfaces implemented in the AMUNATCOLL project (Nowak *et al*. 2022).

7. The iconographic data import process

The iconographic data import process, to some extent, resembles that of species data. In this case, metadata are accompanied by multimedia (scans, pictures), and from the import perspective, they are one. That means that if one of the two (a record in the file or a multimedia file) is missing, the import will not begin. In the first step of importing iconographic data, the files are placed in predefined directories in the data buffer, and then they are imported to the dLibra digital library. The import is carried out at the level of a single directory, and the files in it are processed before being imported. The directory contains a metadata file that describes the attributes of each file in that directory.

7.1. Temporary buffer

A virtual disk, based on the Seafile system, the data buffer (Seafile 2021), is used to store files ready for import. The data that is ready for import is stored in the previously prepared directory structure corresponding to either the collection names with respect to multimedia related to organisms (*icon-taxon*) or to the author names in the case of pictures unrelated to specimens (*icon-varia*), though they present objects in wider contexts, such as habitats or landscapes. The directories in the cloud are linked with the server where the import procedure is carried out.

7.2. The preparation of multimedia files

Before importing, multimedia files are processed, which allows them to be made eligible for the AMUNATCOLL system (preparation in terms of design assumptions and requirements related to the software used) and to protect them against unauthorised use. In the case of still images, files are subjected to the following operations:

- If the picture is delivered in the CR2 (RAW) format, conversion to the “TIFF” format is applied.
- If a photo has nonstandard orientation information, it is rotated.

- A photograph is protected by the following operations: cropping (frame of several pixels), population of EXIF data (EXIF 2021), at the end, a watermark and a hologram are implied.
- The photograph is saved as a pyramidal TIFF to ensure a more efficient serving of data to the portal.
- A thumbnail is generated for the prepared photograph.

7.3. Software

The AMUNATCOLL project is highly demanding in terms of the number and size of processed images. It is enough to mention that hundreds of thousands of images with a capacity of hundreds of terabytes were developed during the implementation of the design tasks. Therefore, due diligence was taken in selecting the appropriate software. As a result, the following open-source software was selected and used for graphic manipulation in the importing process, processing and storing multimedia files: libvips (2021) and ImageMagick (ImMag 2021). Libvips is capable of processing large pictures with small memory utilisation since it does it piece-by-piece on small rectangles or groups of lines. Moreover, the processing can be conducted in parallel (threads are run on separate cores) mode, which leads to a better use of the full available computing power. AMUNATCOLL uses libvips to create thumbnails and TIFF pyramid files. ImageMagick also prioritises the performance of conducted operations; thus, a multiple threads computation is viable even on tera-pixel image sizes. As a result of a large number of possible operations, such as converting, filtering, and artistic effects, this software is used as an engine of graphic operations in many professional solutions. In the case of AMUNATCOLL, it is primarily used to implement security measures related to the legal protection of shared graphics. It includes the aforementioned cropping of the image border, adding EXIF metadata, applying a visible watermark and a hologram that is visible only after applying the decoding operation. Both applications are able to run on the Linux system and can be called directly from the command line, which is a favour of the AMUNATCOLL infrastructure and processing routine, and it is also crucial for the project prerequisites.

8. Data protection procedures

Digitised resources of the Faculty of Biology of the AMU are very valuable assets resulting from the work of many people for several months of the project lifetime. In the event of their loss as an outcome of a failure of the storage system where the data are intended to be stored, their restoration would be extremely difficult, if not impossible. Therefore, the key issue is to ensure adequate protection measures by implementing procedures that ensure the performance of backups. This allows us to reinstate resources and restore the full functionality of the system. As already mentioned in the AMUNATCOLL system, we deal with two types of data: metadata and graphic data in the form of scans and photos. Therefore, two different approaches were used, which considered the nature of working with the system and the resources available. Metadata describing the specific features of the collected specimens were stored in a database and were especially prepared for this purpose. Two-level data protection is used here. The first level is the registration of changes to individual records. In the case of a record change, its previous version is saved in the database with the annotation describing who and when the change was made. This allows us to restore the contents of the record, e.g., in case of a mistake. Another type of database content protection is to perform regular snapshots of the database. This is done automatically at least once a day at night when modifications have been detected. This way, the different versions of the database are preserved from many months back. A slightly different approach has been designed and implemented for iconographic data. Due to the considerably greater data capacity compared to metadata, the number of copies has been limited to three. The first copy is the data directly used by the AMUNATCOLL system for daily operational work. The second and third are backups run only in case of an emergency. The difference between them is that the second copy is stored on the disk system in the same location as the main copy, while the third is a geographic copy. In the case of the second copy, its physical proximity facilitates the process of possible data recovery after the failure. Locating the third copy on another system makes it resistant to much more serious threats, such as fire or flooding.

9. Conclusions

The creation of a uniform IT system in the AMUNATCOLL project ensures interoperability, communication with other databases, and the possibility of creating applications based on it that allow for multidimensional data analysis, including geoinformatics. Moreover, it allowed for equipping the research team of the project with tools enabling effective work on digitised biological resources and opening up cooperation with other entities implementing projects in the field of biodiversity. It should be noted that the guarantee of the durability of the operation and further development of the system is provisioned by hardware infrastructure and organisational facilities brought by the project partner Poznan Supercomputing and Networking Center (PSNC 2021).

It worth emphasising that the challenges defined in the AMUNATCOLL project, especially in the field of data processing, are among the most demanding. This is due both to a specific quantitative indicator (over two million records), as well as a capacity indicator (hundreds of terabytes of multimedia). This work focuses on the presentation of the essential elements of the AMUNATCOLL system related to the description of digitised objects, their storage both during development and for later sharing for operational purposes, and the preparation and use of dedicated tools for the automation of work supporting effective processing.

The metadata structure needed to store taxonomic data was based on the ABCD standard and extended with fields specific to the project requirements. The design and structure of the database as also individual, addressing the needs related to the storage of descriptive data, as well as information on iconography and supporting the functionality of the portal. Carrying out the digitisation process on such a large scale with the participation of a large team would not have been possible without technological support. For this purpose, a set of tools and infrastructural, application and procedural solutions supporting these activities were developed. Not only logistical issues but also the different nature of the data being processed (concerning specimens or iconography) were taken into account. The need to properly secure the results of the work carried out was done in developing appropriate backup procedures that would allow the system to be properly restored in the event of technical problems.

Acknowledgements. This work has been supported by the AMUNATCOLL project and has been partly funded by the European Union and Ministry of Digital Affairs from the European Regional Development Fund as part of the Digital Poland Operational Program under grant agreement number: POPC.02.03.01-00-0043/18. This paper expresses the opinions of the authors and not necessarily those of the European Commission and Ministry of Digital Affairs. The European Commission and Ministry of Digital Affairs are not liable for any use that may be made of the information contained in this paper.

Author Contributions. All authors contributed to the study design, material preparation, data collection and analysis. The concept of the publication and the first draft of the manuscript were presented by Marcin Lawenda. All authors participated in the development of subsequent versions. Marcin Lawenda and Bogdan Jackowiak read and approved the final manuscript.

References

ABCD. 2021. Access to Biological Collections Data standard. https://www.tdwg.org/standards/abcd/

ANC. 2022. AMUNATCOLL project. http://anc.amu.edu.pl/eng/index.php

ANCPortal. 2022. AMUNATCOLL Portal. https://amunatcoll.pl/

ANDR. 2022. AMUNATCOLL Mobile Application – Android. https://play.google.com/store/apps/details?id=pl.pcss.amunatcoll.mobile

Baldwin R. F. & Fouch N. T. 2018. Understanding the Biodiversity Contributions of Small Protected Areas Presents Many Challenges. Land 7(4): (123). https://doi.org/10.3390/land7040123

Ballard H. L., Robinson L. D., Young A. N., Pauly G. B., Higgins L. M., Johnson R. F. & Tweddle J. C. 2017. Contributions to conservation outcomes by natural history museum-led citizen science: Examining evidence and next steps. Biol Conserv 208: 87-97.

Beaman R., Wieczorek J. & Blum S. 2004. Determining Space from Place for Natural History Collections. D-Lib Magazine 10 (5). https://herbarium.millersville.edu/471/Beaman_et_al_2004.pdf

BioCASe. 2021. Biological Collection Access Service. http://www.biocase.org/

BIS. 2021. Biodiversity Information Standards. https://www.tdwg.org/

Butchart S. H. M., Scharlemann J. P. W., Evans M. I., Quader S., Aricò S., Arinaitwe J., Balman M., Bennun L. A., Bertzky B., Besançon Ch., Boucher T. M., Brooks T. M., Burfield I. J., Burgess N. D., Chan S., Clay R. P., Crosby M. J., Davidson N. C., De Silva N., Devenish Ch., Dutson G. C. L., Día z Fernández D. F., Fishpool L. D. C., Fitzgerald C., Foster M., Heath M. F., Hockings M., Hoffmann M., Knox D., Larsen F. W., Lamoreux J. F., Loucks C., May I., Millett J., Molloy D., Morling P., Parr M., Ricketts T. H., Seddon N., Skolnik B., Stuart S. N., Upgren A. & Woodley S. 2012. Protecting Important Sites for Biodiversity Contributes to Meeting Global Conservation Targets. PLoS ONE 7(3). e32529. https://doi.org/10.1371/journal.pone.0032529

DCMI. 2022. The Dublin Core™ Metadata Initiative. https://www.dublincore.org/

DL. 2022. Digital library. https://en.wikipedia.org/wiki/Digital_library

dLibra. 2021. Digital Library Framework. https://www.psnc.pl/digital-libraries-dlibra-the-most-popular-in-poland/

DwC. 2021. Darwin Core standard. https://dwc.tdwg.org/

DwCS. 2009. Darwin Core Standard terms. http://rs.tdwg.org/dwc/terms.htm

EXIF. 2021. Exchangeable Image File Format. https://en.wikipedia.org/wiki/Exif

FAIR. 2022. FAIR Principles. https://www.go-fair.org/fair-principles/

FBAMU. 2022. Faculty of Biology of the Adam Mickiewicz University in Poznań. http://biologia.amu.edu.pl/

Feest A., van Swaay Ch., Aldred T. D. & Jedamzik K. 2011. The biodiversity quality of butterfly sites: A metadata assessment. Ecol Indic 11(2): 669-675.

GBIF. 2021. Global Biodiversity Information Facility. https://www.gbif.org/en/

Güntsch A., Berendsohn W. G. & Mergen P. 2007. The BioCASE Project – a Biological Collections Access Service for Europe. Ferrantia 51: 103-108.
https://www.researchgate.net/publication/263083477_The_BioCASE_Project_–_a_Biological_Collections_Access_Service_for_Europe

Hardisty A. R., Michener W. K., Agosti D., García E. A., Bastin L., Belbin L., Bowser A., Buttigieg P. L., Canhos D. A. L., Egloff W., De Giovanni R., Figueira R., Groom Q., Guralnick R. P., Hobern D., Hugo W., Koureas D., Ji L., Los W., Manuel J., Manset D., Poelen J., Saarenmaa H., Schigel D., Uhlir P. F. & Kissling W. D. 2019. The Bari Manifesto: An interoperability framework for essential biodiversity variables. Ecol Infor 49: 22-31.

ImMag. 2021. ImageMagick – cross-platform image processing software. https://imagemagick.org/

IOS. 2021. AMUNATCOLL Mobile Application – iOS. https://apps.apple.com/pl/app/amunatcoll/id1523442673

Jackowiak B., Błoszyk J., Dylewska M., Nowak M. M., Szkudlarz P., Lawenda M. & Meyer N. 2022. Digitization and online access to data on natural history collections of Adam Mickiewicz University in Poznan: assumptions and implementation of the AMUNATCOLL project. Biodiv. Res. Conserv. 65: 23-34.

Kacprzak E., Koesten L., Ibáñez L.D., Blount T., Tennison J. & Simperl E. 2018. Characterising dataset search – An analysis of search logs and data requests. J Web Semant 55: 37-55.

Kays R., McShea W. J. & Wikelski M. 2020. Born-digital biodiversity data: Millions and billions. Divers Distrib 26(5): 644-648.

libvips. 2021. libvips – An image processing library. https://www.libvips.org/

Löffler F., Wesp V., König-Ries B. & Klan F. 2021. Dataset search in biodiversity research: Do metadata in data repositories reflect scholarly information needs? PLoS ONE 16(3): e0246099. https://doi.org/10.1371/journal.pone.0246099

MMDBMS. 2022. Multimedia Database Management System. https://en.wikipedia.org/wiki/Multimedia_database

MSSQL. 2022. Microsoft SQL Server. https://www.microsoft.com/en-us/sql-server

MySQL. 2022. MySQL – database management system. https://www.mysql.com/

Nowak M. M., Słupecka K. & Jackowiak B. 2021. Geotagging of natural history collections for reuse in environmental research. Ecol Indic 131: 108131.

Nowak M. M, Lawenda M., Wolniewicz P., Urbaniak M. & Jackowiak B. 2022. The Adam Mickiewicz University Nature Collections IT system (AMUNATCOLL): portal, mobile application and graphical interface. Biodiv. Res. Conserv. 65: 49-67.

Oracle. 2022. Oracle Database. https://www.oracle.com/database/

PostgreSQL. 2021. PostgreSQL database web site. https://www.postgresql.org/

PSNC. 2021. Poznan Supercomputing and Networking Center. https://www.psnc.pl/

Seafile. 2021. Open Source File Sync and Share Software. https://www.seafile.com/en/home/

Silva da J. R., Castro J. A., Ribeiro C., Honrado J., Lomba A. & Gonçalves J. 2014. Beyond INSPIRE: An Ontology for Biodiversity Metadata Records. In: R. Meersman, H. Panetto, A. Mishra, R. Valencia-García, A. L. Soares, I. Ciuciu, F. Ferri, G. Weichhart, T. Moser, M. Bezzi & H. Chan (eds.). OTM 2014 Workshops, LNCS 8842, pp. 597-607. Springer-Verlag Berlin Heidelberg.

Torres R. d. S., Medeiros C. B., Gonçcalves M. A. & Fox E. A. 2006. A digital library framework for biodiversity information systems. International Journal on Digital Libraries 6: 3-17.

UTM. 2022. The Universal Transverse Mercator. https://en.wikipedia.org/wiki/Universal_Transverse_Mercator_coordinate_system

Walls R. L., Deck J., Guralnick R., Baskauf S., Beaman R., Blum S., Bowers S., Buttigieg P. L., Davies N., Endresen D., Gandolfo M. A., Hanner R., Janning A., Krishtalka L., Matsunaga A., Midford P., Morrison N., Tuama É. Ó., Schildhauer M., Smith B., Stucky B. J., Thomer A., Wieczorek J., Whitacre J. & Wooley J. 2014. Semantics in Support of Biodiversity Knowledge Discovery: An Introduction to the Biological

Collections Ontology and Related Ontologies. PLOS ONE 9(3). e89606. https://doi.org/10.1371/journal.pone.0089606

Whitlock M. C. 2011. Data archiving in ecology and evolution: best practices. Trends Ecol Evol 26: 61-65.

XML. 2016. Extensible Markup Language. https://www.w3.org/XML/

Zając A. 1978. Atlas of distribution of vascular plants in Poland (ATPOL). Taxon 27: 481-484.

Zalani L. A. M., Jye K. S., Balakrishnan S. & Dhillon S. K. 2019. Biodiversity Databases and Tools. Encyclopedia of Bioinformatics and Computational Biology 2: 1110-1123.