provide universal presentation template, keeping in mind decrease of information volume and value, in comparison to original source digital services. Further phase, having place currently at UWr, is tuning the system according to users experiences and expectations. We are redesigning indexing layer now, providing new mapping of different metadata fields in accordance to analysed users journeys.

ACKNOWLEDGEMENT
Leopoldina online was implemented as part of the Digital Poland Operational Program for 2014 – 2020, Priority Axis 2 "E-administration and open government", Actions No. 2.3 "Digital availability and usefulness of public sector information", Sub-measure 2.3.1 "Digital access to sector information public from administrative sources and science resources "(type II of the project: digital access to science resources).

Implementation period: 1.11.2017 r. – 31. 10. 2020 r.
Project value: 14 040 2Project co-financing from the EU: 11 882 300, 67 PLN

References

1.Europeana: https://www.europeana.eu
2. Federacja Bibliotek Cyfrowych: https://fbc.pionier.net.pl/?lang=en
3. https://leopoldina.pl/
4. Uniwersytet Wrocławski: https://uni.wroc.pl/en/
5. Poznańskie Centrum Superkomputerowo-Sieciowe (PSNC): https://www.psnc.pl/

BUILDING TAXONOMIES ON THE BASE OF VOCABULARY IN REAL USE IN THE SYSTEM, LEOPOLDINA.PL USE CASE

Agnieszka Łętowska; agnieszkaletowska7@gmail.com; (Freelance programmer and linguist, collaborator of Leopoldina online project)

Taxonomies are attempts of ordering world in semantic field. They are also a useful tool of knowledge organization, facilitating domain information retrieval. Taxonomy or thesauri creation requires extensive and thorough analytical process, engaging information and knowledge from variety of sources. Typically it is a long-term activity. Such traditional approach failed for knowledge organization gathered on Leopoldina.pl platform. The aim of the service is to present heterogenic and differentiated digital resources of University of Wrocław (UWr). Due to lack of success of creation an universal taxonomy in given period of time, an bottom-up approach was proposed. The entire range of vocabulary indexed in Leopoldina.pl platform was divided into discipline categories which were used as the base for taxonomies and thesauri creation. In current paper we describe the bottom-up process of taxonomy creation. Programming tools (in Python) used for domain dictionaries creation are presented. We provided also evaluation of external organized knowledge sources (like Wikipedia, GBIF – the Global Biodiversity Information Facility and other domain-specific thematic portals) for automatic knowledge handling and taxonomies creation.

https://doi.org/10.52036/1335793X.2021.1-2.37-41

INTRODUCTION

Taxonomy is a hierarchical way of describing the world. For centuries, the term taxonomy was used only to describe the systematics of living organisms. Nowadays, this term is also used to classify knowledge from various fields in Knowledge Organization Systems (KOS). In information websites, taxonomy is the division of website content into general

categories, from which detailed categories are distinguished. However, if there are mutual relations between particular concepts in our system – apart from hierarchical ones, also equivalence (synonymy) and/or associative relations (association, semantic kinship), we are already talking about a thesaurus. The thesaurus is a conceptual dictionary in which the meaning of terms is described by means of a network of semantic connections with other terms[1].

Leopoldina online is designed for aggregation metadata and presentation information on wide scope of Cultural Heritage (CH henceforth) and other digitised resources from various sources. Thus there are few separate subject fields that need controlled vocabulary support and representation. For example documents, books, rocks and minerals, fossils, living nature or nature exhibits. There is also area of administrative units of UWr.
Most of the objects collected at the University of Wrocław are related to each other by means of descriptions in many ways, creating a large semantic network. For example, a specific plant species will appear in the Natural History Museum as a dried specimen, in the Botanical Garden as a living plant, it may also appear in many books collected in the libraries of several departments.

Leopoldina.pl indicates all these information and their sources, as well as presents the systematics of a given genre. Generally speaking, one may say, this project is a huge thesaurus containing comprehensive information on digitized resources of physical exhibits accumulated in the University of Wrocław during the hundreds of years of its existence.

PREPARATORY WORKS DESCRIPTION

Building thesaurus based on data collected in one, even big, organization, shouldn't be a difficult task. We have countable amount of objects to include, in theory we know, what do we exactly have. But it is only a theory. The objects have been collected for many years and that implicates a lot of problems in knowledge organization.

Heritage of University of Wrocław isn't only a collection of exhibits accumulated in last 75 years. At first the academy was founded by an Austrian emperor Leopold I in the beginning of XVIII century. Than it has belonged to Prussia, Germany and, finally, after II World War, to Poland. Each owner has collected exhibits of different kinds. They have been organized in collections of different structures. Each single structure represented other knowledge representation system.

Preparing taxonomy which fully covers vocabulary scope of the subject or discipline is time consuming task. It involves team of professionals, many knowledge sources and a lot of work. Some of deliverables of Leopoldina online project was preparing infrastructure for handling flat and polyhierarchical dictionaries, including appropriate APIs for vocabulary manipulation. As the University of Wrocław was already supporting the maintenance and development of Midas Browser system, the system was chosen to support dictionaries related tasks. The application itself bases on MIDAS system, well recognized and used for cataloguing of physical culture objects[2].

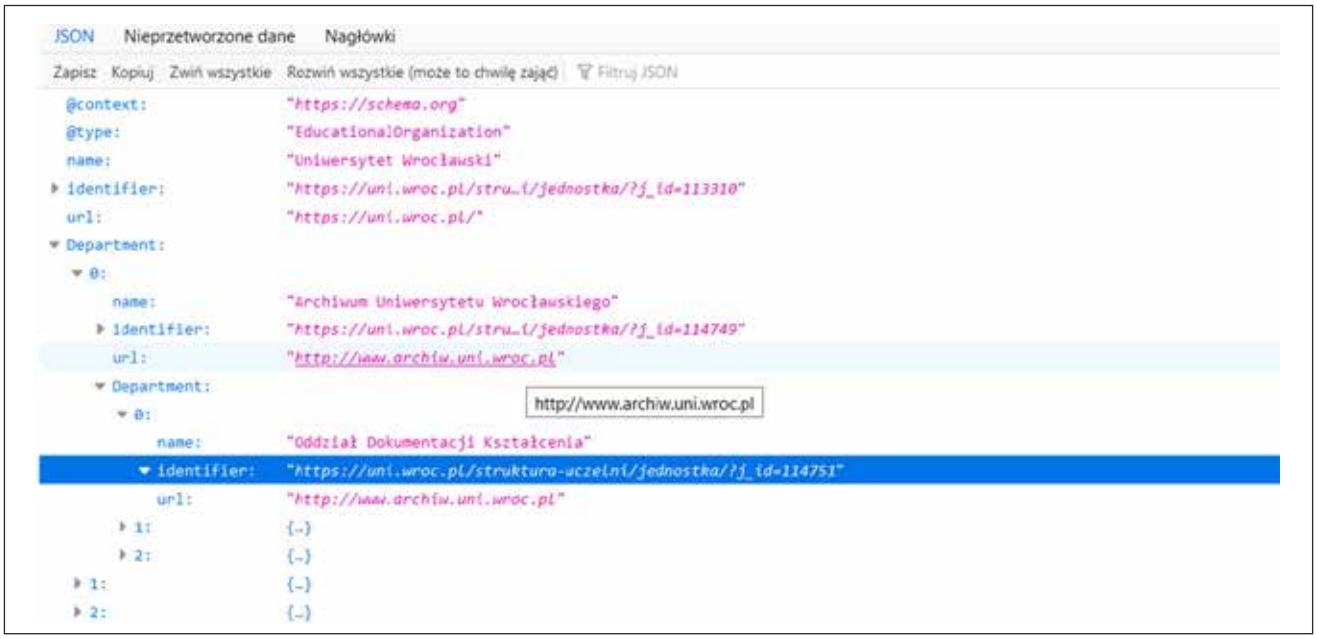
The first two years of work on thesauri was not successful, however. The main obstacle were issues with defining semantic scope of vocabulary in different, distinct subject areas. Scholars working with subject terminology have been using the same terms in different meaning than the theoretical, dictionary one. Thus a lot of draft thesauri or taxonomies, proper in terms of semantic meaning of used vocabulary, was not accepted by users, because of different meaning of the same words in scientific communication. The problem was, among the others, that theoretically proper taxonomies would require to change a lot of scientific articles from many years, due different meanings of terms in the controlled vocabulary and in papers.

Thus an botton-up approach was suggested. We have decided not to produce a full spectrum, wholistic taxonomy or thesauri but to build knowledge representation structures on the strict base of vocabulary already introduced and indexed in Leopoldina online portal.

The first attempt was made for creating a taxonomical representation of UWr administrative organization. All units are listed (unfortunately only in Polish): https://uni.wroc.pl/en/school-organization/university-of-wroclaw-units/?j_id=113310. Each subpage of the listing consists of subunits respectively, in a recursive manner. For each unit its place in the structure was presented in a nested JSON file, and additionally website address was attached, which are not originally registered in the listing. A python script was prepared

for arranging purposes, the final JSON structure was then manually verified. JSON file was chosen for its popularity, computer-readability and capability of storing nested structures. Out of JSON it is possible to build thesauri-like files, e.g. in MIDAS format. Even this, one could say easy, work brought some surprises as University structure is evolving. There were some subunits which do not exist any longer, and some quite new. These changes are not up-to-date reflected in official University structure listing. The problem was also lack of unified policy of domain addresses for University units web pages, clear especially among faculties. Thus it was not possible to collect web addresses automatically. This added informative value was thus stored and registered in hybrid manner. A piece of JSON file is presented on the picture 1.

Next structures to represent were taxonomies of exhibits from Natural History Museum of UWr and from Botanical Garden. Each exhibit has been already described with taxonomical path of living organism. Some paths were full, but there were also a lot of them with partial paths. As mentioned in the introduction part, the taxonomical descriptions come from different



Picture 1. Example of JSON file with UWr administrative structure

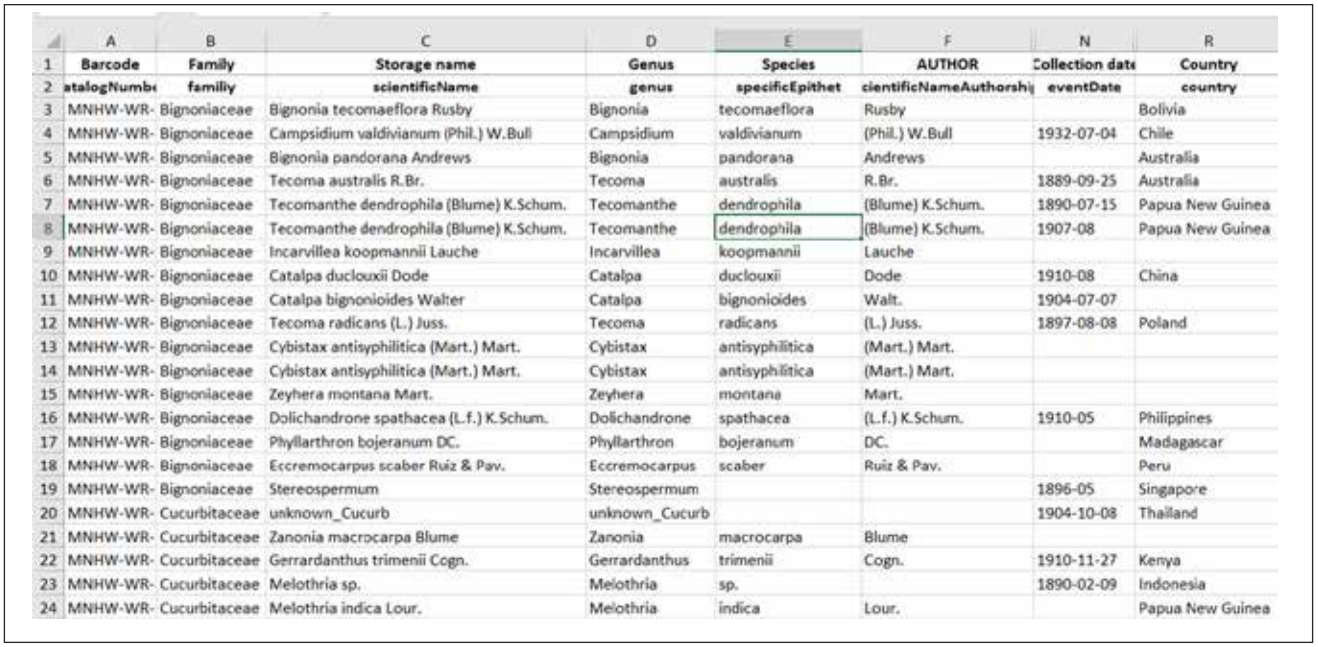
historical times, which means they represent different knowledge organization systems as well as different living organism classification systems or approaches. Some of the organisms have changed their class or family or other systematical level attachment due to current scientific analyses or discoveries on morphological or DNA level. Additionally, what is quite regular situation in biological exhibits descriptions from a wide scope of time, there were a lot of partially resolved taxonomical paths. It is result of many factors, like haste and lack of time resulting in registration only the core data, with hope to fulfill the rest later. A hope, which, we know, is deceptive. Another reason could be insufficient knowledge of describing person, like a young student e.g.

For exhibits, after consultation with natural history collection curators, we have decided to provide the basic taxonomical path, as follow:

Kingdom -> Phylum -> Class -> Order -> Family -> Genus -> Species

Those information were provided, according to scientific rules, in Latin. Additional information were Scientific name, Specific name and common popular name (in Polish). Picture 2. presents an example of original notation of taxonomical path.

Using Python programming language it was quite easy to write a code for matching the scientific name of the exhibit organism with taxonomical path provided in a chosen authority web source. The main problem was finding reliable, full and commonly accepted authority source. Indeed, there are several dozen of such sources. However, each of them states they are not authority, as do not cover taxonomies for each



	A	B	C	D	E	F	N	R
1	Barcode	Family	Storage name	Genus	Species	AUTHOR	Collection date	Country
2	atalogNumb	familiy	scientificName	genus	specificEpithet	cientificNameAuthorshi	eventDate	country
3	MNHW-WR-	Bignoniaceae	Bignonia tecomaeflora Rusby	Bignonia	tecomaeflora	Rusby		Bolivia
4	MNHW-WR-	Bignoniaceae	Campsidium valdivianum (Phil.) W.Bull	Campsidium	valdivianum	(Phil.) W.Bull	1932-07-04	Chile
5	MNHW-WR-	Bignoniaceae	Bignonia pandorana Andrews	Bignonia	pandorana	Andrews		Australia
6	MNHW-WR-	Bignoniaceae	Tecoma australis R.Br.	Tecoma	australis	R.Br.	1889-09-25	Australia
7	MNHW-WR-	Bignoniaceae	Tecomanthe dendrophila (Blume) K.Schum.	Tecomanthe	dendrophila	(Blume) K.Schum.	1890-07-15	Papua New Guinea
8	MNHW-WR-	Bignoniaceae	Tecomanthe dendrophila (Blume) K.Schum.	Tecomanthe	dendrophila	(Blume) K.Schum.	1907-08	Papua New Guinea
9	MNHW-WR-	Bignoniaceae	Incarvillea koopmannii Lauche	Incarvillea	koopmannii	Lauche		
10	MNHW-WR-	Bignoniaceae	Catalpa duclouxii Dode	Catalpa	duclouxii	Dode	1910-08	China
11	MNHW-WR-	Bignoniaceae	Catalpa bignonioides Walter	Catalpa	bignonioides	Walt.	1904-07-07	
12	MNHW-WR-	Bignoniaceae	Tecoma radicans (L.) Juss.	Tecoma	radicans	(L.) Juss.	1897-08-08	Poland
13	MNHW-WR-	Bignoniaceae	Cybistax antisyphilitica (Mart.) Mart.	Cybistax	antisyphilitica	(Mart.) Mart.		
14	MNHW-WR-	Bignoniaceae	Cybistax antisyphilitica (Mart.) Mart.	Cybistax	antisyphilitica	(Mart.) Mart.		
15	MNHW-WR-	Bignoniaceae	Zeyhera montana Mart.	Zeyhera	montana	Mart.		
16	MNHW-WR-	Bignoniaceae	Dolichandrone spathacea (L.f.) K.Schum.	Dolichandrone	spathacea	(L.f.) K.Schum.	1910-05	Philippines
17	MNHW-WR-	Bignoniaceae	Phyllarthron bojeranum DC.	Phyllarthron	bojeranum	DC.		Madagascar
18	MNHW-WR-	Bignoniaceae	Eccremocarpus scaber Ruiz & Pav.	Eccremocarpus	scaber	Ruiz & Pav.		Peru
19	MNHW-WR-	Bignoniaceae	Stereospermum	Stereospermum			1896-05	Singapore
20	MNHW-WR-	Cucurbitaceae	unknown_Cucurb	unknown_Cucurb			1904-10-08	Thailand
21	MNHW-WR-	Cucurbitaceae	Zanonia macrocarpa Blume	Zanonia	macrocarpa	Blume		
22	MNHW-WR-	Cucurbitaceae	Gerrardanthus trimenii Cogn.	Gerrardanthus	trimenii	Cogn.	1910-11-27	Kenya
23	MNHW-WR-	Cucurbitaceae	Melothria sp.	Melothria	sp.		1890-02-09	Indonesia
24	MNHW-WR-	Cucurbitaceae	Melothria indica Lour.	Melothria	indica	Lour.		Papua New Guinea

Picture 2. Example of original exhibit description

living organism or are not up-to-date with current findings. Finally, due its complexity and coverage of living organisms kingdom the GBIF was chosen as referential point.

GBIF – Global Biodiversity Information Facility – an international network and data infrastructure funded by the world's governments and aimed at providing anyone, anywhere, open access to data about all types of life on Earth[3].

It provides taxonomical path, biological characteristics, pictures, geolocation data for huge amount of species all over the world. The added value are synonyms and old names of particular specie. The data are available through the website as well as dedicated API.

A script was prepared for matching scientific name from UWr records with the same level of nomenclature present in GBIF. Successful match was then exploded to full taxonomical path, chosen systematical units were extracted and replaced in local record. The part of code is listed in Picture 3.

Unfortunately, GBIF is excellent, but not perfect source of biological knowledge. Some organisms don't exist there, so we had to search for them in other sources. Some of missing descriptions we have found in Wikipedia, some in WoRMS: World Register of Marine Species, some were available only in PDFs of old, biological books from XIX or XX century etc.

The matching and replacing operations were made recursively for consecutive systematical units in case of partially provided descriptions in local records. During building taxonomy process we did also removed duplicates. Finally a set of 27 515 unique species was selected and exploded into taxonomy.

Analogical process was conducted for Geological Museum exhibits. This collection consist of rocks and minerals as well as fossils. In this case as referential source of knowledge we used Wikipedia – presenting Table of geologic time, according to International Commission on Stratigraphy (ICS). Presented fossils as a species were decoded on the base of GBIF data and their taxon paths were recorded.

CONCLUSIONS

Experience with Leopoldina online resources has proven we need Knowledge Organisation Systems dedicated for different fields of knowledge represented in such project. We have proved that creating thesauri and taxonomies on the base of already indexed vocabulary is reliable and practical approach. The main obstacle is serious lack of universal, referential authority knowledge source. Thus strong need of combining knowledge from different sources appears.

ACKNOWLEDGEMENT

Leopoldina online was implemented as part of the Digital Poland Operational Program for 2014 – 2020, Priority Axis 2 "E-administration and open government", Actions No. 2.3 "Digital availability and usefulness of public sector information",

Picture 3. Python code matching species name and building taxonomical path

Sub-measure 2.3.1 "Digital access to sector information public from administrative sources and science resources "(type II of the project: digital access to science resources).
Implementation period: 1. 11. 2017 r. – 31. 10. 2020 r.
Project value: 14 040 293,84 zł
Project co-financing from the EU: 11 882 300, 67 PLN

References

[1] Barbara Sosińska-Kalata (2005). Systemy organizacji wiedzy w środowisku sieciowym. Available at https://www.researchgate.net/publication/284023385_Systemy_organizacji_wiedzy_w_srodowisku_sieciowym

[2] Midas: Marburger Informations-, Dokumentations- und Administrations-System (MIDAS): Handbuch und CD (Literatur und Archiv; 4). - 4. überarbeitete Auflage, DOI: https://doi.org/10.11588/artdok.00003770:

[3] From What is GBIF?, available: https://www.gbif.org/what-is-gbif

Geologic time scale. In: Wikipedia: Geologic time scale [online]. [Viewed date: 01.06.2021]. Available from: <https://en.wikipedia.org/wiki/Geologic_time_scale>

Midas: Marburger Informations-, Dokumentations- und Administrations-System (MIDAS): Handbuch und CD (Literatur und Archiv ; 4). - 4. überarbeitete Auflage [online]. [Viewed date: 01.06.2021]. Available from: < https://doi.org/10.11588/artdok.00003770>

Sosińska-Kalata Barbara. Systemy organizacji wiedzy w środowisku sieciowym [online]. Warszawa: Wydawnictwo SBP, 2005. [Viewed date: 01.06.2021]. Available from: <https://www.researchgate.net/publication/284023385_Systemy_organizacji_wiedzy_w_srodowisku_sieciowym>

What is GBIF?. In: GBIF: Global Biodiversity Information Facility [online]. Copenhagen: GBIF [Viewed date: 01.06.2021]. Available from: <https://www.gbif.org/what-is-gbif>

WoRMS: World Register of Marine Species [online]. [Viewed date: 01.06.2021]. Available from: https://www.marinespecies.org/