HAL
archives-ouvertes.fr

Towards a Shared Reference Thesaurus for Studies on History of Zoology, Archaeozoology and Conservation Biology

Cécile Callou, Franck Michel, Catherine Faron Zucker, Chloé Martin, Johan Montagnat

▶ **To cite this version:**

Cécile Callou, Franck Michel, Catherine Faron Zucker, Chloé Martin, Johan Montagnat. Towards a Shared Reference Thesaurus for Studies on History of Zoology, Archaeozoology and Conservation Biology. Extended Semantic Web Confenrece 2015, workshop Semantic Web For Scientific Heritage (SW4SH), May 2015, Portoroz, Slovenia. Proceedings of the ESWC'15 workshops. <hal-01146638>

HAL Id: hal-01146638
https://hal.archives-ouvertes.fr/hal-01146638

Submitted on 28 Apr 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Towards a Shared Reference Thesaurus for Studies on History of Zoology, Archaeozoology and Conservation Biology

Cécile Callou[1], Franck Michel[2], Catherine Faron-Zucker[2], Chloé Martin[1], and Johan Montagnat[2]

[1] Archéozoologie et archéobotanique (UMR 7209), BBEES (UMS 3468), Sorbonne Universités, Muséum National d'Histoire Naturelle, CNRS, France
[2] Univ. Nice Sophia Antipolis, CNRS, I3S (UMR 7271), France

Abstract. This paper describes an ongoing work on the construction of a SKOS thesaurus to support multi-disciplinary studies on the transmission of zoological knowledge throughout historical periods, combining the analysis of ancient literature, iconographic and archaeozoological resources. We first describe the I2AF, a national archaeozoological and archaeobotanical inventory database integrating data from archaeological excavation reports. Then we describe the TAXREF taxonomical reference designed to support studies in Conservation Biology, that was enriched with bioarchaeological taxa from I2AF. Finally we describe the TAXREF-based SKOS thesaurus under construction and its intended use within the Zoomathia research network.

Keywords: I2AF, TAXREF, SKOS, History of Zoology

1 Introduction

Animal bones and plant remains from archaeological excavations are a rich and original source of information on the history of biodiversity and its interaction with human societies. When compared with the knowledge about diversity and current locations of human populations, these remains help to figure out the scenarios of past extinction, biological invasions and anthropic impact. This is particularly true during the Holocene, when the influence of human activities overrode that of climatic factors. Therefore, gathering archaeozoological and archaeobotanical data in a sustainable bioarchaeological database, publicly available, represents a major challenge for Natural Sciences and Conservation Biology. *Archaeozoological and Archaeobotanical Inventories of France* database [1] (I2AF) aims to address this challenge.

Historians address a related challenge. Identifying the reported species in ancient literary and iconographic resources, and assessing the documentation is a momentous issue of the History of Zoology. An increasing amount of primary material (such as textual or iconographic resources) is encoded in domain-specific

digital formats. For instance, the SourcEncyMe[3] and Ichtya[4] projects aim to encode mediaeval encyclopedias in the XML-TEI standard[5] while adding manual annotations with regards to mediaeval compilers, author sources and taxa. These works succeed in making material about mediaeval scientific knowledge more easily exploited by a broad scientific community, and support researchers studying e.g. the transmission of zoological knowledge throughout historical periods. Yet, the sharing with related scientific communities remains hampered by the lack of formal semantic reference and terminological standards. For instance, the dolphin is a research topic for modern studies on biodiversity, for archaeozoologists, as well as for studies on Greek mythology wherein the dolphin played an important symbolic role [2]. Nevertheless, when the dolphin is identified in the TEI annotation of the *Hortus Sanitatis* mediaeval encyclopedia[6] or in Pliny the Elder work (*Historia Naturalis*), how to know whether this refers to the same animal? How to know which species is targeted, since the Latin word *delphinus* is used in the textual tradition at least for all Mediterranean regular species of Delphinidae, and labels many different modern taxa (Tursiops truncatus, Delphinus delphis, Stenella coeruleoalba, etc.)? How to relevantly relate those terms with the Delphininae subfamily of modern zoological taxonomy, or even upper terms in the classication (family Delphinidae, order Cetacea)? More generally, how to simultaneously query various archaeozoological, zoological and historical data sources, crosscheck the evidences and make sure that concepts share the same meaning across data sources?

Those challenging questions can be addressed through the use of controlled and widely accepted semantic references. A reference thesaurus shared by sibling scientific disciplines would help to clear the many misinterpretations or conflations made by ancient authors and debated at length in modern critic literature referring to Ancient sources (from P. Belon, 1551, to I. Geoffroy Saint-Hilaire, 1841). The Zoomathia research network[7] addresses this challenge, specifically on the study of rich mediaeval compilation literature on Ancient zoological knowledge, supported by archaeological and iconographic knowledge. The Semantic Web provides powerful models and technologies for connecting and sharing pieces of data while making their semantics explicit. RDF facilitates the combination and sharing of different data sets thanks to the underlying Web technologies and the subsequent Linked Data paradigm. Zoomathia intends to leverage those technologies to annotate and link together various medieval compilations such as the *Hortus Sanitatis*[8], archaeozoological data (I2AF database) and iconographic material. In this context, we chose the TAXREF [3] zoological and botanical

[3] http://atelier-vincent-de-beauvais.irht.cnrs.fr/encyclopedisme-medieval/programme-sourcencyme-corpus-et-sources-des-encyclopedies-medievales
[4] http://www.unicaen.fr/recherche/mrsh/document_numerique/projets/ichtya
[5] http://www.tei-c.org/index.xml
[6] https://www.unicaen.fr/puc/sources/depiscibus/accueil
[7] http://www.cepam.cnrs.fr/zoomathia/
[8] This very popular text that enjoyed numerous editions and translations between 1491 et 1547 is not only a landmark in the history of encyclopedias, but also, concerning the naturalistic knowledge, representative of the whole medieval tradition. It provides

taxonomy to build a SKOS thesaurus supporting the integration of these heterogeneous data sets.

This paper is organized as follows: Section 2 presents the I2AF project. Section 3 describes the TAXREF taxonomical reference. Then, section 4 presents our ongoing work on the construction of a SKOS thesaurus based on TAXREF. Finally, section 5 concludes and suggests leads for future works.

2 I2AF: Archaeozoological and Archaeobotanical Inventories of France

During the eighties decade, it was acknowledged that the access to archaeological data by researchers was increasingly challenged by the growing amount of data produced, and hampered by its scattering. The risk of permanent loss was even more worrying. Thus, it appeared obvious that data in archaeological reports had to be systematically and sustainably collected and inventoried, in a heritage perspective, while making them available to all potential users. From 2003 on, several programs supported by multiple French institutes designed, deployed and maintained such a national inventory database. Today, the I2AF is a collection of the French *National Museum of Natural History* (MNHN). It is continuously and increasingly populated with data on flora and fauna from reports of all excavations performed in French territories, whether the bioarchaeological material was already studied or not. Since January 2014, the inventory and knowledge dissemination effort has been actively sustained by a national multi-institute network of bioarchaeologists[9]. When data from excavation reports is imported into the I2AF, it is aligned on two thesauri: a chronocultural thesaurus provides temporal terms with regards to cultural periods (the oldest records date back to the Middle Palaeolithic), and a taxonomic thesaurus of zoological and botanical names, namely the TAXREF taxonomical reference (see section 3).

As the national reference for nature and biodiversity, the MNHN is responsible for scientific and technical coordination of the natural heritage inventory. To this end, it develops and distributes the TAXREF taxonomical reference, and maintains the *National Inventory of Natural Heritage*[10] (INPN), an information system that gathers current (contemporary) occurrence data on fauna and flora of metropolitan France and overseas departments and collectivities. To date, INPN gathers data from approximately 800 data sources aligned on TAXREF. In this context, the I2AF was naturally identified as a potential data contributor to the INPN. This was however challenging due to the discrepancies between both databases in terms of temporal periods and inventoried species. Indeed, while the INPN gathers actual environmental data on wild life, the I2AF also

most of the data available between 1260 and 1320 in western Europe, derived from the late antiquity compilations.

[9] GDR 3644 BioArcheoDat, "Societies, biodiversity and environment: archaeozoological and archaeobotanical data and results on the French territory".

[10] Inventaire National du Patrimoine Naturel: http://inpn.mnhn.fr. Muséum National d'Histoire Naturelle [Ed]. 2003-2015.

provides archaeological data on domestic species, exotic species (not inventoried on any French territory, notably imported by menageries as soon as Roman Antiquity) and possibly extinct species. This issue was solved progressively by enriching TAXREF with new taxa along with the integration of I2AF data into the INPN. As examples we can cite extinct species such as the mammoth and the cave bear, domestic species such as the dog and the ox, and exotic species such as the Barbary macaque.

3 TAXREF: a Taxonomic Reference in Conservation Biology

TAXREF[3] is the French national taxonomic reference for fauna, flora and fungus of metropolitan France and overseas departments and collectivities. It is developed and distributed by the MNHN in the context of the Information System on Nature and Landscapes[11]. TAXREF aims to (i) give an unambiguous unique scientific name for each taxon inventoried on the territory, that marks a national and international consensus; (ii) enable interoperability between databases in (archaeo)zoology and (archaeo)botany, to help the study of biodiversity and support strategies of natural heritage conservation; and (iii) manage the taxonomic changes (synonymy, taxonomic hierarchy).

TAXREF can be browsed on the INPN web site, and downloaded in TSV format (tab-separated values). An on-going work aims to set up a Web service allowing to query the taxonomy and retrieve results in XML or JSON formats. TAXREF is organized as a unique, controlled, hierarchical list of scientific names. Conceptually, it consists of a table wherein one row uniquely describes one scientific name. All taxonomical names are presented in the same way, whatever their taxonomical rank. Most salient fields are listed below:

- *CD_NOM*: unique identifier of the scientific name.
- *CD_SUP*: identifier of the upper taxon in the classification.
- *CD_REF*: identifier of the reference taxon. This may be either the same as *CD_NOM* or a different one. In the latter, *CD_NOM* identifies a synonym of the reference name identified by *CD_REF*.
- *Nom*: taxon scientific name.
- *Nom_Vern* and *Nom_Vern_Eng*: French and English vernacular names.
- *Auteur*: taxon authority (author name and publication year).
- *Rang*: taxonomical rank (phylum, class, order, family, gender, species...), represented by a code of two to four letters.
- *HABITAT*: type of habitat in which the taxon usually lives marine, fresh water, terrestrial...) coded as values from 1 to 8.
- A set of biogeographical statuses, one for each geographical region (metropolitan France and overseas departments and collectivities). E.g.: P stands for present, E for endemic, X for extinct, etc.

[11] http://www.naturefrance.fr/sinp/presentation-du-sinp

As an example, Listing 1.1 shows a JSON excerpt describing the common dolphin using its reference scientific name *Delphinus delphis*, and its synonym *Delphinus tropicalis*. Annotation `"HABITAT":1` states that it lives in a marine habitat. Annotation `"Rang":"ES"` states that the taxon belongs to the *species* taxonomical rank (*ESpèce* in French). Annotation `"GUA":"P"` states that its biogeographical status is P (present) in Guadeloupe, a French overseas department. A comprehensive description of allowed values for the habitat, taxonomical rank and biogeographical status is provided in [3].

```
{  "CD_NOM":60878, "CD_REF":60878, "CD_SUP":191591,
   "Nom":"Delphinus delphis",
   "Nom_Vern":"Dauphin commun a bec court",
   "Nom_Vern_Eng":"Short-beaked common dolphin",
   "Auteur":"Linnaeus, 1758",
   "HABITAT":1, "Rang":"ES",
   "FR":"P", "GUA":"P", "REU":"B", (...)
},
{
   "CD_NOM":60881, "CD_REF":60878, "CD_SUP":191591
   "Nom":"Delphinus tropicalis",
   "Nom_Vern":"Dauphin commun d'Arabie",
   "Nom_Vern_Eng":"Arabian common dolphin",
   "Auteur":"Van Bree, 1971",
   "HABITAT":1, "Rang":"ES",
   "FR":"P", "GUA":"P", "REU":"B", (...)
}
```

Listing 1.1. Example of a JSON representation of TAXREF entries

Currently, more than 450.000 taxa are registered, covering the continental and marine environments. From the temporal perspective, all current living beings are considered as well as those of the close natural history, that is, from the Palaeolithic until now. Usage statistics[12] attest the large variety of people using TAXREF, far beyond the research community: botanic conservatories, associations, public institutions and collectivities, private companies, individuals. Given its wide adoption in various communities, we chose it to build a SKOS reference thesaurus that should be published and linked on the Linked Data.

4 A TAXREF-based Thesaurus for the Linked Data

In this section we present our ongoing work on the creation of a SKOS vocabulary faithfully representing the TAXREF taxonomical reference. SKOS[13] is the acronym of Simple Knowledge Organization System; it is a W3C standard designed to represent controlled vocabularies, taxonomies and thesauri. It is used extensively to bridge the gap between existing knowledge organisation systems and the Semantic Web and Linked Data.

[12] TAXREF usage statistics are not published publicly but can be provided on demand.
[13] http://www.w3.org/2009/08/skos-reference/skos.html

```
@prefix skc: <http://www.w3.org/2004/02/skos/core#>.
@prefix skx: <http://www.w3.org/2008/05/skos-xl#>.
@prefix tc: <http://lod.taxonconcept.org/ontology/txn.owl#>.
@prefix gn: <http://www.geonames.org/ontology#> .
@prefix nt: <http://purl.obolibrary.org/obo/ncbitaxon#> .
@prefix taxr: <http://inpn.mnhn.fr/taxref/>.

<http://inpn.mnhn.fr/taxref/taxon/60878> a skc:Concept;
  skx:altLabel <http://inpn.mnhn.fr/espece/cd_nom/60881>;
  skx:prefLabel <http://inpn.mnhn.fr/espece/cd_nom/60878>.
  skc:broader <http://inpn.mnhn.fr/taxref/taxon/191591>;
  taxr:hasHabitat <http://inpn.mnhn.fr/taxref/habitat#Marine>;
  taxr:bioGeoStatusIn [
   taxr:bioGeoStatus  <http://inpn.mnhn.fr/taxref/bioGeoStat#P>;
   gn:locatedIn <http://sws.geonames.org/3017382/> ];
  taxr:bioGeoStatusIn [
   taxr:bioGeoStatus  <http://inpn.mnhn.fr/taxref/bioGeoStat#P>;
   gn:locatedIn <http://sws.geonames.org/3579143/> ];
  taxr:bioGeoStatusIn [
   taxr:bioGeoStatus  <http://inpn.mnhn.fr/taxref/bioGeoStat#B>;
   gn:locatedIn <http://sws.geonames.org/935317/> ].

<http://inpn.mnhn.fr/espece/cd_nom/60878> a skx:Label;
  taxr:isPrefLabelOf <http://inpn.mnhn.fr/taxref/taxon/60878>:
  skx:literalForm "Delphinus delphis";
  tc:authority "Linnaeus, 1758";
  nt:has_rank <http://inpn.mnhn.fr/taxref/taxrank#Species>;
  taxr:vernacularName "Dauphin commun a bec court"@fr;
  taxr:vernacularName "Short-beaked common dolphin"@en.

<http://inpn.mnhn.fr/espece/cd_nom/60881> a skx:Label;
  taxr:isAltLabelOf <http://inpn.mnhn.fr/taxref/taxon/60878>;
  skx:literalForm "Delphinus tropicalis".
  tc:authority "Van Bree, 1971";
  nt:has_rank <http://inpn.mnhn.fr/taxref/taxrank#Species>;
  taxr:vernacularName "Dauphin commun d'Arabie"@fr;
  taxr:vernacularName "Arabian common dolphin"@en.

<http://inpn.mnhn.fr/taxref/taxrank#Species> a skc:Concept;
 skc:prefLabel "Species"@en;
 skc:exactMatch
   <http://purl.obolibrary.org/obo/NCBITaxon_species>;
 skc:exactMatch
   <http://rdf.geospecies.org/ont/geospecies#TaxonRank_species>.

<http://inpn.mnhn.fr/taxref/habitat#Marine> a skc:Concept;
 skc:prefLabel "Marine habitat"@en;
 skc:relatedMatch
   <http://lod.taxonconcept.org/ontology/txn.owl#MarineHabitat>;
 skc:exactMatch
   <http://purl.obolibrary.org/obo/ENVO_00002227>.
```

Listing 1.2. Example SKOS representation of TAXREF entries

Listing 1.2 shows the proposed SKOS representation of the taxon presented in Listing 1.1, using the Turtle RDF syntax. The keystone of our modelling of TAXREF in SKOS is as follows. Each taxon is represented by a SKOS concept (line 8); its URI is in namespace `http://inpn.mnhn.fr/taxref/taxon/`, which local name is CD_NOM, the TAXREF taxon identifier (see section 3). The `skc:broader` property is used to model the relationships between a taxon and the upper taxon in the classification (CD_SUP). The reference scientific name of a taxon and its synonyms are defined as values of properties `skx:prefLabel` and `skx:altLabel` respectively (lines 9 and 10). They are URIs in namespace `http://inpn.mnhn.fr/espece/cd_nom/`. These URIs have been defined by INPN; today they are dereferenced to a Web page providing a HTML description of the taxon. The label literal values themselves are defined with property `skx:literalForm` (lines 25 and 33). The habitat and biogeographical status are represented by a property value which subject is the URI representing the taxon (lines 12 to 21), while the authorities, taxonomical rank, and vernacular names are attached to labels (lines 26 to 29 and 34 to 37).

We identified existing vocabularies that can be reused in our context, keeping in mind that we wish to link the TAXREF thesaurus with existing, well-adopted data sets, in particular within the Linking Open Data cloud. We first focussed on classes and properties that represent taxon characteristics (habitat, taxonomical rank, name authority, etc.). For example, taxonomical ranks are defined in various ontologies such as the NCBI taxonomic classification[14] and the GeoSpecies ontology[15]. Similarly, the type of habitat is commonly defined in several ontologies such as the ENVO[16] environment ontology. To keep full control over the TAXREF vocabulary, we chose to define terms (SKOS concepts) for the taxonomical ranks (lines 39 to 44), types of habitat (lines 46 to 51) and biogeographical statuses in a specific TAXREF namespace (`http://inpn.mnhn.fr/taxref/`), and align them with concepts of existing vocabularies using the `skc:exactMatch` or `skc:closeMatch` properties. In future works, we intend to align the TAXREF taxa themselves with taxa in other well-adopted taxonomies.

To perform the translation of TAXREF into a SKOS vocabulary, we use xR2RML [4], a declarative mapping language designed to address the mapping of a large and extensible scope of databases (RDB, NoSQL, XML native database, object oriented, etc.) into RDF, by flexibly adapting to various data models and query languages. The produced RDF graph can reuse existing domain vocabularies. A prototype implementation of the xR2RML mapping language, Morph-xR2RML, supports the translation of data from relational databases and from the MongoDB[17] NoSQL document store. To deal with TAXREF, we import its JSON version into a MongoDB instance. Then, we write the xR2RML mapping that describes how to map the result of queries to the MongoDB instance into RDF triples. Finally, the Morph-xR2RML tool coordinates the whole process: it

[14] http://www.ontobee.org/browser/index.php?o=NCBITaxon
[15] http://datahub.io/dataset/geospecies
[16] http://www.ontobee.org/browser/index.php?o=ENVO
[17] http://www.mongodb.org/

parses the mapping description, performs the queries against the database and produces the resulting target SKOS vocabulary according to the mapping.

5 Conclusion and Future Works

In this paper, through a few simple example questions, we have highlighted today's needs of some scientific disciplines, as diverse as Conservation Biology, Bioarchaeology, and Ancient literature, to gather and make sense of heterogeneous data and material. Then, we have described I2AF, a national archaeozoological and archaeobotanical inventory database integrating data from archaeological excavation reports. We have presented the TAXREF taxonomical reference designed to support studies in Conservation Biology. To meet the needs of Archaeozoology and Archaeobotany, TAXREF was progressively extended with taxa from I2AF. It is the first taxonomical reference used to integrate data from Bioarchaeology and Conservation Biology[5].

Then we have presented our ongoing work on the construction of a SKOS thesaurus based on TAXREF. In the context of the Zoomathia research network, we aim to use this thesaurus to support multi-disciplinary studies on the history and transmission of zoological knowledge throughout historical periods, combining the analysis of ancient and mediaeval literature, iconographic and archaeozoological resources. This will require the enrichment of the TAXREF-based thesaurus with philological and cultural information and its geographical extension to other Mediterranean areas (Greece, Italy, etc.). Besides, in order for a large community to benefit from this work, and to spur its adoption by linked-data based applications, future works target the automatic creation of links with other well-adopted open data sets and thesaurus, may they be non-specialized like DBpedia, or domain-specific like the NCBI taxonomical reference.

References

1. C. Callou, I. Baly, C. Martin, and E. Landais, "Base de données I2AF: Inventaires archéozoologiques et archéobotaniques de France," *Archéopages*, vol. 26, 2009.
2. E. Voultsiadou and A. Tatolas, "The fauna of Greece and adjacent areas in the Age of Homer: evidence from the first written documents of Greek literature," *Journal of Biogeography*, vol. 32, no. 11, 2005.
3. P. Gargominy, S. Tercerie, C. Régnier, T. Ramage, C. Schoelinck, P. Dupont, E. Vandel, P. Daszkiewicz, and L. Poncet, "TAXREF v8.0, référentiel taxonomique pour la France: Méthodologie, mise en oeuvre et diffusion," in *Rapport SPN 2014 - 42*, 2014.
4. F. Michel, L. Djimenou, C. Faron-Zucker, and J. Montagnat, "Translation of relational and non-relational databases into RDF with xR2RML," in *Proc. of 11th International Conference on Web Information Systems and Technologies (WEBIST)*, 2015.
5. C. Callou, I. Baly, O. Gargominy, and E. Rieb, "National Inventory of Natural Heritage website : recent, historical and archaeological data," *The SAA Archaeological Record*, vol. 11, no. 1, 2011.