The Development of the ROSSIO Thesaurus: Supporting Content Discovery and Management in a Research Infrastructure*

Bruno Almeida[1,2][0000−0002−5777−5574], Nuno Freire[1][0000−0002−3632−8046], and
Daniel Monteiro[1]

[1] ROSSIO Infrastructure, NOVA FCSH, Lisbon, Portugal
{brunoalmeida,nunofreire,danielmonteiro}@fcsh.unl.pt
https://rossio.fcsh.unl.pt/
[2] NOVA CLUNL, NOVA FCSH, Lisbon, Portugal

Abstract. The ROSSIO Infrastructure aims at aggregating, organising and contextualising digital objects whose metadata descriptions are provided by a consortium of Portuguese academic, public and private sector institutions in the domains of social sciences, arts and humanities. ROSSIO is the Portuguese representative of DARIAH, the European research infrastructure for arts and humanities. A platform is being developed to provide search and content curation services for researchers and the general public, namely digital collections, digital exhibitions and a virtual research environment. The ROSSIO Thesaurus is being developed to support content discovery and management in the platform. The thesaurus is being modelled as a SKOS concept scheme, and it leverages on existing unstructured vocabularies, such as subject headings and thesauri, produced by the ROSSIO data providers. This paper describes the modelling process of the ROSSIO Thesaurus, its integration and role in the infrastructure, its publication as Linked Open Data, and the results of this work after 6 months of development.

Keywords: Thesauri · SKOS · Linked Open Data · ROSSIO Infrastructure.

1 Introduction

The ROSSIO Infrastructure aims at aggregating, organising and contextualising digital objects based on their metadata descriptions, which are provided

* Supported by Infraestrutura ROSSIO – Ciências Sociais, Artes e Humanidades (project no. 22139), with funding from POR Lisboa 2020, the European Regional Development Fund (ERDF) and Fundação para a Ciência e Tecnologia (FCT).

by a consortium. The latter is composed of seven Portuguese academic, public and private sector institutions that provide access to digitised collections in the domains of social sciences, arts and humanities (SSAH). The consortium is coordinated by the NOVA School of Social Sciences and Humanities (NOVA FCSH), and further includes the Municipal Archive of Lisbon, the Portuguese Film Archives, the Directorate General for Books, Archives and Libraries, the Directorate General for Cultural Heritage, the Calouste Gulbenkian Foundation and the D. Maria II National Theatre. ROSSIO is the Portuguese representative of DARIAH, the European research infrastructure for arts and humanities.[3]

ROSSIO will offer free and open access to a platform that will provide search and content curation services based on the aggregated datasets. These services, aimed at researchers and the general public, will include the curation of digital exhibitions, digital collection and access to a virtual research environment. The ROSSIO infrastructure will provide functionalities to support researchers in the execution of the research data management plans of their projects and publish any research data in compliance with the FAIR principles.[4] The datasets will be described according to DCAT 2 (Data Catalogue Vocabulary), a W3C recommendation for enhancing the discoverability of datasets and data services in the Web [1], which will allow for interoperability with relevant aggregation services such as RCAAP (Scientific Open Access Repository of Portugal), OpenAire and the European Data Portal.

The ROSSIO Thesaurus is being developed as an effort to support content discovery and management in the platform. The thesaurus is modelled in SKOS (Simple Knowledge Organisation System), the W3C recommendation for sharing and linking to controlled vocabularies in the Semantic Web [14]. This paper describes the modelling process of the ROSSIO Thesaurus, its role in the infrastructure and publication as Linked Open Data (LOD).[5] Section 2 will review related work based on the modelling of thesauri and other controlled vocabularies, as well as on their applications in other research infrastructures and as part of LOD initiatives in cultural heritage. Section 3 will describe the modelling methodology and tools implemented in the infrastructure for managing and publishing vocabularies. Section 4 shows the preliminary results of our research. Section 5 presents our conclusions and outlook for future work.

2 Related work

Thesauri and other knowledge organisation systems (KOS), e.g. thesauri, taxonomies, classification schemes, are extensively used in the GLAM sector (Galleries, Libraries, Archives and Museums) for search and information retrieval applications, including subject indexing, keyword-based search in local databases and federated search across multiple databases [4, 7]. SKOS was born out of the

[3] https://www.dariah.eu/

[4] https://www.go-fair.org/fair-principles/

[5] https://lod-cloud.net/

need for standardising a common RDF vocabulary for modelling KOS, following several European projects in the late 1990's and early 2000's, and became a W3C recommendation in 2009 [2]. SKOS is presently one of the most used vocabularies in LOD initiatives, and is extensively reused by other Semantic Web vocabularies. The 'skosification' of controlled vocabularies has enabled their use for Semantic Web applications, such as query expansion, search term recommendations and semantic search engines [13, 15, 19]. A notable example is The Getty Vocabularies Project, which includes reference vocabularies in cultural heritage, such as the Art and Architecture Thesaurus (AAT).[6]

Research infrastructures in Europe make extensive use of SKOS vocabularies for knowledge organisation and interoperability purposes. DARIAH, the European research infrastructure on which ROSSIO is integrated, makes available a vocabulary repository (Vocabs).[7] Among the hosted vocabularies, we highlight the Backbone Thesaurus (BBT), a top-level thesaurus for alignment of domain or project-specific SKOS vocabularies [18]. The Finnish LOD Infrastructure for Digital Humanities (LODI4DH), a joint initiative of several academic institutions, is anchored on a thesaurus and ontology service (Finto)[8], which is managed by the National Library of Finland [9]. Finto consists of a collection of aligned SKOS concept schemes, including general and domain-specific vocabularies [8]. As part of LODI4DH, Finto's vocabularies are employed for annotation and search in a series of semantic portals.

A notable application of Semantic Web vocabularies is Europeana, which provides access to over 50 million digitised cultural heritage objects provided by European institutions in the GLAM sector.[9] The Europeana Data Model (EDM) includes several SKOS elements for the enrichment of provided cultural heritage objects based on so-called 'contextual entities', i.e. agents, places and concepts [5]. The contextualisation efforts of Europeana have included the design of an 'Entity Collection', a knowledge graph that re-uses third-party data and maps entities to EDM [3]. A further initiative is the evaluation of third-party knowledge bases for enrichment and disambiguation purposes, e.g. VIAF with regard to agents [6]. These examples highlight the potential of SKOS vocabularies for content discovery, management and contextualisation in digital libraries and research infrastructures. This constitutes our motivation for the development of a project-specific vocabulary, the ROSSIO Thesaurus, which will be integrated as a component of ROSSIO's LOD infrastructure.

[6] http://vocab.getty.edu/

[7] https://vocabs.dariah.eu/

[8] https://finto.fi/

[9] https://www.europeana.eu/

3 Development of the ROSSIO Thesaurus

3.1 Methodology and tools used

The development of ROSSIO Thesaurus is following the stages shown in Fig. 1, which are based on the methodology for developing multilingual thesauri outlined in ISO 25964-1 [11].

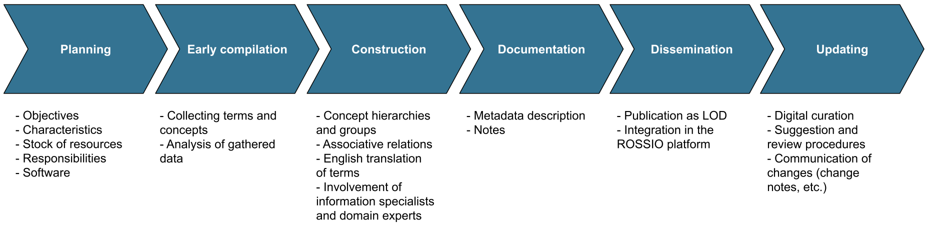
Planning
- Objectives
- Characteristics
- Stock of resources
- Responsibilities
- Software
Early compilation
- Collecting terms and concepts
- Analysis of gathered data
Construction
- Concept hierarchies and groups
- Associative relations
- English translation of terms
- Involvement of information specialists and domain experts
Documentation
- Metadata description
- Notes
Dissemination
- Publication as LOD
- Integration in the ROSSIO platform
Updating
- Digital curation
- Suggestion and review procedures
- Communication of changes (change notes, etc.)

Fig. 1. Development stages of ROSSIO Thesaurus.

Currently, we are entering the construction stage, after having compiled several vocabulary resources from our partner institutions, along with reference thesauri and other relevant KOS from the SSAH. The latter are relevant not only as canonical sources for terms and concepts for the ROSSIO Thesaurus, but also as targets for mapping/alignment. For its publication as LOD, the concepts it comprises should be mapped, whenever possible, to reference and top-level concept schemes, such as the AAT and the BBT.

An important part of our methodology consists of collaborating with information specialists and subject domain experts. This will be fundamental for: (i) supply of local vocabulary resources (e.g. subject-heading lists, thesauri); (ii) validation of third-party vocabulary resources; (iii) feedback regarding the macro- and microstructural components of the ROSSIO Thesaurus.

At this stage, the ROSSIO Infrastructure has already adopted several tools for the development and dissemination of SKOS concept schemes. VocBench 3,[10] an open source platform developed at the University of Rome Tor Vergata [16], has been deployed for managing the ROSSIO Thesaurus development. Among its functionalities, we highlight its flexibility in creating and editing vocabularies, and the possibility of defining several roles for collaborative development.

With regard to dissemination, Skosmos, an open source platform developed at the National Library of Finland [17], is being deployed for the publication of the ROSSIO Thesaurus. This platform offers several possibilities for browsing and information retrieval, including keyword-search with auto-completion and Linked Data access through several RDF serialisations (RDF/XML, Turtle and JSON-LD).

[10] http://vocbench.uniroma2.it/

3.2 Relevant characteristics

ROSSIO Thesaurus' model is based on SKOS and on the ISO 25964 model. Fig. 2 shows the elements that are more relevant for the modelling process. The thesaurus is represented as an instance of skos:ConceptScheme, whose relevant metadata will be described through Dublin Core elements/terms (e.g. creator, description, licence). The main elements of the thesaurus are instances of skos:Concept. The latter are lexicalised by terms (i.e., preferred, alternative and hidden labels) in Portuguese. As a minimum requirement, the preferred labels are translated to English, either through automatic translation and verification or by reusing English terms from reference KOS or other vocabulary resources. There will be the possibility of documenting concepts (e.g. through scope notes) and identifying their sources, which will be mostly relevant for identifying the corresponding terms provided by our partner institutions. Concepts in the ROSSIO Thesaurus are mapped to reference concept schemes in domains within SSAH by means of SKOS mapping properties (e.g. skos:exactMatch, skos:broadMatch). Concepts will be aggregated in collections, including: (i) arrays of sibling concepts (e.g. <Art by period>) and (ii) groups/facets (e.g. Physical Objects).

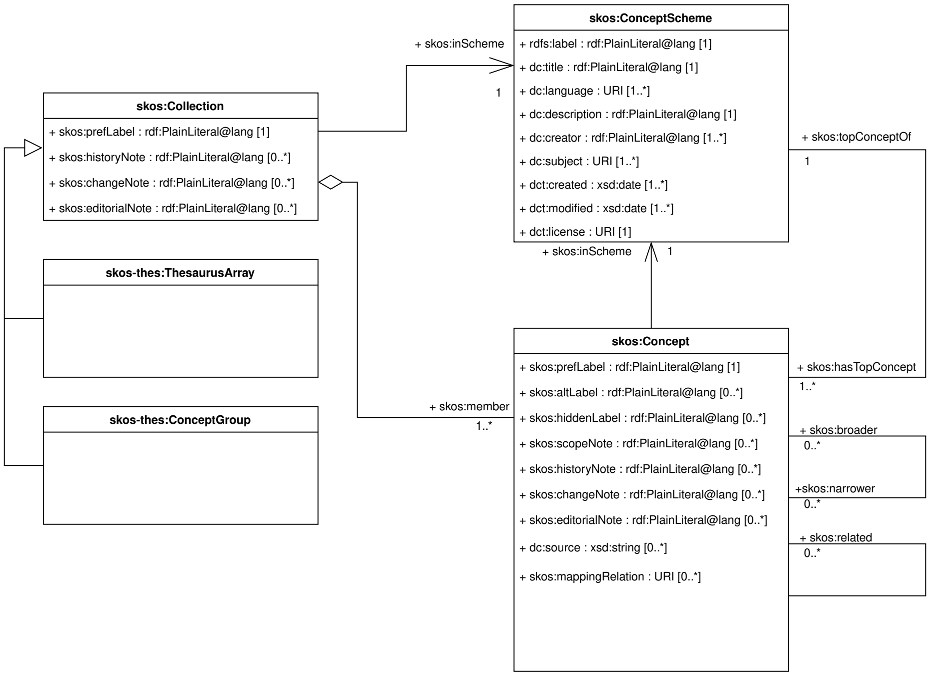
skos:ConceptScheme
+ rdfs:label : rdf:PlainLiteral@lang [1]
+ dc:title : rdf:PlainLiteral@lang [1]
+ dc:language : URI [1..*]
+ dc:description : rdf:PlainLiteral@lang [1]
+ dc:creator : rdf:PlainLiteral@lang [1..*]
+ dc:subject : URI [1..*]
+ dct:created : xsd:date [1..*]
+ dct:modified : xsd:date [1..*]
+ dct:license : URI [1]
+ skos:inScheme
1
skos:Collection
+ skos:prefLabel : rdf:PlainLiteral@lang [1]
+ skos:historyNote : rdf:PlainLiteral@lang [0..*]
+ skos:changeNote : rdf:PlainLiteral@lang [0..*]
+ skos:editorialNote : rdf:PlainLiteral@lang [0..*]
+ skos:topConceptOf
1
skos-thes:ThesaurusArray
+ skos:inScheme
1
skos-thes:ConceptGroup
skos:Concept
+ skos:prefLabel : rdf:PlainLiteral@lang [1]
+ skos:altLabel : rdf:PlainLiteral@lang [0..*]
+ skos:hiddenLabel : rdf:PlainLiteral@lang [0..*]
+ skos:scopeNote : rdf:PlainLiteral@lang [0..*]
+ skos:historyNote : rdf:PlainLiteral@lang [0..*]
+ skos:changeNote : rdf:PlainLiteral@lang [0..*]
+ skos:editorialNote : rdf:PlainLiteral@lang [0..*]
+ dc:source : xsd:string [0..*]
+ skos:mappingRelation : URI [0..*]
+ skos:member
1..*
+ skos:hasTopConcept
1..*
+ skos:broader
0..*
+skos:narrower
0..*
+ skos:related
0..*

Fig. 2. UML class diagram of the ROSSIO Thesaurus.

The ROSSIO Thesaurus includes general concepts, i.e., units of knowledge that correspond to a plurality of objects, which are usually designated by common nouns and noun phrases (e.g. ‘pottery artefacts’). Several individuals are included, however, corresponding to names of disciplines (e.g. ‘sociology’), styles (e.g. ‘Rococo’) and movements (e.g. ‘Dadaism’), among other possibilities.[11]

As part of the ROSSIO Infrastructure, the thesaurus should be complemented by other concept schemes comprising, e.g., place names, person and organisation names and time-periods, as well as domain-specific vocabularies used within the infrastructure, such as the thesaurus on cultural heritage that is being developed by one of our partner institutions, the Directorate General for Cultural Heritage. This approach, if followed, should result in a collection of aligned vocabularies, similar to the example of Finto, and would facilitate the use and reuse of vocabularies for LOD applications.

4 Results

4.1 Development of the thesaurus

The top-level concepts of the ROSSIO Thesaurus were put forward based on the analysis of top-level or reference thesauri and ontologies, such as the BBT, CIDOC CRM, BFO and DOLCE ontologies. The resulting top-level concepts are the following: Agents, Places, Temporal Entities, Physical Objects and Conceptual Objects.

At this time, three institutions of the ROSSIO consortium have contributed to the development of the thesaurus by providing vocabulary resources. Some of the more specific domains represented by these resources include heritage studies and performing arts. The early results described in this section are based on the analysis of one such resource, namely a list of more than 45,000 indexing terms and their frequency of use in the shared catalogue of the Division of Libraries and Documentation of NOVA FCSH. This list covers most fields in the social sciences and humanities.

The list was analysed by performing simple queries, starting with words from key semantic fields, such as ‘arte’ (art) or ‘património’ (heritage). This allowed to retrieve multiword terms with the same base word (e.g. ‘arte portuguesa’ [Portuguese art]), and in which the searched word is part of modifier phrases (e.g. ‘história da arte’ [art history]). Term candidates for associated concepts were also searched (e.g. ‘estética’ [aesthetics], ‘preservação’ [preservation]), and its respective occurrences as heads or modifiers of multiword expressions were retrieved (e.g. ‘preservação digital’ [digital preservation]). Terms and concepts were structured in a spreadsheet and transformed to RDF/SKOS in VocBench. During this process, the preferred terms in Portuguese were automatically translated to English, and the translations were manually verified.

At this time, mappings were carried out to the BBT and the AAT. Since the BBT is a small vocabulary, with only 30 concepts, the mapping was carried out manually. This task was facilitated by the fact that many of the BBT concepts

[11] The notions of general and individual concepts are defined in the ISO standards for terminology work [10].

are reused in the ROSSIO Thesaurus. Mappings to the AAT were automatically discovered through Silk Workbench. The latter task required the extraction of English and Portuguese AAT terms through the SPARQL endpoint of the Getty Vocabularies Project. Terms from the AAT and the ROSSIO Thesaurus were then processed (lowercased and tokenised) and compared based on a similarity measure (Jaccard index) [12]. The resulting RDF alignment was then imported in VocBench, where the mappings were validated.

Here are some relevant metrics regarding the thesaurus project:

- No. of triples: 22,686
- No. of concepts (instances of skos:Concept): 3255
- No. of Portuguese terms (SKOS labels): 4143
- No. of English terms (SKOS labels): 3256
- No. of links to the AAT (SKOS mapping properties): 567
- No. of links to the BBT (SKOS mapping properties): 23

4.2 Application in research data management

The ROSSIO Infrastructure aims to provide functionalities to support researchers in the execution of the research data management plans of their projects. Research data is managed within the infrastructure based on a profile of the Data Catalog Vocabulary (DCAT). The dataset-level metadata is also published by ROSSIO as linked data, as the means to ensure compliance with several of the FAIR principles for research data.

The ROSSIO Thesaurus also has a role in ROSSIO's compliance with the FAIR principles. The dataset metadata is classified with terms from the thesaurus, in order to support the FAIR principles F2-"data are described with rich metadata", and I2-"(meta)data use vocabularies that follow the FAIR principles (as described in Section 3, the ROSSIO Thesaurus is published in compliance with the FAIR principles).

In ROSSIO's dataset metadata profile, concepts from the thesaurus must be used for dcat:theme properties of the dataset. They may also be used for expressing the spatial/geographical coverage of the dataset (property dcterms:spatial) and the temporal period that the dataset covers (with property dcterms:temporal).

5 Conclusion and future work

In this paper, we have presented the work towards the development of the ROSSIO Thesaurus. The latter aims at supporting content discovery and management in the ROSSIO Infrastructure, which will offer access to diversified contents in social sciences, arts and humanities. The thesaurus further aims at being a contribution to the LOD cloud, which we find to be under-represented with regard to linguistic resources in the Portuguese language, as well as to be a reference in domains of the SSAH.

There is still considerable work to be done in the construction and publication of the ROSSIO Thesaurus. We expect to make available the first full version of the thesaurus by September 2021. Other work to be carried out consists on the integration of the thesaurus in the ROSSIO platform, which should allow researchers to annotate virtual exhibitions, virtual collections and other content produced in the platform. We will further explore the use of the thesaurus for search expansion in the platform, and also for semantic search, starting with a smaller dataset provided by a research unit of NOVA FCSH in the field of musicology.

References

1. Albertoni, R., Browning, D., Cox, S., Gonzalez Beltran, A., Perego, A., Winstanley, P.: Data Catalog Vocabulary (DCAT) - Version 2 (Feb 2020), https://www.w3.org/TR/vocab-dcat-2/
2. Baker, T., Bechhofer, S., Isaac, A., Miles, A., Schreiber, G., Summers, E.: Key Choices in the Design of Simple Knowledge Organization System (SKOS). Journal of Web Semantics **20**, 35–49 (May 2013). https://doi.org/https://doi.org/10.1016/j.websem.2013.05.001
3. Charles, V., Manguinhas, H., Isaac, A., Freire, N., Gordea, S.: Designing a Multilingual Knowledge Graph as a Service for Cultural Heritage – Some Challenges and Solutions. In: International Conference on Dublin Core and Metadata Applications Proceedings. pp. 29–40. DCMI, [S.l.] (2018), https://dcpapers.dublincore.org/pubs/article/view/3966
4. Coudyzer, E.: First release GLAM sector reference terminologies (Sep 2013), https://www.athenaplus.eu/getFile.php?id=187
5. Europeana Foundation: Europeana Data Model Primer (Jul 2013), https://pro.europeana.eu/page/edm-documentation
6. Freire, N., Manguinhas, H., Isaac, A.: An Observational Study of Equivalence Links in Cultural Heritage Linked Data for agents. In: Hall, M., Merčun, T., Risse, T., Duchateau, F. (eds.) Digital Libraries for Open Knowledge. pp. 62–70. Springer International Publishing, Cham (2020). https://doi.org/https://doi.org/10.1007/978-3-030-54956-5_5
7. Harpring, P.: Introduction to controlled vocabularies: terminology for art, architecture, and other cultural works. Getty Research Institute, Los Angeles, CA (2010)
8. Hyvönen, E.: Preventing ontology interoperability problems instead of solving them. Semantic Web Journal **1**(1-2), 33–37 (2010). https://doi.org/10.3233/SW-2010-0014
9. Hyvönen, E.: Linked Open Data Infrastructure for Digital Humanities in Finland. In: Reinsone, S., Skadiņa, I., Baklāne, A., Daugavietis, J. (eds.) Proceedings of the Digital Humanities in the Nordic Countries 5th Conference (DHN 2020). pp. 254–259. CEUR-WS, CEUR, Riga (2020)
10. ISO 1087: Terminology work and terminology science – Vocabulary. ISO, Geneva (2019)
11. ISO 25964-1: Information and documentation - Thesauri and interoperability with other vocabularies - Part 1: Thesauri for information retrieval. ISO, Geneva (2011)
12. Jaccard, P.: Nouvelles recherches sur la distribution florale. Bulletin de la Societe Vaudoise des Sciences Naturelles **44**(163), 223–270 (1908)
13. Koutsomitropoulos, D., Solomou, G., Kalou, K.: Federated Semantic Search Using Terminological Thesauri for Learning Object Discovery. Journal of Enterprise Information Management **30**(5), 795–808 (2017). https://doi.org/https://doi.org/10.1108/JEIM-06-2016-0116

14. Miles, A., Bechhofer, S.: SKOS Simple Knowledge Organization System Reference (Aug 2009), http://www.w3.org/TR/skos-reference
15. Nagy, H., Pellegrini, T., Mader, C.: Exploring structural differences in thesauri for SKOS-based applications. In: Proceedings of the 7th International Conference on Semantic Systems. pp. 187–190. ACM, New York (2011). https://doi.org/https://doi.org/10.1145/2063518.2063546
16. Stellato, A., Fiorelli, M., Turbati, A., Lorenzetti, T., Gemert, W., Dechandon, D., Laaboudi-Spoiden, C., Gerencsér, A., Waniart, A., Costetchi, E., Keizer, J.: VocBench 3: a Collaborative Semantic Web Editor for Ontologies, Thesauri and Lexicons. Semantic Web Journal **11**(5), 855–881 (2020). https://doi.org/10.3233/SW-200370
17. Suominen, O., Ylikotila, H., Pessala, S., Lappalainen, M., Frosterus, M., Tuominen, J., Baker, T., Caracciolo, C., Retterath, A.: Publishing SKOS vocabularies with Skosmos (Jun 2015), http://skosmos.org/publishing-skos-vocabularies-with-skosmos.pdf
18. Thesaurus Maintenance WG, D.E.: DARIAH Backbone Thesaurus (BBT): Definition of a model for sustainable interoperable thesauri maintenance (May 2019), https://www.backbonethesaurus.eu/version/bbt-version-122
19. Yang, Y., Xiong, J., Wang, S.: A Semantic Search Engine Based on SKOS Model Ontology in Agriculture. In: Li, D., Liu, Y., Chen, Y. (eds.) Computer and Computing Technologies in Agriculture IV. pp. 110–118. Springer, Berlin (2011). https://doi.org/https://doi.org/10.1007/978-3-642-18333-1_15