Controlled vocabularies and knowledge organisation for the digital humanities

Online workshop
12 July 2021

Organising committee:
Bruno Almeida (ROSSIO Infrastructure / NOVA CLUNL)
Rute Costa (NOVA FCSH. Department of Linguistics / NOVA CLUNL)
Filipa Medeiros (Art Library of the Calouste Gulbenkian Foundation / CIDEHUS – University of Évora / Information Systems in Museums WG of the Portuguese Association of Librarians, Archivists and Information Professionals)

ORGANISED:

Controlled vocabularies and knowledge organisation for the digital humanities : proceedings

Editors:

Bruno Almeida
ROSSIO Infrastructure / NOVA CLUNL
brunoalmeida@fcsh.unl.pt

Rute Costa
NOVA FCSH. Department of Linguistics / NOVA CLUNL
rute.costa@fcsh.unl.pt

Filipa Medeiros
Art Library of the Calouste Gulbenkian Foundation / CIDEHUS – University of Évora / Information Systems in Museums WG of the Portuguese Association of Librarians, Archivists and Information Professionals
f.medeiros@gulbenkian.pt

Funding:

Infraestrutura ROSSIO – Ciências Sociais, Artes e Humanidades, Projeto n.º 22139. Financiada pelo Programa Operacional Regional de Lisboa, apoiado pelo Fundo Europeu de Desenvolvimento Regional (FEDER) e pela Fundação para a Ciência e Tecnologia, I.P/MCTES, através de fundos nacionais (PIDDAC). Código da Operação: LISBOA-01-0145-FEDER-022139 //
NOVA CLUNL - unidade de investigação financiada pela Fundação para a Ciência e a Tecnologia (FCT) através de fundos nacionais (UIDB/LIN/03213/2020 e UIDP/LIN/03213/2020)

Lisbon, December 2021

ISBN: 978-972-9347-41-2

DOI: https://doi.org/10.34619/pgtp-upne

Table of contents

The BBT meta-thesaurus model: building interoperable thesauri for humanities researchers
Helen Goulis 4

De los vocabularios terminológicos a los sistemas de organización de objetos
Jose Antonio Moreiro González 8

The ROSSIO vocabularies: development and publication as linked open data
Bruno Almeida, Nuno Freire, Ângela Salgueiro, Daniel Monteiro 16

How to use controlled lists to organize and manage information: a toolkit for cultural heritage institutions
Filipa Medeiros, Juliana Rodrigues Alves, Natália Jorge, Susana Medina, Marlene Rocha 21

Vocabulaires de recherche, vocabulaire contrôlé et modèle de données : une chaîne opératoire pour le partage des données archéologiques
Sébastien Durost, Guillaume Reich, Jean Pierre Girard 25

Thesaurus de Acervos Científicos em Língua Portuguesa: building a controlled vocabulary in a collaborative mode
Susana Medina.. 30

“Tu que me lês,”
Teresa Barreto Borges 35

The BBT meta-thesaurus model: building interoperable thesauri for humanities researchers

Helen Goulis[1]

[1] Academy of Athens, Athens, Greece

Abstract. The Backbone Thesaurus (BBT for short) is the research outcome of work undertaken by the Thesaurus Maintenance working group of the DARIAH EU in an effort to design and establish a coherent overarching meta-thesaurus for the Humanities, under which specialist thesauri and structured vocabularies used across scholarly communities can be integrated and form a thesaurus federation. Its core feature is that it promotes alignment of cutting-edge terminology to the well-formed terms of the meta-thesaurus capturing general meanings. The BBT favours a loose integration of multiple thesauri, by offering a small set of top-level concepts (facets and hierarchies) for specialist thesauri terms to map to. This way, it enables cross-disciplinary resource discovery, while ensuring compatibility with thesauri that cover highly specific scientific domains and areas of knowledge in development. One of its major advantages is the potential of a sustainable and manageable expansion of the thesauri into new areas of knowledge, in which it continues to be effective and efficient, without forcing the experts to abandon their terminology.

Keywords: humanities, thesaurus building, thesaurus management

The Thesaurus Maintenance Working Group (TMWG), established in the framework of DARIAH EU, has developed a model for existing thesauri to become interoperable and be maintained in a sustainable and scalable way. It has designed a coherent overarching thesaurus for the humanities, a meta-thesaurus or Backbone Thesaurus (BBT), under which all the vocabularies and terminologies in use in the domain can be aligned. It is a faceted classification scheme that enables intersubjective and interdisciplinary classification development and integration. It aims at allowing access, compatibility and comparison across heterogeneous classification systems and enhances collaboration across the various humanities fields without forcing researchers to abandon their own terminology.

1 Methodology

The main issue tackled during the process is the lack of concrete and consistent classification systems which enable the modeling decisions to be made. Each field develops its own system of classification, so that it is impossible to have access to all the information related to a user's query or to map the information from different fields in such a way that the same term will be used to denote the same meaning.

The working group elaborated methods for the development of semantically interoperable thesauri. These methods entail a collaborative approach and involve the use of ontology-driven faceted analysis. First of all, the working group had to define the epistemological principles and the methodological rules to be followed in order to build sustainable, interoperable thesauri, which facilitate the harmonisation of different local terminologies and the sustained collaboration between the thesauri maintainers.

It decided to define an initial set of top-level concepts (facets and hierarchies) based on evidence from existing vocabularies that thesauri maintainers provided us for testing. The latter lead to the building of the methodological principles of the BBT and the development of its tools and services, to date after testing the functionality of the methodological and technical tools and their applicability in real use.

Facets are the most general concepts, whose properties are inherited by all possible hierarchies and narrower terms of each facet. The facets are further subdivided into an open number of hierarchies (expressed by the hierarchy top terms), which inherit the properties of the facet and additionally exhibit at least one specific feature which is characteristic of a certain type of terms within this hierarchy.

A bottom-up approach was adopted– rather than by theoretical argument; top-level concepts (target terms)[1] are developed by adequate abstraction from existing terminological systems (source terms)[2]. This requires an effective methodology in order to decide, if a more generic concept has the power to effectively subsume enough narrower terms from different thesauri and to determine whether it is comprehensible enough in its definition to allow experts from different subdisciplines to align their terms effectively under these concepts. For this methodology, the working group exploits all the advantages offered by categorical semantics, to define the essential properties of the general concepts under which more specific terms are subsumed. The subsumption of narrower under broader term is formulated as an inference supporting the inheritance of the properties of potential instances of the broader term to all the instances of the narrower terms (IsA relationship). The definition of the essential properties of the top level-concepts, which are acceptable regardless of the scientific field in which they apply, enables the classification in a consistent and objective way. The intensional properties of the source terms are utilised to reveal hierarchical relationships that lead to broader categories.

[1] They express types of subjects of attribution i.e., universals whose properties reveal the intensionality of the source terms. They should be context independent.

[2] The scientific terminology that consists of a finite set of general concepts which are used by experts in order to describe their scientific methods, results, tools etc.

These concepts constitute a first operational draft, which will be adapted and extended as the integration of more terminologies will introduce new concepts not yet covered or not well covered, or when finer distinctions into hierarchies or subhierarchies may be needed. The updating and revising of the proposed classification and definitions is an ongoing process.

Ten facets along with their hierarchies, top terms and narrower terms' examples have been defined so far. The labels and the definitions for the facets and hierarchies of the BBT are available in four (4) languages –namely, English, French, German, and Greek. One of the major advantages of this kind of classification is the potential of a sustainable and manageable expansion of the thesauri into new areas of knowledge, in which it continues to be effective and efficient, without forcing the experts to abandon their terminology. Furthermore, it enables collaboration, cross-disciplinary resource discovery, and detection of common principles and ensures compatibility with other thesauri that are restricted to particular areas of knowledge.

2 The BBT federation

The controlled vocabularies/thesauri the concepts of which have been mapped to the BBT to this date are the DAI Thesaurus, the DYAS Humanities Thesaurus and the Parthenos Vocabularies. At the same time, members of the working group are working towards integrating the Language of Bindings Thesaurus, PACTOLS, the Taxonomy of Digital Research Activities in the Humanities (TADiRAH) and the Arts and Architecture Thesaurus (AAT) with the BBT, at least partially.

3 Curation/Publication

The BBT is systematically curated by a multidisciplinary team of experts coming from organisations participating in the TMWG (AA, DAI, FORTH, FRANTIQ/CNRS), through BBTalk, an online service designed to support collaborative, interdisciplinary development and extension of thesauri. The BBT versions are regularly uploaded to the DARIAH-EU vocabularies along with all their connections to local thesauri via the ACDH-ÖAW service. During 2020 the team recorded a user's experience video in order to outline the challenges encountered in the curatorial process. This process served as an overall evaluation of the work performed thus far.

References

1. Goulis, H., Tsouloucha, E.: BBT User stories. Benefits from joining the thesaurus federation. In: Scholarly Primitives - DARIAH Annual Event 2020. Zenodo, Zagreb, Croatia (2020).
2. Nouvel, B., Sinigaglia, E., Humbert. V.: Deconstructing for reconstructing: the use of the BackBone Thesaurus for reorganising the PACTOLS Thesaurus. In: FAIR Heritage: Digital Methods, Scholarly Editing,and Tools for Cultural and Natural Heritage. Consortium

MASA, mémoire des archéologues et des sites archéologiques; Programme Intelligence des Patrimoines (ARD 2020). En ligne, France (2020).

3. Goulis, H., Tsouloucha, E.: A web of thesauri. Aligning specialist thesauri to the Backbone Thesaurus. Thesaurus Maintenance WG meeting. In: DARIAH Annual Event 2019: Humanities Data, Book of Abstracts, pp. 17-18. Warsaw, Poland (2019).
4. Bruseker, G., Goulis, H., Tsouloucha, E.: The BackBone Thesaurus – evolving a meta-thesaurus for the humanities through collaborative, distributed thesauri maintenance and development. In: DARIAH Annual Event 2019: Humanities Data, Book of Abstracts, pp. 77-79. Warsaw, Poland (2019).
5. Georgis, Ch., Bruseker, G. Tsouloucha, E.: BBTalk: An Online Service for Collaborative and Transparent Thesaurus Curation, ERCIM News 116, (2019).
6. Daskalaki, M., Charami, L.: A Back Bone Thesaurus for Digital Humanities, ERCIM News 111, (2017).
7. Daskalaki, M, Doerr, M. Philosophical background assumptions in digitized knowledge representation systems. In Dia-noesis: A Journal of Philosophy, 2017, (3), 17-28.
8. Doerr, M., Daskalaki, M., Bekiari, Ch., Katsiadakis, H., Goulis, H., Terzis, Ch.: Thesaurus Maintenance Methodological Outline. Thesaurus Maintenance Working Group, Greece (2015).
9. Doerr, M., Iorizzo, D.: The dream of a global knowledge network – A new approach. In: Journal on Computing and Cultural Heritage 1(1), 2008.
10. Smith, B.: Ontology. In: Floridi, L. (ed.) The Blackwell Guide to the Philosophy of Computing and Information. Blackwell, Oxford (2004), p. 158.

De los vocabularios terminológicos a los sistemas de organización de objetos

José Antonio Moreiro González[1][0000-0002-8827-158X]

[1] Universidad Carlos III de Madrid

Resumen. Se atiende a la adaptación digital y la gestión en la Web de los SOC. En especial a la intersección de las taxonomías y de los tesauros cuando se les dota de características ontológicas, así como con otros tipos de vocabularios a través de la interoperabilidad, los estándares y los programas de gestión informática, el formato SKOS y las características semánticas. La comparación se detiene para considerar los atributos de la utilización de las taxonomías facetadas en empresas y organizaciones, así como la evolución de las folksonomías hacia estructuras taxonómicas. El carácter ontológico de los tesauros conceptuales se aprovecha para representar sus relaciones a través del modelo de grafos, por el que los conceptos se enlazan mediante verbos en una arquitectura que se mantiene en las sentencias RDF, el modelo sintáctico para procesar metadatos y conseguir la interoperabilidad entre SOC. Es de gran importancia la asociación de objetos de contenido mediante el empleo de expresiones verbales para exponer las gráficas cognitivas que desembocan en los Topic Maps-ontologías. El funcionamiento enlazado de taxonomías y tesauros proporciona mayor riqueza semántica, tanto desde el punto de vista de las relaciones entre objetos de contenido.

Palabras-clave: SOC, Transformación digital, Taxonomías, Tesauros, Ontologías, Redes semánticas, Convergencia funcional.

1 Situación

La generalización del carácter abierto y global de Internet, con su capacidad para conectar a un número ilimitado de usuarios, transformó en los años finales del siglo XX los modelos económicos y sociales. El desarrollo comercial de Internet desde 1995, en especial a través de los navegadores, junto con la expansión del mercado de las telecomunicaciones estableció la situación definitiva para que los servicios de Internet se desarrollaran a un ritmo vertiginoso [1]. En este nuevo marco los SOC transformaron en profundidad sus elementos compositivos, estructuras y aplicaciones, perdieron su delimitación anterior a los campos científico-técnicos y pasaron a dar servicio a otras necesidades de la sociedad, por lo que se adaptaron a todo tipo de usuarios y clientes en muy diversas funciones comunicativas.

Desde la conformación de la Ciencia de la Información, los tesauros fueron los vocabularios terminológicos que intermediaron con los ordenadores en la búsqueda y recuperación de la información. Sus bases de funcionamiento requerían siempre el arbitraje de los expertos en un campo semántico para entender los descriptores, sustantivos de significado unívoco, con los que realizar la búsqueda de manera competente en las bases de datos. Si bien, cuando la información pasó a gestionarse en la red, los problemas causados por la estructura estática, su restricción a la expresión léxica y una representación limitada al campo semántico correspondiente [2], e incluso a una institución concreta, obligó a que los vocabularios terminológicos se transformasen para aprovechar todas sus posibilidades de indización, navegación y recuperación en la Web.

El eje organizativo de las taxonomías y los tesauros se reconfirmó en la ANSI/NISO Z39.19 [3], la BSI Group [4] y la ISO 25964-1 [5] sobre los tres tipos de relación jerárquica en los vocabularios: la genérico-específica, la partitiva y la enumerativa. Si bien, la auténtica transformación se reconoció al fijarse los procesos de interoperabilidad con otros vocabularios mediante la norma ISO 25964-2 [6]. Aunque para favorecer la interoperabilidad, el intercambio y la puesta en común de los vocabularios se requería el concurso simultáneo de las directrices y recomendaciones del W3C (*World Wide Web Consortium*), en especial de SKOS, por parte de la mayoría de los programas informáticos que gestionan tesauros y taxonomías [7]. Sus efectos se notaron en práctica de nuevos tesauros y taxonomías se traduce en la utilidad que supone reutilizar partes suyas, compartirlos y enlazarlos, para lo que se necesita un formato estándar de SKOS, como la expresión en tripletas RDF [8]. También se debe a SKOS el cambio de mentalidad respecto a la idea clave de los vocabularios semánticos que pasó de los términos a los conceptos. Resulta curioso que, en este punto, los conceptos puedan determinarse ontológicamente, pero también a través de la reunión de todos sus significantes, como sucede en los anillos de sinónimos con la representación de las ideas mediante todas sus etiquetas sinónimas. Este hecho repercute asimismo en la aplicación extensiva de los conceptos a organizar, ahora como objetos de contenido, a todo tipo de empresas y organizaciones.

2 Las taxonomías, modelos de uso y de convergencia con otros SOC

La organización jerárquico-categorizadora se representa por lo común ahora mediante un nombre muy actual y, a la vez, tan positivista como los taxones. Se debe a la eficacia de las taxonomías para clasificar por derivación semántica objetos de contenido, por lo que está presente en diferentes SOC, ya que la mayoría de los vocabularios controlados parten de una categorización temática que es, después de todo, una taxonomía de términos, conceptos u objetos de contenido. Aunque el rejuvenecimiento del término taxonomía implica la participación de Internet y de la organización informática de objetos, no aporta a los vocabularios mayor extensión que la determinada clásicamente por las estructuras jerárquicas o clasificatorias, por más que su uso frecuente y su polivalencia las hayan llevado a representar la organización jerárquica. Pues, su indudable éxito se debe a que se incluyen en los esquemas de metadatos, los vocabularios controlados, los modelos de conceptos y los *Topic maps* - ontologías [3-5]. Sin olvidar que, en la práctica, por encima de jerarquizar, ordenan visualmente los objetos y facilitan su recuperación.

Las taxonomías describen el modelo conceptual de un campo semántico, pero no se quedan en la representación terminológica, sino que conectan con la experiencia y el conocimiento del personal propio de una empresa o institución. De modo que organizan los objetos de contenido teniendo en cuenta la naturaleza, ramo, sección, misión, tipo de objetos y disposición de estas, para adaptarse a los servicios que prestan, a los productos que elaboran y a los recursos humanos con los que cuentan. Por lo que se aplican más allá de los conceptos hasta ahora usuales del conocimiento, información y datos.

A la hora de diseñar la estructura de una taxonomía la expansión de la jerarquía por los niveles intermedios ofrece cada vez más posibilidades por la tendencia a acoplar distintos métodos para pasar de un nivel a otro [9]. En este punto debe destacarse la elaboración que incluye facetas, ya más frecuente que la de taxonomías sólo jerárquicas. Las taxonomías facetadas no se limitan a ordenar y encontrar los objetos siguiendo los niveles de jerarquía o rastreando sus características y relaciones. Tampoco su confección se destina a una utilización científica. Es parte de la gestión de los probables objetos de contenido y la recuperación de su información de empresas y otras entidades en el comercio o en los servicios electrónicos para conseguir que la experiencia de usuario sea útil de verdad. Además, en los diferentes niveles jerárquicos, se pueden integrar facetas comunes a un sector de actividad con otras individualizadas de una empresa concreta para la que se diseñan a medida. Por lo que lo que su reutilización es más complicada. Las mismas las facetas se organizan a su vez por jerarquías. Se adaptan así se adaptan a la categoría del producto y pueden referirse al tamaño, el color, el tipo de usuario, la tecnología y las categorías específicas de las características de un producto.

Pese a lo cual, persisten las dificultades de los SOC tradicionales para hacer navegables las grandes colecciones generadas por los usuarios, en especial cuando crecen muy deprisa y el diseño de una categorización profesional es demasiado costoso.

La información sobre el contenido de los objetos puede tener dos orígenes:

1. Los atributos de las taxonomías con estructura arborescente.
2. Las etiquetas de folksonomías asignadas de forma libre y con bajo coste, pero defectuosas.

Se alcanza así un modelo híbrido que integra el conocimiento de la taxonomía y la folksonomía, para mejorar las recomendaciones del comercio o de los servicios electrónicos [10]. Esta solución busca mitigar los defectos de las folksonomías a través de anotaciones semánticas, por ejemplo, con metadatos estructurados que permitan gestionarlas y navegar. Sin que esto oculte la dificultad de encontrar soluciones semánticas para problemas originados por el uso de vocabularios libres y de elaboración gratuita. De modo que a la folksonomía se le dota con una estructura taxonómica originada en el criterio de los propios etiquetadores. Por lo que se genera un SOC conceptual para dar respuesta a la tensión que ese establece entre la capacidad de mejorar la disponibilidad y su comportamiento, cuando la empresa o los usuarios demandan más recursos, y el valor del criterio humano [11]. Finalmente, empujada por el estudio real de sus datos, una folksonomía puede evolucionar hacia una organización jerárquica por clases desde unos objetos con granularidad bien especificada [12].

3 Aproximación de las taxonomías a los tesauros. Su combinación con ontologías

La convergencia de tipos diferentes de SOC es una propensión. Así sucede en la convergencia natural de taxonomías y tesauros, pues para construir y gestionarlos tesauros siempre hay que basarse en el orden y la categorización que suponen las taxonomías. Los programas informáticos de gestión de vocabulario admiten sin distinción taxonomías y tesauros, por lo que han propiciado la aproximación de ambos SOC, incluso de ellos con las ontologías. Así *MultiTes Pro*; *PoolParty*; *Synaptica* (con componentes para la construcción de taxonomías digitales); *Protégé* (enfocado a la construcción de esquemas para la web semántica) o *Cognatrix*. Por otra parte, los métodos de indización automatizada dedicados a grandes taxonomías han valorado el empleo de relaciones asociativas para etiquetar, al situar en contexto el significado de los objetos de conocimiento desde la cercanía semántica de otros objetos tanto genéricos como relacionados por asociación [13].

De esta forma se ha potenciado la inclinación a juntar las características de ambos SOC: una estructura jerárquica dominante como en las taxonomías, y el beneficio de las relaciones asociativas (no jerárquicas) adicionales como las soportadas en los tesauros. En una situación que se extiende a las ontologías. Las relaciones asociativas se establecen entre objetos de forma pragmática, pues sirven de guía a los usuarios que navegan y a los que realizan indizaciones mediante etiquetado manual para identificar conceptos de interés relacionados. Lo que ha llevado a que se apliquen a las taxonomías. Pese a que puede hacerse de forma más inconsistente que en los tesauros. Esta es otra circunstancia que ayuda a que taxonomías y tesauros confluyan, ya que por definición los tesauros cuentan con relaciones asociativas y las taxonomías no, pero se puede acabar creando una taxonomía/tesauro tan sólo con algunas relaciones asociativas. Por otra parte, la diferenciación entre tesauros y taxonomías basada en las relaciones asociativas se hace más por funciones que por estructuras, pues los tesauros siguen más centrados en los vocabularios mientras que las taxonomías en la visualización organizada de objetos.

Al margen de que las clases forman la taxonomía de una ontología, se ha vuelto frecuente que se elaboren taxonomías con características ontológicas, propiciado por la aplicación de anotaciones semánticas, como puede ser mediante la integración de metadatos estructurados. Hay un interés creciente por dotar a las taxonomías de características ontológicas que definan sus relaciones semánticas y los atributos personalizados. Se trata de definir formalmente los objetos de un campo semántico, sus tipos, propiedades y las principales relaciones entre objetos. A lo que ayudan algunos de los programas antes mencionados. No se trata de alcanzar una ontología complicada, si no de aplicar una ontología elemental como capa semántica a una taxonomía o a partes de ella. Además, a la hora de visualizar y diseñar las taxonomías, es muy clara la participación de las ontologías, ya que constituyen la base de los grafos con los que se representa el conocimiento [14]. El interés por las ontologías va en aumento, pues se emplean cada vez más como base de los grafos de conocimiento-*Topic maps*.

4 Los tesauros como modelos de conceptos

Interoperabilidad es sinónimo de reutilización terminológica y esquemática y determina la capacidad de dos o más sistemas o componentes de intercambiar información y de usar la información que se ha intercambiado [3-5]. Fue el paso definitivo para volver semánticos los vocabularios controlados y, por tanto, darle una salida ontológica a su funcionamiento en red. Desde el momento en que las búsquedas y recuperaciones se hacen en pantalla, también el procesamiento se vuelve digital. Se enmarca aquí la posibilidad de mejorar las taxonomías y los tesauros estáticos de sustantivos, mediante la inclusión de verbos para expresar las asociaciones existentes entre conceptos, sin otro límite que el del uso lingüístico y utilizando todas las posibilidades del lenguaje natural [15]. La manera de relacionar los conceptos mediante verbos se había llamado mapas conceptuales (*concept maps*) en el ámbito educativo. No son ningún tipo de KOS, pero sirven para representar el conocimiento en gráficas cognitivas que conforman redes de conceptos con las que se amplía el número de relaciones de asociación de los tesauros y se evita que planteen cualquier ambigüedad.

Estos mapas comprenden la representación del conocimiento mediante frases simples y estructuradas que se componen entre los conceptos (nodos, puntos o vértices) y la unión que se establece entre los nodos (arcos, extremos o satélites) y que muestran, incluso, la dirección que sigue la relación:

sujeto → verbo → predicado; en cuanto asociación:
concepto → relación → concepto

Actuando así los verbos para definir las infinitas acciones que puede realizar un sujeto y que puede recibir un objeto. Lo que amplía las potenciales relaciones mucho más allá de las determinadas en la norma ISO 25964-1 [5]. Desde este esquema de representación básica del lenguaje se ha reproducido la relación entre los datos en el modelo orientado a grafos, que es la base de las herramientas para visualizar las relaciones que se dan, sobre todo, en el ámbito empresarial e institucional, sobre todo entre los clientes o usuarios y los productos. E incluso alcanza a ser la razón subyacente, que no su modelo de representación, en la relación entre los conceptos del modelo relacional.

El empleo de los verbos para nombrar las relaciones se normalizó cuando los *Topic maps* se convirtieron en una norma [16]. En inicio estaban pensados para fusionar índices con tesauros, aunque tal como sucede con las directrices y recomendaciones del W3C para la Web semántica, el desarrollo de su aplicación pasó a gestionarse desde el consorcio *TopicMaps.org*. Se transformaron así en un modelo para representar visualmente la información mediante el establecimiento de una red de enlaces semánticos que no sólo representaba el mapa conceptual de un objeto de contenido, sino que lo enlazaba con otros conjuntos de objetos, llegando a relacionarlo con las *ocurrences* o apariciones de esa red semántica en recursos informativos. Las posibilidades que ofrecen los *Topic* *maps* de hacer inferencias y de navegabilidad mediante estructuras semánticas, proporcionan capas ontológicas a las estructuras del conocimiento con las que se funden [16].

Topic maps y RDF son dos estándares desarrollados de forma independiente para la representación, el intercambio y el manejo de conocimiento que se basan en dar respuesta a la gestión en la Web desde la facilidad que ella misma propicia y que en ambos casos se fundamenta en un modelo de grafos.

La representación lingüística del conocimiento en frases simples apoyó primero la construcción de mapas de conceptos, luego heredada por los *Topic maps*, y es asimismo el origen de la creación de las sentencias con RDF, como modelo sintáctico para para procesar metadatos y obtener interoperabilidad, mediante el que se representan las propiedades y sus valores, cuyas partes se denominan, en coincidencia con lo antes visto:

- Sujeto: recurso o cosa sobre el que versa la declaración
- Predicado: propiedad o característica del sujeto que se expresa mediante esa declaración (creador, idioma, fecha de creación)
- Objeto: valor de la propiedad a la que se refiere el predicado

De forma que la estructura de las sentencias en RDF se expresa mediante triples de relación sujeto-predicado-objeto, en un grafo de representación donde sujetos y objetos son nodos y su arco de relación son las propiedades o predicados. Los triples RDF se expresan: *sujeto → predicado → objeto*, mediante un grafo unidireccional. Es evidente el reflejo de los mapas de conceptos y del esquema de representación básica del lenguaje.

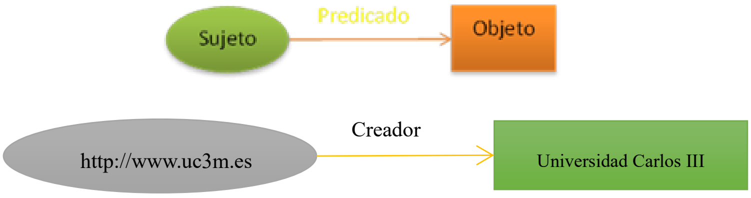

Fig. 1. Triple RDF

5 Rasgos diferenciales LD – VES

Mientras los lenguajes controlados eran terminológicos y sus componentes sustantivos que se caracterizaban por su univocidad de significado y tenían una estructura estática de representación arborescente y ordenación categorial. Su aplicación estaba limitada al necesitarse la intermediación humana en dominios limitados por la especialización. La búsqueda y recuperación eran léxicas. Los vocabularios y esquemas semánticos se basan en conceptos que se pueden asociar mediante verbos y visualizar con adjetivos, lo que les concede mayores posibilidades semánticas. Se aceptan todos los significantes y expresiones de un concepto. Su estructura de representación es dinámica en redes semánticas que determinan la situación contextual del conocimiento. Su aplicación es enlazada y se expanden mediante asociaciones semánticas. La desambiguación de los objetos de contenido se realiza mediante *Schemas* formales.

Referencias

1. Delgado Kloos, C.; García Rubio, C. (2002). Historia de Internet. En Fernández Ordóñez, M.; Cremedes García, J.; Illescas Ortiz, R. (Coord.). Régimen jurídico de Internet (87-100). La Ley.
2. Sánchez-Cuadrado, S., Colmenero-Ruiz, M. J. y Moreiro-González, J. A. (2012). Tesauros: estándares y recomendaciones. El profesional de la información, 2(3), 229-235. https://doi.org/10.3145/epi.2012.may.02
3. Z3919:2005. ISO. ANSI/NISO. Z39.19-2005. Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies. Bethesda, Maryland: NISO Press, 2005. [En línea] http://www.niso.org/standards/index.html [En línea]: http://www.niso.org/standards/resources/Z39-19-2005.pdf.
4. BSI Group. Structured vocabularies for information retrieval: guide. London: BSI, 2005-2007. (BS 8723/1-4).
5. ISO 25964-1. 2011. Information and documentation - Thesauri and interoperability with other vocabularies - Part 1: Thesauri for information retrieval. Geneva: ISO.
6. ISO 25964-2. 2013. Information and documentation - Thesauri and interoperability with other vocabularies - Part 2. Interoperability with other vocabularies. Geneva: ISO.
7. Pastor-Sánchez, J. A., Martínez-Méndez, F. J. y Rodríguez-Muñoz, J. V. (2012). Aplicación de SKOS para la interoperabilidad de vocabularios controlados en el entorno de linked open data. El profesional de la información, 21(3), 245-253. https://doi.org/10.3145/epi.2012.may.04
8. Méndez, Eva; Greenberg, Jane. (2012). Linked data for open vocabularies and HIVE's global framework. El profesional de la información, 21(3), 236-244. https://doi.org/10.3145/epi.2012.may.03
9. Moreiro-González, J. A. (2018). Adaptación de los vocabularios documentales al am-biente digital en red: léxico, significado y relaciones semánticas. Informação & Socie-dade: Estudos, 28(1), 35-46. DOI: 10.22478/ufpb.1809-4783.2018v28n1.37864
10. Bullard, J. (2019). Curated folksonomies: Three implementations of structure through human judgment. KO Knowledge Organization, 45(8), 643-652. doi.org/10.5771/0943-7444-2018-8-643
11. Bolaños-Mejías, C. y Moreiro-González, J. A. (2018). Folksonomy Indexing From the Assignment of Free Tags to Setup Subject: A Search Analysis into the Domain of Legal History. KO Knowledge Organization, 45(7), 574-585. doi.org/10.5771/0943-7444-2018-7-574
12. Mocnik, F. B., Zipf, A., y Raifer, M. (2017). The OpenStreetMap folksonomy and its evolution. Geo-spatial Information Science, 20(3), 219-230. https://doi.org/10.1080/10095020.2017.1368193
13. Pastor-Sánchez, J. A., y Martínez-Méndez, F. J. (2009). Aplicación de tesauros, taxonomías y ontologías en los sistemas de gestión de contenidos mediante tecnologías de la Web Semántica. Ibersid: revista de sistemas de información y documentación, 3, 143-153. https://www.ibersid.eu/ojs/index.php/ibersid/article/view/3734
14. Tramullas, J., Sánchez-Casabón, A. I. y Garrido-Picazo, P. (2009). Gestión de información personal con software para mapas conceptuales. Profesional de la Información, 18(6), 601-612. https://doi.org/10.3145/epi.2009.nov.03

15. Morato, J., Marzal, M. A., Lloréns, J., y Moreiro, J. (20-23 de enero, 2004). Wordnet applications [Comunicación en congreso]. GWC 2004. Second International WordNet Conference. Brno (República Checa). Sojka, P. et al. (eds). GWC 2004: Proceedings of the Second International WordNet Conference. Brno: Masaryk University: 270-278.
16. ISO/IEC 13250. (2003). Information technology: SGML applications topic maps. Geneva: ISO. https://www.iso.org/standard/38068.html.

The ROSSIO vocabularies: development and publication as linked open data

Bruno Almeida[1,2][0000-0002-5777-5574], Nuno Freire[1][0000-0002-3632-8046],
Ângela Salgueiro[1,3][0000-0001-8053-4050], and Daniel Monteiro[1]

[1] ROSSIO Infrastructure, NOVA FCSH, Lisbon, Portugal
[2] NOVA CLUNL – Linguistics Centre of NOVA University of Lisbon, NOVA FCSH, Lisbon, Portugal
[3] NOVA IHC – Institute of Contemporary History of NOVA University of Lisbon, NOVA FCSH, Lisbon, Portugal
rossio@fcsh.unl.pt

Abstract. The ROSSIO Infrastructure brings together a consortium of Portuguese academic, public and private institutions that provide access to unique and diversified digital collections in social sciences, arts and humanities. A platform is being developed for the aggregation and enrichment of the digital objects' metadata provided by consortium members and partner institutions. This platform will provide free and open access to services based on the aggregated datasets, including a discovery portal, digital collections, digital exhibitions and a virtual research environment. The ROSSIO vocabularies, a set of controlled vocabularies modelled in SKOS (Simple Knowledge Organisation System), will facilitate the discoverability of the datasets and their semantic enrichment with Agents, Places, Periods and Topics. These vocabularies will be published as linked open data, being mapped to reference thesauri, gazetteers and other knowledge organisation systems in the linked open data cloud. In this presentation, we provide an overview of the methodology, resources and tools that are currently being used for modelling and publishing the ROSSIO vocabularies as linked open data.

Keywords: controlled vocabularies, research infrastructures, linked open data.

1 Background

The ROSSIO Infrastructure brings together a consortium of Portuguese academic, public and private institutions that provide access to unique and diversified digital collections in the social sciences, arts and humanities (SSAH). The consortium is coordinated by the NOVA School of Social Sciences and Humanities (NOVA FCSH). The other members of the consortium are the Municipal Archives of Lisbon, the Portuguese Film Archives, the Art Library of the Calouste Gulbenkian Foundation, the D. Maria II National Theatre, the Directorate-General for Cultural Heritage of Portugal, and the Directorate-General for Books, Archives and Libraries. ROSSIO also includes partner institutions, such as the Portuguese Web Archive (Arquivo.pt) and the Diplomatic

Institute of the Portuguese Ministry of Foreign Affairs. A platform is being developed for providing services based on the aggregated metadata descriptions of digital objects provided by consortium members and partner institutions. These services include a discovery portal, digital collections, digital exhibitions, and a virtual research environment.

As described in [1], the approach followed in the development and implementation of the platform is based on metadata aggregation through an OAI-PMH instance[1]. During the ingestion process, the metadata are normalised and enriched based on the ROSSIO vocabularies, a set of controlled vocabularies in SSAH, which are being modelled in SKOS (Simple Knowledge Organisation System) and published as linked open data through the infrastructure's vocabulary services[2]. This process is expected to facilitate semantic search in the platform, based on Agents, Places, Periods and Topics, as well as to promote the semantic interoperability of the datasets in the linked open data cloud by mapping the ROSSIO vocabularies to reference thesauri, gazetteers and other knowledge organization systems (KOS) in the linked open data cloud.

The work carried out in the ROSSIO Infrastructure regarding controlled vocabularies is in line with similar initiatives carried out in Europe. During the past decades, controlled vocabularies modelled as multilingual SKOS concept schemes have been applied in several European research infrastructures and linked open data portals for interoperability and semantic search in cultural heritage and digital humanities[3]. These applications of controlled vocabularies are made possible by the transition from term-based, 'legacy' vocabularies to concept schemes published as linked data, which is evidenced by the evolution of thesauri standards in the past decades [2].

2 The ROSSIO vocabularies

2.1 Vocabulary services in the ROSSIO Infrastructure

The ROSSIO platform will include services for constructing and managing controlled vocabularies, as well as for publishing them as linked open data. These services are based on two open-source applications:

- **VocBench 3**. Application for collaborative development and management of vocabularies within the platform. Authorised users will be able to create projects and manage vocabularies.
- **Skosmos**. Application for browsing and publishing vocabularies as linked open data. Users may search and browse the hierarchies and/or alphabetic indexes of concepts

[1] The latest working prototype of the harvesting and ingestion application gathered a total of 37 datasets comprising more than 4 million items.

[2] Available from http://vocabs.rossio.fcsh.unl.pt/en/, last accessed 2021/07/28.

[3] Among the several applications of vocabularies in research infrastructures, we highlight the case of DARIAH-EU, of which ROSSIO is the Portuguese node. DARIAH makes available a vocabulary repository for dataset interoperability in arts and humanities: https://vocabs.dariah.eu/en/, last accessed 2021/06/17.

and terms, while machines may access RDF data through an included SPARQL endpoint.

2.2 Development and publication of controlled vocabularies as linked open data

The core ROSSIO vocabularies comprise the following concept schemes:

- **ROSSIO Agents**. Vocabulary of personal and organisational names.
- **ROSSIO Places**. Vocabulary of toponyms, including names of geopolitical entities, geographical features, areas, buildings and other landmarks.
- **ROSSIO Periods**. Vocabulary of historical, geological, cultural and artistic periods.
- **ROSSIO Thesaurus**. Vocabulary of topics in social sciences, arts and humanities.

As in the case of any vocabulary modelled in SKOS, concepts are the focal elements [3, 4]. In terminology and information science a distinction is generally drawn between general concepts, which are typically designated by common nouns (e.g., ‘artists’) and individual concepts, designated by proper names (e.g., ‘Amália Rodrigues’) [5, 6]. To facilitate the use of the ROSSIO vocabularies for semantic web applications, each concept in the ROSSIO vocabularies is also an instance of one of the following BIBFRAME[4] classes: `Person`, `Organization`, `Place`, `Temporal` and `Topic`.

The ROSSIO Thesaurus plays a central role in the core vocabularies under development [7]. It provides concepts for metadata enrichment, information retrieval and knowledge organisation within the platform. For example, it enables subject indexing of resources produced within the platform, including the core ROSSIO vocabularies, digital collections and exhibitions. The ROSSIO Thesaurus also enables the classification of agent and place types in ROSSIO Agents and ROSSIO Places, respectively, by means of properties reused from the Getty Vocabulary Program ontology[5]. For example, the entry for the Portuguese fado singer Amália Rodrigues in ROSSIO Agents is linked to the ‘fado singers’ concept in the ROSSIO Thesaurus through the agent type property.

The development of the core vocabularies leverages existing structured and unstructured vocabulary resources, such as lists of index terms and in-development thesauri provided by partner institutions, as well as sections of established vocabularies in SSAH, such as the Getty’s Art and Architecture Thesaurus (AAT)[6]. As a minimum requirement, the concepts within the ROSSIO vocabularies are designated by Portuguese labels with English language equivalents. The preferred labels generally follow the standards of thesauri for information retrieval [8], e.g., the plural is used for count nouns (e.g., ‘disciplines), while the singular is used for non-count nouns (e.g.

[4] BIBFRAME ontology, https://id.loc.gov/ontologies/bibframe, last accessed 2021/07/23.

[5] Getty Vocabulary Program ontology, http://vocab.getty.edu/ontology, last accessed 2021/07/28.

[6] Art and Architecture Thesaurus, http://www.getty.edu/research/tools/vocabularies/aat, last accessed 2021/07/28.

‘architecture’). Portuguese standards are also used for determining the headings for personal, organisational and geographic names [9, 10].

As linked data, the vocabularies must include links to external resources identified through URIs. In the case of SKOS vocabularies, this is achieved by declaring mapping properties between concepts in the ROSSIO vocabularies and external KOS, which enables their interoperability [11]. Concepts in ROSSIO Thesaurus and ROSSIO Periods are being mapped to the AAT, either manually or semi-automatically through alignment tools, such as Silk Workbench[7]. The ROSSIO Thesaurus is also aligned with the Backbone Thesaurus (BBT)[8], a top-level thesaurus for interoperability of different vocabularies in arts and humanities, which is managed by a working group within DARIAH-EU. Finally, ROSSIO Agents is aligned with VIAF, the Virtual International Authority File[9], while ROSSIO Places is aligned with place gazetteers, namely GeoNames[10] and the Getty Thesaurus of Geographic Names[11].

2.3 Partner institution and third-party vocabularies

The ROSSIO vocabulary services will include relevant third-party vocabularies for metadata enrichment, whose use will be promoted among the ROSSIO Infrastructure’s partner and member institutions. Presently, we have published a SKOS version of the Lexvo.org knowledge base[12], which includes the ISO 639 two- and three-character codes for languages. At this time, our vocabulary repository also includes a resource types vocabulary from the Confederation of Open Access Repositories[13]. A select number of concepts from this vocabulary should be used for classifying the aggregated metadata descriptions in the platform (e.g., ‘text’, ‘image’, ‘video’, ‘other’).

In the near future, the ROSSIO vocabulary services will also enable the development, publishing and alignment of subject-specific vocabularies by our partner and member institutions. This is the case of the SIPA Thesaurus, a vocabulary focussing on architectural heritage, which is currently under development by the Directorate-General for Cultural Heritage. Enabling the development and publishing of these vocabularies should lead the vocabulary services to become a local hub for developing, publishing and promoting the use of controlled vocabularies in SSAH projects, in line with similar initiatives elsewhere in Europe.

[7] Silk Workbench, http://silkframework.org/, last accessed 2021/07/28.

[8] Backbone Thesaurus, https://vocabs.dariah.eu/backbone_thesaurus/, last accessed 2021/07/28.

[9] VIAF: The Virtual International Authority File, https://viaf.org/, last accessed 2021/07/28.

[10] GeoNames, http://www.geonames.org/, last accessed 2021/07/28.

[11] Thesaurus of Geographic Names, http://www.getty.edu/research/tools/vocabularies/tgn/, last accessed 2021/07/28.

[12] Lexvo.org, http://lexvo.org/, last accessed 2021/07/28.

[13] COAR Resource Types Vocabulary, https://vocabularies.coar-repositories.org/resource_types/, last accessed 2021/07/28.

Acknowledgement. This research was funded by Programa Operacional Regional de Lisboa 2020, FEDER and Fundação para a Ciência e a Tecnologia I.P., grant number 22139.

References

1. Silva, G., Glória, A., Almeida, B., Monteiro, D., Freitas, M., Freire, N.: ROSSIO Infrastructure: a digital research tool for social sciences, arts and humanities. In: Proceedings of the ICTeSSH 2021 conference (2021). https://doi.org/10.21428/7a45813f.579bb144.

2. Dextre Clarke, S., Zeng, M.: From ISO 2788 to ISO 25964: the evolution of thesaurus standards towards interoperability and data modeling. Information standards quarterly. 24, 20–26 (2012). https://doi.org/10.3789/isqv24n1.2012.04.

3. Miles, A., Bechhofer, S.: SKOS Simple Knowledge Organization System Reference, http://www.w3.org/TR/skos-reference, last accessed 2020/07/28.

4. 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 (2013). https://doi.org/10.1016/j.websem.2013.05.001.

5. ISO 1087: Terminology work and terminology science – Vocabulary. ISO, Geneva (2019).

6. ISO 5127: Information and documentation - Foundation and vocabulary. ISO, Geneva (2017).

7. Almeida, B., Freire, N., Monteiro, D.: The development of the ROSSIO Thesaurus: supporting content discovery and management in a research infrastructure. In: Dosso, D., Ferilli, S., Manghi, P., Poggi, A., Serra, G., and Silvello, G. (eds.) Proceedings of the 17th Italian Research Conference on Digital Libraries. pp. 138–146. CEUR-WS, Aachen (2021).

8. ISO 25964-1: Information and documentation - Thesauri and interoperability with other vocabularies - Part 1: Thesauri for information retrieval. ISO, Geneva (2011).

9. Sottomayor, J.C. ed: Regras de catalogação : descrição e acesso de recursos bibliográficos nas bibliotecas de língua portuguesa. APBAD, Lisboa (2008).

10.Área de Classificação e Indexação da Biblioteca Nacional: SIPORbase : Sistema de Indexação em Português : manual. Biblioteca Nacional, Lisboa (1998).

11.ISO 25964-2: Information and documentation - Thesauri and interoperability with other vocabularies - Part 2: Interoperability with other vocabularies. ISO, Geneva (2013).

How to use controlled lists to organize and manage information: a toolkit for cultural heritage institutions

Filipa Medeiros[1], Juliana Rodrigues Alves[2], Natália Jorge[2],
Susana Medina[3] and Marlene Rocha[4]

[1] Art Library of the Calouste Gulbenkian Foundation
[2] CITCEM – University of Porto
[3] Faculty of Engineering of the University of Porto
[4] Câmara Municipal of Porto, Museums Division

Abstract. The aim of this article is to present the work of the Information Systems in Museums Working Group of the Portuguese Association of Librarians, Archivists and Information Professionals, namely a toolkit proposal to build lists of controlled terms to organize and manage information in cultural heritage institutions. This toolkit comprises six steps: 1) define the scope; 2) select information sources; 3) select terms; 4) vocabulary standardization; 5) use the vocabulary (execute); 6) evaluate and review. This article also addresses issues related to the requirements of developing a standardization projects, in particular goals and scope; dimension/coverage; target audience; timeline; data structure; procedures; terminology; human resources; software; evaluation and maintenance. Finally, some prospects for the future are proposed, among which the development of the toolkit to other controlled vocabularies (e.g., taxonomies and thesaurus), the promotion of this tool at the national and international level or the organization of training actions for the professional community, promoting the sharing of experiences and resources.

Keywords: Knowledge organization, Controlled vocabularies, Controlled lists, Cultural heritage institutions, Information Systems in Museums

1 The Information Systems in Museums Working Group

The Information Systems in Museums Working Group (WG) was formed in 2012 within the Portuguese Association of Librarians, Archivists and Information Professionals (BAD) [1]. This WG is organized into task forces to develop documentation and standardization projects. In the framework of the knowledge organization and the controlled vocabularies has particular importance the Terminology task force, whose goals are: create technical tools and documents; provide guidelines for the use of controlled vocabularies; organize seminars and workshops; publish scientific literature related to the controlled vocabularies; provide dissemination projects related to controlled vocabularies and creation of tools in Portuguese language.

All technical and scientific production of the WG is available in several platforms (such as Zotero and Zenodo) [2-3], with emphasis on the translation of international normative documents (e.g., Spectrum 5.0) and technical guides like the practical Guidelines about controlled vocabularies for knowledge organization in cultural heritage institutions [4].

2 Standardization in cultural heritage institutions: what for?

In this context, it is important to reflect on the relevance of standardization in cultural heritage institutions, considering in particular the following elements: provide categorization, indexing and retrieval of information; manage collections data; ensure access to collections by users (internal and external); and enable interoperability between similar information systems [5].

In this domain, the WG takes as a reference some basic and framework concepts established by the ISO 5127:2017: Information and Documentation-Foundation and vocabulary [6]. Therefore, the designations "controlled vocabulary", "concept", "term" and "controlled/pick list used in the scope of the WG projects are in line with the spirit of this same standard.

3 Requirements for the development of standardization projects

Before developing a standardization project, a set of essential elements should be defined, in particular: goals and scope; dimension/coverage; target audience; timeline; data structure; procedures; terminology; human resources; software; evaluation and maintenance. In this context, it is advisable to set up a working group to manage the project and monitor its different stages [4-5,7].

4 Controlled lists to organize and manage information: a toolkit in 6 steps

The WG has developed a toolkit in six steps to build controlled lists to organize and manage information in cultural heritage institutions. This toolkit is only a proposal and, as such, it is possible to be discussed, adjusted and developed [8-9, 15-16]. For the construction of the toolkit, we followed the guidelines of the "Forward Planning Toolkit" of the Arts Council England [10] and, naturally, the guidelines prescribed in the standards related to the controlled vocabularies [11-14].

Currently, the WG is working in a more extensive version of the toolkit, which will be released shortly. At this stage, the toolkit comprises six steps and each of them has several different actions, specifically:

- Step 1: Define the scope:
  - Action 1: Define what data in the Information System (IS) requires terminological control/controlled vocabularies;

  - Action 2: Choose a project manager and create multiple WG to analyze in detail each data category which must have terminological control.
- Step 2: Select information sources:
  - Action 1: Search, identify and select controlled vocabularies (already existing) of the thematic scope of the collections;
  - Action 2: Approach similar institutions to find out which tools they use, in the case they use them;
  - Action 3: Comply with the requirements of the national and international standards of the scope of the KOS – Knowledge Information Systems [12-14].
- Step 3: Select terms:
  - Action 1: Select the terms representative of the existing themes of the collection, by each data category;
  - Action 2: Create a preview alphabetic list of terms extracted from the information sources;
  - Action 3: Discuss the first version of the list of terms between the several working groups.
- Step 4: Vocabulary standardization:
  - Action 1: compare the selected terms with the concepts already existing in natural language in the database, if they already exist;
  - Action 2: remake the list with the preferred terms collected from the selected controlled vocabularies and the concepts that do not exist in the selected controlled vocabularies. These concepts will have to be standardized;
  - Action 3: formal standardization (morphological control - language, gender and number / syntactic control (structure of the terms - simple/compound nouns);
  - Action 4: semantic standardization (ambiguity control/ homographs - use qualifiers/scope notes – and synonyms or quasi-synonyms - use hierarchical relations).
- Step 5: Use the vocabulary (execute):
  - Action 1: Insertion of the list of terms in the IS;
  - Action 2: Create a procedure manual to manage the use of the list of terms, especially the fields that will use controlled language;
  - Action 3: Select one sample of records and apply them the list of terms already standardized to register and categorize information, by category of data;
  - Action 4: Discuss the results within the different WG's of the project.
- Step 6: Evaluate and review:
  - Action 1: Develop some research and retrieval experiences with the records that make up the sample;
  - Action 2: In group, identify inconsistences and make adjustments/corrections to the controlled list of terms.

5 Prospects for the future

In the mid-term, the WG intends to achieve the following actions: develop, test and adjust the toolkit in different cultural heritage Portuguese institutions; extend the toolkit to other controlled vocabularies, namely taxonomies and thesaurus; promote this tool, at the national and international level, through participation in seminars, webinars and workshops; organize training actions for the professional community, promoting the sharing of experiences and resources; continue to translate international reference standards related to knowledge organization and controlled vocabularies.

References

1. Grupo de Trabalho Sistemas de Informação em Museus, https://www.bad.pt/noticia/category/informacaomuseus/, last accessed 2021/07/10.
2. Grupo de Trabalho Sistemas de Informação em Museus - Zotero Library, https://www.zotero.org/groups/81851/gt-sim, last accessed 2021/07/10.
3. Grupo de Trabalho Sistemas de Informação em Museus – Zenodo, https://zenodo.org/communities/gt-sim_bad?page=1&size=20, last accessed 2021/07/10.
4. Jorge, N., Medeiros, F., Alves, J., Medina, S. Os vocabulários controlados na organização e gestão de informação sobre património cultural: orientações práticas (2017), https://www.zotero.org/groups/81851/gt-sim/search/Nat%C3%A1lia/titleCreatorYear/items/R6XXVEVD/item-list, last accessed 2021/07/10.
5. Harpring, Patricia. Introdução aos vocabulários controlados: terminologia para arte, arquitetura e outras obras culturais. São Paulo, Secretaria da Cultura do Estado (2016).
6. ISO 5127:2017 Information and documentation — Foundation and vocabulary. Geneva, ISO (2017).
7. Smit, J. A informação no museu: não basta organizar, ainda é preciso preservar e comunicar! In: II Seminário Serviços de Informação em Museus, pp.349-354. São Paulo, Pinacoteca do Estado de São Paulo (2012).
8. Jagerman, E. Creating, maintaining and applying taxonomies. Zoetermeer, Author's Ed. (2006).
9. Moreiro González, J. Linguagens documentárias e vocabulários semânticos para a web: elementos concetuais. Salvador, EDUFBA (2011).
10. Arts Council England (ACE). Forward Planning Toolkit (2009), https://www.artscouncil.org.uk/advice-and-guidance-library/toolkits, last accessed 2021/07/10.
11. NP 4036 – Tesauros monolingues: directivas para a sua construção e desenvolvimento. Lisboa, Instituto Português da Qualidade (1992).
12. Z3919:2005. Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies. Bethesda, Maryland, NISO Press (2005), http://www.niso.org/standards/index.html [En línea]: http://www.niso.org/standards/resources/Z39-19-2005.pdf, last accessed 2021/07/10.
13. ISO 25964-1: 2011. Information and documentation - Thesauri and interoperability with other vocabularies - Part 1: Thesauri for information retrieval. Geneva, ISO (2011).
14. ISO 25964-2. 2013. Information and documentation - Thesauri and interoperability with other vocabularies - Part 2. Interoperability with other vocabularies. Geneva, ISO (2011).
15. CIDOC. Diretrizes internacionais de informação sobre objetos: categorias de informação do CIDOC. In Grant A., Nieuwenhuis, J., Petersen T. (eds.) Declaração de Princípios de Documentação em Museus e Diretrizes internacionais de informação sobre objetos: categorias de informação do CIDOC, pp. 23-76. São Paulo, Secretaria de Estado de Cultura; Associação de Amigos do Museu do Café; Pinacoteca do Estado de São Paulo (1995/2014).
16. Zaharee, M. Building controlled vocabularies for metadata harmonization. Bulletin of the American Society for Information Science and Technology. 39 (2), 6 p. (2013).

Vocabulaires de recherche, vocabulaire contrôlé et modèle de données : une chaîne opératoire pour le partage des données archéologiques

Sébastien Durost[1], Guillaume Reich[2][0000-0001-5268-5708], Jean Pierre Girard[3][0000-0003-4225-8678]

[1] BIBRACTE EPCC - Centre archéologique européen, F-58370, Glux-en-Glenne
s.durost@bibracte.fr
[2] Maison des Sciences de l'Homme et de l'Environnement Claude-Nicolas Ledoux - UAR 3124, CNRS, Université Bourgogne Franche-Comté, F-25000, Besançon
dr.guillaume.reich@gmail.com
[3] Archéorient - UMR 5133 - Environnements et sociétés de l'Orient ancien - Maison de l'Orient et de la Méditerranée - Jean Pouilloux – F-69365, Lyon
jean-pierre.girard@mom.fr

Abstract. Un consortium d'acteurs français de l'archéologie mène, depuis trois ans, une démarche pour trouver, accéder, interopérer et réutiliser (FAIR) des données archéologiques. Le vocabulaire est apparu comme le pivot de l'intercompréhension des travaux et de leurs comptes-rendus. Les projets *HyperThésau* et *Bibracte numérique* ont donc posé l'usage du vocabulaire comme axe de l'interopérabilité des données archéologiques tout au long de leur cycle de vie. Pour ce faire, l'usage de la forme normalisée du thésaurus – via la plateforme open source *Opentheso* – fournit un outil déjà adapté au web sémantique, qui devient le support d'une chaîne d'interopérabilité « humain-machine-humain » complète, mise au point dans le cadre du projet *Bibracte Ville Ouverte*. Pour aller plus loin vers l'interopérabilité "*machine to machine*", les obstacles en matière d'alignement entre les thésaurus documentaires et un thésaurus de la pratique scientifique soulignent le besoin d'un espace sémantique faisant office de « pivot » entre les deux univers. Cette voie n'est pas sans soulever un défi : l'espace d'incertitude inhérent à la prise en compte dans les alignements de concepts scientifiques proches et néanmoins distincts. La résolution des incertitudes linguistiques supposera sans doute de faire appel à des outils d'intelligence artificielle appliqués au contenu sémantique des définitions et à l'analyse des objets graphiques associés.

Keywords: archéologie, thésaurus, interopérabilité.

1 Vers le partage des données archéologiques

1.1 Une approche centrée sur le vocabulaire

Un consortium d'acteurs français de l'archéologie mène, depuis trois ans, une démarche pour trouver, accéder, interopérer et réutiliser des données archéologiques (*FAIR data*). Les principes en ont été exprimés en 2018 pour le projet HyperThésau, porté par le laboratoire Archéorient (CNRS-univ. Lyon 2). En centrant la réflexion sur le rapport de l'humain-utilisateur à des "objets de savoir" archéologiques en contexte numérique, a été abordée la question du cycle de vie des données acquises, organisées, étudiées, archivées et partagées par des professionnels : un flux collaboratif scientifique pour l'enrichissement sémantique et la documentarisation de données archéologiques. Le projet *HyperThésau* (2018-2020) a permis la constitution de "micro-thésaurus pivots" pour l'archéologie, issus des pratiques-métier des équipes de recherche pour des sous-ensembles de la chaîne opératoire et alignés sur les grands référentiels du web sémantique (Library of Congress Subject Headings, data.bnf.fr, IdREf, etc.). Cette démarche a ensuite été mise en pratique par le Centre européen d'archéologie de Bibracte dans le cadre du projet *Bibracte Numérique*, un programme pluriannuel (2018-2021) de développement d'outils numériques pour ses différents métiers et les différentes catégories de ses usagers (archéologues, étudiants, chercheurs et grand public).

Le vocabulaire nous est apparu comme le pivot de l'intercompréhension des travaux et de leurs comptes-rendus. En effet, les seules données réellement "brutes" sont l'unité stratigraphique ou le vestige exhumé, tel l'objet physique que l'on peut directement observer. Dans la pratique ne sont ensuite manipulées que des "traces primaires", descriptions graphiques (minute dessinée ou orthophotographie, par ex.) ou textuelles. Or, Il n'existe pas de norme ISO pour la représentation de la donnée archéologique, ni de consensus en tenant lieu. « Lire » des libellés revient donc à les interpréter, selon différents paradigmes, dans un univers sémantique à définir [1].

L'usage d'un thésaurus (norme ISO 25964). Les projets *HyperThésau* et *Bibracte numérique* ont donc posé l'usage du vocabulaire comme axe de l'interopérabilité des données archéologiques tout au long de leur cycle de vie. Pour ce faire, l'usage de la forme normalisée du thésaurus – via la plateforme *Opentheso* – fournit un outil déjà adapté au "Web des données" [2]. Néanmoins, son usage a rapidement soulevé la question des paradigmes présidant à l'élaboration d'un vocabulaire propre, par chaque (groupe de) scientifique(s). La norme ISO 25964, conçue pour la gestion et l'interopérabilité des langages documentaires, se révèle assez souple pour être mise en œuvre selon différents "points de vue". Mais la mise en cohérence de ces derniers au moyen d'alignements permettant cette interopérabilité nécessite une méthodologie de coopération régulée en vue d'interfacer différentes granularités sémantiques : le signalement des recherches, la description des données "primaires" (*raw data*), une passerelle ou "pivot" entre les deux. Les défis qui restent à relever sur cette voie n'empêchent pas l'outil thésaurus d'être, d'ores et déjà, un support adapté à une interopérabilité complète entre scientifiques, dont la chaîne opératoire a été mise au point dans le cadre du projet *Bibracte Ville Ouverte*.

1.2 Vocabulaire de recherche ou thésaurus documentaire ?

Selon la nature de sa problématique, l'archéologue sélectionne et observe plus particulièrement certains paramètres intrinsèques de son objet d'étude, qui lui semblent pertinents pour son travail général de modélisation de la connaissance. Il utilise les plus petites unités descriptives congruentes pour rendre compte de ce qu'il observe sur le terrain ou sur le mobilier, puis les hiérarchise en fonction de la résolution ou granularité de sa problématique initiale. L'agglomération des différentes caractéristiques repérées lui permettent de construire un *type*, c'est-à-dire une réalité matérielle couplée avec une définition univoque (vocabulaire et généralement une ou des représentations iconographiques). Tout ce réseau d'informations est synthétisé par l'archéologue dans des typo-chronologies (inscription des types sur une échelle temporelle) dont les concepts s'appuient sur des définitions très détaillées. À l'inverse, pour la création d'index bibliographiques structurés selon la norme ISO-25964-1, le contexte socio-technique de leur mise en œuvre [3] a focalisé toute l'attention des documenta-listes sur la différenciation intralinguistique des libellés, la sémantique n'étant qu'une apparence formelle, une instance du dyptique « organisation + message » [4] dépourvue de signification réelle.

Expérimentation. À partir d'une publication de synthèse des artefacts céramiques du site de Bibracte [5], un prototype d'arborescence explicitant les besoins sémantiques spécifiques de l'archéologue a été formalisé et décrit dans un thésaurus nommé *Bibracte_Thesaurus* (http://ark.mom.fr/ark:/39676/srvtxcg5zrhv8). Cette expérimentation a permis de comprendre que les deux approches – celle de l'archéologue et celle du documentaliste – sont irréductibles l'une à l'autre, en raison d'une intentionnalité d'usage différente de la norme ISO 25964-1 : pour l'archéologue, le concept est la somme d'une définition, d'un identifiant et d'un terme préféré placé au sein d'une arborescence. C'est la définition associée au terme préféré (libellé) qui permet de différencier et de hiérarchiser les concepts pour rendre compte du raisonnement par des relations hiérarchiques ou associatives. L'ensemble de ces relations forme un graphe répertoriant et contextualisant des ensembles de concepts (référentiel archéologique) et traduisant un état du savoir, par nature situé dans le temps/millésimé, dans une branche du *Bibracte_Thesaurus*.

2 Une première étape opérationnelle

2.1 Une chaîne opératoire XML

Sur cette base a été construite une chaîne opératoire, appuyée sur une modélisation en XML, un vocabulaire contrôlé (hiérarchie formelle + définition) et le gestionnaire de thésaurus *Opentheso*, sur un dépôt ouvert (Nakala) et des métadonnées DublinCore incluant des identifiants pérennes (URI). L'ensemble constitue une chaîne “FAIR” humain-machine-humain, qui permet le partage et la réutilisation effective des données entre et par les scientifiques mais dont la modélisation XML, fondée sur le vocabulaire, n'est pas appariée avec une ontologie normalisée telle que le Cidoc-CRM.

2.2 Défis méthodologiques

Pour aller plus loin vers l'interopérabilité "*machine to machine*", une tentative d'alignement entre les deux modèles d'usage de la norme a été opérée, en rapprochant, toujours via l'exemple de la céramique, le vocabulaire en usage à Bibracte [5] avec le thésaurus d'indexation bibliographique PACTOLS 2. Cette tentative a rencontré plusieurs difficultés épistémologiques liées à l'écart ou à la tentative de résorber l'écart entre deux pratiques, scientifique et documentaire, irréductibles l'une à l'autre, ce qui empêche un alignement complet et satisfaisant (*exactMatch*).

La nécessité d'un vocabulaire "pivot". Les obstacles en matière d'alignement entre les deux systèmes de pensée et de pratique soulignent le besoin d'un espace sémantique faisant office de "pivot" [6] et susceptible de générer une forme de consensus, tel que le thésaurus HyperThesau (https://thesaurus.mom.fr/). Ce niveau de normalisation sémantique doit être construit dans un dispositif de coopération régulée qui associe un ou des locuteurs de la langue du domaine, spécialistes-experts (ici : des archéologues) et un ingénieur de la connaissance maîtrisant l'usage de la norme dont le rôle n'est pas celui d'arbitrer des choix mais d'être un facilitateur méthodologique de l'élaboration d'une sémantique réellement collégiale [1].

Prendre en compte l'incertitude. Cette voie nous confronte à un autre défi, plus dur à relever : l'espace d'incertitude inhérent à la prise en compte de « points de vue » scientifiques proches et néanmoins distincts. Un précédent inspirant existe toutefois : le référentiel PeriodO (https://perio.do/en/) qui organise, non pas des concepts de périodes *in abstracto*, mais des définitions sourcées de périodes associées à un intervalle de dates, ce qui rend cette incertitude calculable et permet l'organisation automatisée des « points de vue » chronologiques. La prise en compte d'incertitudes linguistiques supposera sans doute de les rendre pareillement calculables, donc de faire appel, peu ou prou, à des outils d'intelligence artificielle (IA) appliqués tant à la structuration des concepts (graphes) qu'au contenu sémantique de leurs définitions et à l'analyse des objets graphiques associés.

References

1. Bachimont B. Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en Ingénierie des connaissances. In : Charlet J., Zacklad M., Kassel G. et Bourigault D. (éds.), Ingénierie des connaissances, évolutions récentes et nouveaux défis, Eyrolles., Paris (2000). [https://www.utc.fr/~bachimon/dokuwiki/_media/fr/ontologie-icbook.pdf]
2. Hudon M. ISO 25964 : pour le développement, la gestion et l'interopérabilité des langages documentaires. In : Documentation et bibliothèques, 58 (3), Paris (2012), pp. 130–140. [https://www.erudit.org/fr/revues/documentation/2012-v58-n3-documentation01721/1028903ar.pdf]
3. Stiegler B., La Mémoire et le temps, t. 1 : la faute d'Épiméthée, Éditions Galilée, Paris (1994) - La Mémoire et le temps, nouvelle édition augmentée ,Fayard, Paris (2018).
4. Shannon C. E., Weaver W., A Mathematical Theory of Communication, Bell System Technical Journal 27(3), New York (1949). [http://people.math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf]

5. Barrier S., Luginbühl T. La vaisselle céramique de Bibracte. De l'identification à l'analyse, Bibracte 31, Glux-en-Glenne (2021).
6. Perrin E., Girard J.-P., Rousset M.-O., Durost S. Thésaurus et terminologies aux sources de l'interopérabilité des données archéologiques. In : Actes de l'Atelier DAHLIA Conférence EGC, Bruxelles (2020).
   [https://www.egc.asso.fr/wp-content/uploads/egc2020_atelier_Dahlia.pdf]

Thesaurus de Acervos Científicos em Língua Portuguesa: building a controlled vocabulary in a collaborative mode

Susana Medina[1]

[1] Museu da Faculdade de Engenharia da Universidade do Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal / CITCEM – Centro de Investigação Transdisciplinar Cultura, Espaço e Memória - Faculdade de Letras da Universidade do Porto,

smedina@fe.up.pt

Abstract. This paper will present a fruitful collaboration between the Museum of Astronomy and Related Sciences in Rio de Janeiro, the National Museum of Natural History and Science of the University of Lisbon and other ten Portuguese and Brazilian university museums: the *Thesaurus de Acervos Científicos em Língua Portuguesa* project. This long cooperation (2006-2011) has contributed to the promotion of cultural heritage of science and technology between institutions from Brazil and Portugal. A synthesis of the joint efforts and main outcomes of the project seemed important and relevant; so this paper will present a brief description of the task and will outline most relevant aspects of the collaborative network benefits for projects of this kind.

Keywords: Scientific Heritage, Controlled Vocabularies, Collaborative Network.

1 Introduction

Preserving scientific heritage is a major challenge in present-day society. Dispersed through a multitude of institutions – from universities to schools and research laboratories – and unprotected by cultural heritage legislation, the preservation of scientific heritage needs to gradually change from a museum-based approach to an attitude increasingly oriented towards information management and access, sustained by networks and partnerships at national and international scale [1].

These museum sector networks tend to be informal, in the sense that they are usually based on personal or professional contacts. For the purposes of this paper we may define these networks as a combination of professional and inter organizational relations, professional knowledge and work processes brought together to achieve a common purpose. They constitute communication channels that give museum staff access to information and knowledge about assets and collections.

However, communication barriers may have a significant impact on museum's professional networks. The way people interact with each other can vary depending on demographic and cultural differences. Communication barriers among science and technology museums may also occur due to problems of terminological clarification and standardization, related to general terms such as collection, collection, cataloging, numbering, documentation, inventory, etc., or specific terminology for science and technology museums that made it difficult to identify and classify their collections [2].

University museum professionals from Portugal and Brazil were well aware of these barriers. From 2006 to 2011, and inspired by similar experiences in France and Italy, Museu de Astronomia e Ciências Afins from Rio de Janeiro (MAST) and Museu de Ciências da Universidade de Lisboa (MCUL) were involved in the production of a thesaurus of scientific instruments in Portuguese. A network of institutions undertook the task of creating a terminology control and access tool for museums or other institutions with collections of scientific instruments:

- from Portugal, and coordinated by Marta Lourenço (MCUL) - Museu de Ciência (Coimbra), Museu de Ciência (Porto), Museu da Faculdade de Engenharia (Porto), Museu de Física do Instituto Superior de Engenharia (Lisboa), and Museu do Instituto Superior de Engenharia (Porto);
- from Brazil, and coordinated by Marcus Granato (MAST) - Colégio Pedro II, Museu da Escola Politécnica (UFRJ), Museu Dinâmico de Ciência e Tecnologia (UFJF/Juiz de Fora), Centro de Memória da Farmácia (UFOP/Ouro Preto), Museu de Ciência e Técnica da Escola de Minas (UFOP/Ouro Preto), and Museu da Farmácia Lucas Marques do Amaral (UFJF/Juiz de Fora).

2 Goals and objectives

The basis of the project entailed the following goals and objectives, in order to develop a terminological thesaurus for scientific collections that could be an information retrieval tool, a facilitator of communication between science and technical museums within the Lusophone sphere, especially Portugal and Brazil [2]:

- to identify terms used in historical scientific collections, namely teaching and researching scientific collections of exact sciences, mathematics and engineering. Medicine and natural history collections were not included in the project;
- to classify, standardize, control and relate terminology;
- to define standard images to represent each scientific instrument, based on the network's own collections;
- to create a dynamic thesaurus with scope notes;
- to make results free and universally available in 2011, to the museum and scientific community, through an website, a book and a DVD.

3 Methodology

The thesaurus was developed by a bottom-up approach, involving the already mentioned institutions.

Since the beginning of the project, working groups within the network were formed to accomplish specific tasks. The working groups had the benefit of a broad range of experience, skills and points of view.

A group of subject experts to serve as advisors was also created and the *Thesaurus des Objets Mobiliers*, also known as Palissy Base [3] was chosen as a reference. The lists of designations and object definitions provided by network partners were thesaurus' primary sources. Original function of the objects (used for...) was adopted as a classification criterion and the project started with the definition of objects' function of the objects in a working glossary.

Lists of terms were then grouped, studied, reviewed and organized, using a variety of resources as aids. Definitions, concepts and relationships were studied in order to establish the thesaurus structure and partners (individually, in working groups and plenary sessions) produced texts and comments, debated terms whose meaning or usage was unclear and provided object-related images.

A draft thesaurus was finally produced, index tested and revised, before it was published.

4 Project development

During project collaboration, there were key moments that had shown important achievements and the completion of major phases within the project. The following Table 1 gives a summary of the main activities and results of each session, already established in other sources [1] [2]:

Table 1. Key moments of project development

Session	Activities	Results
BR team workshops **PT team workshops** 2008 - 2010	Data collection and guidelines to provide introduction and advice on how to build hierarchical tree structures (following standards and practices in partners' museums); Selection and organization of the terms that were included in each team's designations lists from the contributions of the institutions involved (Purge List I); Definition of objects' function; term purging criterions to reduce original listings; eliminating duplicate, redundant, current and out of scope terms,	Study and selection of terns from the instruments list; study and selection of terms from the machines list; thesaurus' structure (technical standards); criterion and procedures manual; image database; study on digital platform for thesaurus; working glossary and thesaurus explanatory notes; activities report and updated procedures manual.

	as well as accessories, parts and cases; Terms arranged into a hierarchical tree structure using parent/child, whole/part or instance relationships (broader term (BT) and narrower term (NT) to denote these relationships; Creation of an Working Group (WG) to refine the list, pre-organize synonyms, related terms and complex names; initial list reduced to a new one (Purge List II); Search for missing terms in other sources about scientific instruments in catalogues of other Science and Technology museums and manufacturers; compiled terms integrated into the Purge II list. Visit to partner museums.	
PT and BR teams workshops Lisboa, 2009 Rio de Janeiro, 2010	Joint working sessions. Merging PT and BR lists and carrying out new purges related to duplicate terms; Debate on adopted methodology and thesaurus structure; Scientific validation, structure and final list of index terms; Comparison and development of a software within a public-private partnership; Classification by knowledge areas based on the classifications that were used in partner's museums and Brazilian National Council of Scientific and Technological Development (CNPq); Creation of the most specialized WGs (Classification, Images, Scope Notes and Glossary, Sources). International seminars.	Merged PT and BR lists; final classification and top terms list; criteria for scope notes, image bank and top terms definitions; final revision of scope notes before experts' validation; updated methodology and procedures manual; thesaurus' software requirements specification (free access and stability). Project website launched. (http://chcul.fc.ul.pt/thesaurus/) *Coleções Científicas de Instituições Luso-Brasileiras: patrimônio a ser descoberto* published [4].

BR team workshops **PT team workshops** **WG coordinator's meeting** 2010 – 2011	Conclusion of Thesaurus (first edition); Thesaural relationships, features and notes were supported and tested in a trial run and updated if records were amended; a review of the constructed thesaurus was undertaken by subject experts and professional users. The review highlighted some gaps in the thesaurus, missing or redundant features as well as some usability issues that needed to be fixed before dissemination.	*Thesaurus de Acervos Científicos em Língua Portuguesa* published (http://thesaurusonline.museus.ul.pt/) in a recognised format that can be reused by other professionals and cultural heritage institutions, both in Lusophony space and communities in the world.

5 Conclusions

The experience resulting from this collaborative network brought many benefits to the involved museum professionals, as it was a source of positive motivation and brought together Brazilian and Portuguese professionals.

As stated by its coordinators [2], the formation of the project network proved to be as important as the thesaurus, bringing together institutions that, for multiple reasons, could hardly have their collections inventoried, registered and researched, revealing unknown collections, revaluating and setting institutional recognition of others.

In addition, the described network of like-minded peers is still a good source of advice among participants of the project.

References

1. Granato, M. et al.: Thesaurus de acervos científicos em língua portuguesa: concepção e resultados preliminares. In: Anais do XI Encontro Nacional de Pesquisa em Ciência da Informação. IBICT, Rio de Janeiro (2010).
2. Granato, M., Lourenço, M.C.: Preservação do patrimônio cultural de ciência e tecnologia: uma parceria luso-brasileira entre o Museu Nacional de História Natural e da Ciência (Portugal) e o Museu de Astronomia e Ciências Afins (Brasil)". Ciência da Informação, 42 (3), 435 - 453. (2013).
3. Ministère de la Culture et de la Communication. Thesaurus des Objets Mobiliers. Éditions du Patrimoine, Paris (2001).
4. Granato, M., Lourenço, M. (eds.): Coleções científicas luso-brasileiras: património a ser descoberto. Museu de Astronomia e Ciências Afins, Rio de Janeiro (2010)

“Tu que me lês,”

Teresa Barreto Borges[1]

[1] Cinemateca Portuguesa-Museu do Cinema, IP, Lisboa, Portugal
teresa.borges@cinemateca.pt

Resumo. Apresentação do trabalho de desenvolvimento de um *thesaurus* para a indexação das imagens em movimento.

Palavras-chave: Thesaurus, Filme, “Não-filme”.

No âmbito do projeto ROSSIO, inscreveu a Cinemateca Portuguesa-Museu do Cinema o desenvolvimento de um *thesaurus* para a descrição e a posterior recuperação da informação das obras de imagens em movimento, até agora descritas por sinopses, por resumos e por palavras-chave.

O modelo conceptual anteriormente decidido (e já refletido na criação e associação dessas palavras-chave prévias) tomava como ponto de partida o Thesaurus FIAF[1], desenvolvido pela Comissão de Documentação da Federação no final dos anos 70 do século passado para a descrição de conteúdos de coleções bibliográficas sobre cinema e televisão. Na Cinemateca Portuguesa, o Centro de Documentação e Informação começou por utilizar este *thesaurus* na sua versão em língua francesa, implementando em 1981 a sua própria versão em língua portuguesa – traduzida por Rui Santana Brito[2] –, e que desde então tem sido revista e atualizada.

Trata-se de um *thesaurus* desenvolvido para, repete-se *e sublinha-se*, ser aplicado a documentos bibliográficos sobre cinema e televisão, abrangendo todos os aspetos da história, teoria, estética e técnica destes domínios. Caracterizando-o brevemente, apresenta uma lista alfabética estruturada e listas de subáreas temáticas (nomes de pessoas e de instituições, nomes de personagens); para controlo dos termos, contém descritores com e sem definição, simples e compostos, listas de não-descritores e notas de aplicação; relativamente à forma, os termos preferenciais são apresentados em maiúsculas e os não-descritores em minúsculas; finalmente, ao nível semântico, apresenta relações de equivalência, hierárquicas e de associação.

No momento em que a Cinemateca procurava reunir as áreas Filme e “Não-filme” num único sistema de informação sob o mote de “cooperar e partilhar”, esta opção, ainda que prévia, seguia este mesmo princípio: desenvolver o *thesaurus* pré-existente e em uso de modo a abranger a indexação de imagens em movimento.

[1] FIAF - Federação Internacional dos Arquivos de Filmes, fundada em 1938 e de que a Cinemateca é membro desde 1956.

[2] Rui Santana Brito (1944-2017), Chefe de Divisão do Centro de Documentação e Informação de 1989 a 1997 e Vice-Presidente da Cinemateca de 1997 a 2005, data em que se reformou.

A hipótese em debate foi então a de estabelecer relações partitivas ou todo-parte neste *thesaurus* combinado (ou desdobrado) da Cinemateca, em que os descritores aplicáveis ao universo "não-filme" operariam como termos de topo dos descritores aplicáveis ao universo "filme"[3]. Tal hipótese contém igualmente em si o princípio de que as imagens tanto são genéricas como específicas[4], ou seja, o princípio de que, por exemplo, uma imagem da Ponte 25 de Abril é também, genericamente, uma imagem de uma ponte.

Verificando-se a necessidade de clarificação – tanto para o editor da informação (indexador) como para o utilizador que pretende recuperar essa informação – a qual das áreas se refere o uso de determinado descritor, a proposta prosseguiu com a opção pela diferenciação da composição textual dos conceitos. Uma vez que no *thesaurus* FIAF os termos compostos incluem os distintivos "preposições-nome" (exemplo: PONTES *NOS FILMES*) e "conjunções-nome" (exemplo: PINTURA *E CINEMA*), estabeleceu-se que a omissão destes distintivos indicaria a parte dentro desse todo[5].

Retomando o exemplo "Ponte", teremos assim a seguinte estrutura:

PONTES NOS FILMES
TEP: PONTE 25 DE ABRIL

O modelo proposto terá ainda de ver provada a sua ampla aplicabilidade às coleções de imagens em movimento da Cinemateca, designadamente pela implementação sistemática da metodologia proposta para a indexação de imagens em movimento que, baseada em camadas de informação, procura responder às questões: *O que se vê? O que se ouve? O que se lê?* Paralelamente, ao nível dos futuros desenvolvimentos deste vocabulário para a descrição e representação de conteúdos de imagens em movimento, prevê-se desde já a necessidade de estabelecer uma clara distinção entre imagens "reais" e imagens ficcionais, isto é, entre imagens de um determinado acontecimento (registo direto, local + data) e a sua representação ficcional.

[3] Na verdade, onde se lê "filme" poder-se-á ler "imagem".

[4] Sobre esta questão, veja-se o trabalho de investigação absolutamente determinante realizado por Sara Shatford Layne.

[5] O trabalho desenvolvido pelos investigadores do projeto ROSSIO que connosco colaboraram consistiu inicialmente na análise dos filmes (através das respetivas representações digitais publicadas no sítio web da Cinemateca) e seleção dos descritores para a representação do seu conteúdo. Posteriormente, procederam à identificação dos conceitos do *thesaurus* FIAF que seriam desdobrados nas partes correspondentes e ao estabelecimento das relações todo-parte.