Knowledge organisation in institutional repositories: A case study on policies and procedures manuals in the Ibero-American environment

Abstract

Purpose - The aim of this work is to analyse the recommendations on knowledge organisation from guidelines, policies, and procedure manuals of a sample of institutional repositories and networks within the Latin American area and observe the level of follow-up of international guidelines.

Methodology – Presented is an exploratory and descriptive study of repositories' professional documents. The study comprised four steps: 1. definition of convenience sample; 2. development of data codebook; 3. coding of data; and 4. analysis of data and conclusions drawing. The convenience sample includes representative sources at three levels: local institutional repositories, national aggregators, and international network and aggregators. The codebook gathers information from the repositories' sample, such as: institutional rules and procedure manuals openly available, or recommendations on the use of controlled vocabularies.

Findings - The results indicate that at the local repository level, the use of controlled vocabularies is not regulated, leaving the choice of terms to the authors' discretion. It results in a set of unstructured keywords, not standardised terms, mixing subject terms with other authorities on persons, institutions, or places. National aggregators do not regulate these issues either, and limit to pointing to international guidelines and policies, which simply recommend the use of controlled vocabularies, using URIs to facilitate interoperability.

Originality - The originality of the study lies in identifying how the principles of knowledge organisation are effectively applied by institutional repositories, at local, national, and international levels.

Keywords: Knowledge organisation, Institutional repositories, Academic repositories, Policies, Procedure manuals, Guidelines, OpenAIRE, Oasisbr, COAR, RCAAP, Recolecta, LA Referencia

Article classification: Research paper

1. Introduction

Repositories emerged in the late 20th century as a new strategy of academic institutions with the aim of storing, managing, and preserving huge collections of resources together with the metadata describing them. Academic repositories began to develop the role of publishers by updating the process of scholarly communication (Villalobos and Gomes, 2018) and improving the dissemination of their intellectual production, through open access to knowledge. Nowadays, the number of repositories and the consequent availability of information have increased enormously; for example, OpenDOAR statistics (JISC, 2022).

Open access to scientific publications in digital format has been emphasised worldwide, mainly in university institutions that need to have control of scientific production for various reasons, but, above all, to give visibility to external evaluators, information systems, classification systems, and the academic community itself. The storage, organisation, and dissemination of scientific information and publicly accessible data have been emphasised globally with the aim of discussing policies to minimise the problems of sharing and reuse of scientific data and publications.

The repository has much importance in the intellectual life and scientific output of an institution (Kounoudes and Zervas, 2011) and is recognized as an essential infrastructure which is in the process of developing guiding principles and best practices.

The practices of institutions, such as libraries, archives, and museums (LAMs), have always been improved by knowledge organization, as they organise, catalogue, and classify digital resources in the long term, and this guarantees systematicity and, mainly, better information retrieval and data management systems. Systematic organisation of data and information is a universal need in the digital age, and lack of knowledge or unfamiliarity with knowledge organisation can lead to problems of sustainability, scalability, or traceability (Gollub *et al*., 2021).

Like digital libraries, academic repositories have specific approaches to information management and need to organise knowledge through representation techniques, which are important for information retrieval (Hjørland, 2021). Today, academic libraries coexist with institutional repositories and usually share administrative functions and structures. Although originally responding to different needs, this close co-existence in administrative structures produces synergies and mutual improvements. While libraries are very much oriented to the technical process of collection management (bibliographic selection to support reference and cataloguing services), repositories focus on digital preservation, availability, and accessibility.

Both information systems have common strategies on accessibility, interoperability, and standards adoption. Therefore, they share basic knowledge organisation processes, such as the attribution of metadata to support access to collections. Villalobos and Gomes (2018) highlighted the lack of efforts to standardise metadata in repositories. Knowledge representation and organisation can be a means to enhance the value of digital repositories (Araújo and Silva, 2021).

A fundamental issue that differentiates repositories from library catalogues is that the former are based on providing the information object, the content, together with the structured metadata describing it. In the case of library catalogues, only access to the metadata, which provides structured information about the object, is offered. By focusing on access to objects, repositories have paid less attention to the knowledge organisation aspects of their resource's collections.

The importance of developing professional documents (repository policies, guidelines, and procedural manuals) is highlighted by Villalobos and Gomes (2018). These documents should include aspects such as: metadata, information processing, controlled vocabularies, auxiliary tables, and thesauri (Sousa, 2012). The use of natural language and keywords produces fewer effective results when compared to a controlled vocabulary search. The larger the collection of a repository, the more sophisticated the subject-controlled vocabulary should be (Phillips *et al*., 2020).

The authors of this paper continue previous research on the organisation of knowledge in institutional repositories (Fujita and Tolare, 2019; Terra *et al.,* 2021) and its treatment in the scientific literature (Fujita *et al.*, 2021, 2023). As a result, the need to also analyse repositories' professional documents was identified.

The aim of this work is to analyse the recommendations on knowledge organisation from the guidelines, policies, and procedure manuals of a sample of institutional repositories and networks within the Latin American area, and observe the extent to which the international recommendations and guidelines on knowledge organisation are followed for the thematic characterization and classification of their contents. Presented is an exploratory study on these repositories' professional documents, and how their integration into national and international repository networks is reflected.

The paper is structured in three main sections: first, the research methodology is explained; second, the main findings are summarised, following the three levels defined in the methodology, local, national, and international; and third, a summary of the main conclusions is presented.

2. Research methodology

The researchers present an exploratory and descriptive study on the above-mentioned repositories' professional documents. The study has been carried out in four steps: 1. definition of a convenience sample; 2. development of a data codebook; 3. coding and comparison of data gathered; and 4. analysis of the data extracted and drawing of conclusions.

- At the local level, the following repositories are chosen for study: Repositório Institucional UNESP (University Estadual Paulista, https://repositorio.unesp.br) from Brazil, Estudo Geral (University of Coimbra, https://estudogeral.sib.uc.pt) from Portugal, and Repositorio Zaguan (University of Zaragoza, https://zaguan.unizar.es) from Spain.
- At the national level, the aggregators and networks from Brazil, Portugal, and Spain are analysed: the Brazilian Open Access Publications and Scientific Data Portal (Oasisbr, https://oasisbr.ibict.br), the Repositórios Científicos de Acesso Aberto de Portugal (RCAAP,

https://www.rcaap.pt), and the Recolector de Ciencia Abierta (RECOLECTA, https://recolecta.fecyt.es).

- Finally, at the international level, the following aggregators and networks are analysed: the Red de Repositorios de Acceso Abierto a la Ciencia (LAReferencia, http://www.lareferencia.info), the OpenAIRE (https://www.openaire.eu), and the Confederation of Open Access Repositories (COAR, https://www.coar-repositories.org). La Referencia and OpenAIRE also serve as aggregators, but this is not so in the case of COAR.

Of these three levels of analysis, the most intense attention has been paid to the local level, the institutional repositories.

Figure 1. Research methodology (source: own elaboration)

In the second step of the methodology, a codebook has been defined for information gathering, such as: institutional rules and procedure manuals openly available; authorship and traceability; institutional organisation and organic responsibility; thematic organisation of the contents; recommendations on subject metadata elements; recommendations to use controlled vocabularies (thesaurus, taxonomies, classification systems). The codebook includes aspects that pay attention to the open availability and regular updating of the institutional rules, procedure manuals, recommendations, and other supporting documents.

The codebook covers a total of 24 aspects, organised around three axes (see Appendix 1):

- the first axis covers descriptive, organisational, and regulatory aspects of each institutional repository;
- the second axis covers issues related to collection organisation and a repository's management, including the user support in the case of self-deposit; and
- the third axis focuses on issues related to the description of repositories' collections in terms of metadata and thematic content vocabularies, which are essential for knowledge organisation.

In the second level of analysis, national repository aggregators are studied (Oasisbr, RCAAP and Recolecta), observing the following aspects: alignment with international policies and guidelines

for metadata collection and validation; the documents, guidelines and recommendations offered to the repositories in the network; the effective application of knowledge organisation methods for content browsing and retrieval.

Finally, at the third level of analysis of this second step, the international level, special attention has been given to the published guidelines, recommendations of good practices, and other frameworks.

In the third step of the methodology, all the data obtained from the observation of institutional repositories, aggregators, and national and international networks, and their public documents analysis, are codified and compared.

In the fourth step, all the results obtained have been compared and analysed in detail, in order to draw precise conclusions about what has been observed regarding the organisation of knowledge in the repositories.

3. Findings

In the following three sections will be presented the main results obtained in the analysis carried out at three levels, local, national, and international, for each of the repositories, networks, and aggregators included in the analysis.

3.1. National level: UNESP, Estudo Geral, and Zaguan repositories

In this section will be discussed the most critical aspects of the rubric, observed in each of the sample repositories at the local level. The designed codebook included a total of 24 elements of analysis, and, in the following lines, those aspects related to the organisation of knowledge in institutional repositories will be discussed. Some of these aspects were expanded and analysed in greater depth, after consulting the policies and best practice guidelines of the networks and aggregators; namely, Recolecta, OpenAIre, and COAR.

3.1.1 UNESP Institutional Repository (Brasil)

The UNESP Institutional Repository aims to store, preserve, disseminate, and provide open access to scientific, academic, artistic, and technical documentation, as well as data and management plans, produced by researchers and students at UNESP.

The repository was established in 2013, using the open source software DSpace for the management of digital collections. It develops its collections through: databases' harvesting processes (Web of Science, Scopus, SciELO, PubMed, Latess Platform, etc.); automatic import of indexed records of the university's researchers production (Digital Library of Theses and Dissertations and the Digital Library of Final Degree Works); and on-demand archiving and self-archiving (Assumpção *et al.*, 2014; Vidotti *et al.*, 2015, 2016).

Currently, the UNESP Repository stores more than 200,000 records organised in seven communities of output types: academic and scientific, administrative, artistic, cultural, technical, and commemorative production. These communities are organised by collections representing the university units, and divided into sub-collections of document types (articles, theses, dissertations, etc.), departments, and student graduate and postgraduate programs.

The UNESP Repository uses natural language (keywords) combined with indexing language for subject representation. It does not use automatic tools for validation or correction of terms or subject names (geographical, persons, identifiers, series, and titles). Nor is there any vocabulary checking of keywords collected from the metadata.

The repository does not perform authority control of records migrated from external sources to the repository, nor does it have a written and formalised indexing policy. Tutorials are provided for the assignment of metadata in the self-archiving process, and the use of the UNESP Thesaurus is recommended to establish the thematic descriptors. The repository also offers multiple channels to launch support queries, via online chat, or forms. The online chat is available for the whole repository.

Keywords are a digital object retrieval option within the repository search menu. It is an alphabetical list that currently counts on more than 360,000 terms in Portuguese, English, French, and Italian. The list presents variations of the same word in singular and plural, in upper and lower case, and so on, as a result of the above-mentioned lack of authority control.

3.1.2 Estudo Geral (University of Coimbra, Portugal)

As for the second example, Estudo Geral (General Study) is the digital repository of scientific production of the University of Coimbra (UC), whose library has the responsibility and function of ensuring the quality of the metadata. This repository was publicly presented in June 2008, after the UC subscribed to the principles of the Declaration of Berlin, in 2007.

Estudo Geral was created with the objective of gathering, storing, preserving, disseminating, and providing access to the UC's scientific production, increasing its visibility and that of its researchers (Miguéis, 2021). For the implementation and development of the repository, DSpace software was adopted. In 2018, it began to operate on a DSpace-CRIS basis, which allowed strengthening the relationship between the General Study and the UC research centres, introducing new functionalities and contributing, in this way, to a scientific information management system (Ferreira *et al.*, 2021; Miguéis and Neves, 2021).

Estudo Geral reflects the organisational structure of the UC. It is organised in communities (faculties and research units), which are subdivided into subcommunities (department, when they exist), where the collections are gathered, created from the different documentary typologies resulting from research and teaching activities. The most frequently deposited document typologies are master's theses, doctoral theses, book chapters, scientific articles, and conference papers (https://estudogeral.sib.uc.pt/cris/explore/publications).

In the first phase, the archiving of scientific production was carried out solely by SIBUC (Integrated Service of Libraries of the University of Coimbra). In a second phase, and from the opening of the project to the academic community, and with the organisation of the scientific communities by areas of knowledge, the conditions were created so that the authors themselves could self-deposit their documents.

Issues concerning copyright have been safeguarded. The authors from the University of Coimbra are encouraged to grant authorization – not exclusive – of displaying digital documents in Estudo Geral. With this grant of non-exclusive authorization, the authors maintain their rights.

Regarding recommendations on the use of vocabularies and classification systems, on the Help page, the General Survey includes information from the DSpace templates on the use of subject categories for content description. It is noted that just the templates from the generic DSpace installation are used, and that this vocabulary has not been implemented to describe the collection. The resources collection is actually described by keywords assigned by the authors by self-deposit or by the repository working team.

The repository's collections of materials include 55,688 items, while the number of keywords reaches 142,927 terms, resulting in an average of 2.56 terms per document. An analysis of the list of keywords used shows that there is a general use of uncontrolled terms, with a variety of expressions, a mixture of capital and lowercase letters, terms in English and Portuguese, a mixture of numbers and signs, erroneous terms from search sentences, which seem to be of little use for launching new search engines.

3.1.3. Zaguan, University of Zaragoza (Spain)

Zaguan is the institutional repository that collects digital objects produced by professors, researchers, and students at the University of Zaragoza. It is the responsibility of the university library – as all the information is inserted into a web page of the library dedicated to open access – and of the Computing and Communications Service of the library, as indicated in the footer.

The only official normative document found has been the *Open Access to Research Results Policy*, approved in 2015 by the university's governing council. It is a declarative and normative

document. Its purpose is to set out the institutional interpretation of open access policies and to establish the commitments of the institution and its researchers.

In the repository policy, which is basically a web page (and that causes difficulties with the traceability of authorship and updates), no further responsibilities for the repository are explicitly determined. Nevertheless, these responsibilities can be inferred since the repository is based on delegated deposit, not self-deposit, and therefore other units of the organisation, and other systems, are somehow involved in the process, and managing the deposit of that part of the collection. At least five responsible parties are involved: secretariats of the centres (theses), vice-rectorate for Science Policy and the Scientific Information Management Unit of the University of Zaragoza Library (scientific publications), doctoral school (dissertations), and the university library itself.

The general secretariat of the university is involved, since the repository contains the official gazettes and a good number of regulatory documents of the institution. And also the library itself, with its historical collections, as well as the Publications Service of the University of Zaragoza.

Zaguan is a system with some important inconsistencies, such as the juxtaposition of administrative units; the user guides repeated in different places; the lack of updates and the lack of systematisation in the organisation of the collections. The repository policies do not state any issues related to the organisation of knowledge. Metadata is mentioned, with a section, but only from the point of view of the use and licensing of metadata.

The repository itself, implemented with the open source software Invenio, contains a list of thematic categories non systematically organised, and in a random sequence of "collections" (academic papers, theses, academic materials, *Official Bulletin of the University of Zaragoza*, rules of the University of Zaragoza, articles, communications and papers, preprints, books, journals, historical collection, personal collection, open data).

However, the search box responds to library logic and allows searching, as in bibliographic records in any field, title, author, abstract, keyword, report number, journal, year, full text, and reference. There are also Search Tips (https://zaguan.unizar.es/help/search-tips?ln=es), with indications similar to search instructions for online databases; instructions and requirements for self-archiving and a help page, with Search Tips and Search Guide, Complete Guide, Submission Help, and Metrics.

Zaguan does not have tutorials, user manuals, or public procedure manuals. Its pages lack authorship, it is not possible to know the traceability of its versions or the dates of updates. It needs to improve the categories of the collections, establish the metadata used in the repository, and adopt knowledge organisation systems for thematic description.

3.2. National aggregators: Oasisbr, RCAAP, Recolecta

At the second level of analysis, the national level, the study has shown that the three national aggregators (Oasisbr, RCAAP, Recolecta) share a common framework of metadata aggregation based on international policies and guidelines.

3.2.1. Oasisbr, Brasil

The Brazilian network and national aggregator, Oasisbr, aggregates the national scientific production in open access into a single portal, harvesting 1,499 sources, such as: journals (1,293), repositories of publications (117), digital libraries of theses and dissertations (67), digital libraries of monographs (4), together with research data repositories and other collections.

The contents aggregated by Oasisbr are also collected by the Open Access Scientific Repository of Portugal (RCAAP), and, in the same way, the contents made available by RCAAP are collected by Oasisbr and made available to the Brazilian scientific community. The contents aggregated by Oasisbr are also collected by the LA Referencia network, which in turn is collected by the European Aggregator OpenAIRE.

Oasisbr requires the repositories use some basic metadata elements in order to be harvested, which includes the use of abstract and keywords. Moreover, Oasisbr offers a number of indicators

(https://oasisbr.ibict.br/vufind/indicators/home), including keywords and areas of knowledge CNPq, the nationally recognized classification of the Brazilian National Council for Scientific and Technological Development (Conselho Nacional de Desenvolvimento Científico e Tecnológico), a body of the Brazilian Ministry of Science, Technology and Innovation. This classification of eight knowledge areas and four levels of hierarchy, is widely implemented in Brazilian repositories.

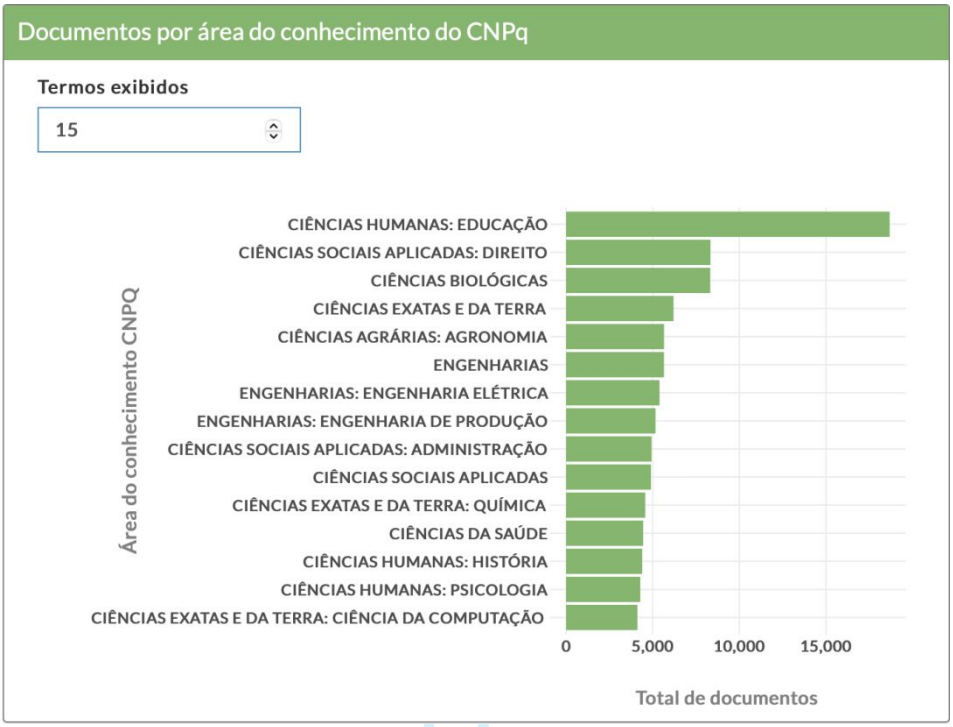

Figure 2. CNPq areas of knowledge (source: https://oasisbr.ibict.br/vufind/indicators/home)

The search form of the Oasisbr aggregator includes as a main query element the title, author, and topic of the document. The advanced search adds more elements including the CNPq area of knowledge. At the results page, it is possible to view the metadata of the resource, which includes keywords in Portuguese and English. It is not possible either to identify the use of other controlled vocabularies to standardise the content of these fields.

3.2.2. RCAAP, Portugal

The RCAAP portal is an aggregator (meta-repository) that collects the description (metadata) of documents deposited in various institutional repositories (53), common repositories (89), journals (246), and other sources (29) in Portugal. The portal saves the full text of these documents to improve the search results but does not save any document to offer to the users, just links to the original resource.

The RCAAP portal is the main service of the project Repositórios Científicos de Acesso Aberto de Portugal, an initiative from the UMIC Knowledge Society Agency, a public institute existing from 2005 to 2012. The RCAAP portal was developed by FCCN Fundação para a Computação Científica Nacional (Portuguese Foundation for National Scientific Computing), which is responsible for its maintenance and execution with the technical and scientific support of Minho University.

FCCN offers several services besides the RCAAP portal, as an institutional repository hosting service (SARI); a common repository; and a repository validation tool, the RCAAP validator. The RCAAP validator explicitly says in its Features page, the validation types followed are: https://validador.rcaap.pt/validator2/features#validation-types. Several validation profiles are available, together with an indication of the purpose of this validation. The basic profile follows the

DRIVER 2.0. guidelines and is used for publication in the RCAAP portal. Also, a validation profile based on OpenAIRE guidelines is applied for European commission funded publications.

The Portuguese aggregator RCAAP, as we have already explained, feeds back to the Brazilian network Oasisbr. Consulting its references to the guidelines, they link to those of OpenAIRE, making it clear that this network is also integrated in OpenAIRE.

Analysing the portal, RCAAP offers an advanced search with four options of querying the harvested content (title, subject, author, and description). When performing a search through subject, the results page offers a list of keywords with a number of occurrences, to limit the search. The details page of each result includes the list of assigned subjects.

The FCCN offers services in several technological areas, and as part of its Knowledge area, they include the RCAAP portal and also Indexar. INDEXAR https://www.indexar.pt is a Directory of repositories and digital journals in the areas of science, technology, and culture. The directory allows to browse the list of repositories and journals by several indexes, including Subject.

In the Support section of RCAAP, several professional documents, guidelines, tutorials and other tools are provided. The requirements for resources aggregation to RCAAP stated that they are based on the DRIVER Guidelines, though no specific comment regarding subject metadata is included.

A quick look through the DRIVER 2.0 Guidelines the project translated into Portuguese in 2009, and which are also linked from the RCAAP Support page at https://projeto.rcaap.pt/apoio/geral/diretrizes-driver-2-0. The DRIVER 2.0. guidelines include a specific section about the "Use of Vocabularies and Semantics". It recommended to deliver information on the classification usage in a repository in the Identify response, and to transport the classification in the element subject "URI-field" using an authoritative namespace.

The DRIVER guidelines did not prescribe any controlled vocabulary, just mentioned that the most frequently used classification schemes were the Library of Congress Classification, the Dewey Decimal Classification (DDC), and the Universal Decimal Classification. Also, the most frequently used subject headings systems in OAI context were the Library of Congress Subject Headings (LCSH) and Schlagwortnormdatei (SWD). If no specific classification scheme was used, DRIVER recommended the Dewey Decimal Classification.

3.2.3. Recolecta, Spain

The Spanish network and harvester, Recolecta, adheres to international policy frameworks for metadata harvesting and validation. Recolecta follows the principles of international initiatives, as part of the OpenAire project network. Moreover, the alliances with other networks are explicit; for example, OpenAire, COAR, LA Referencia.

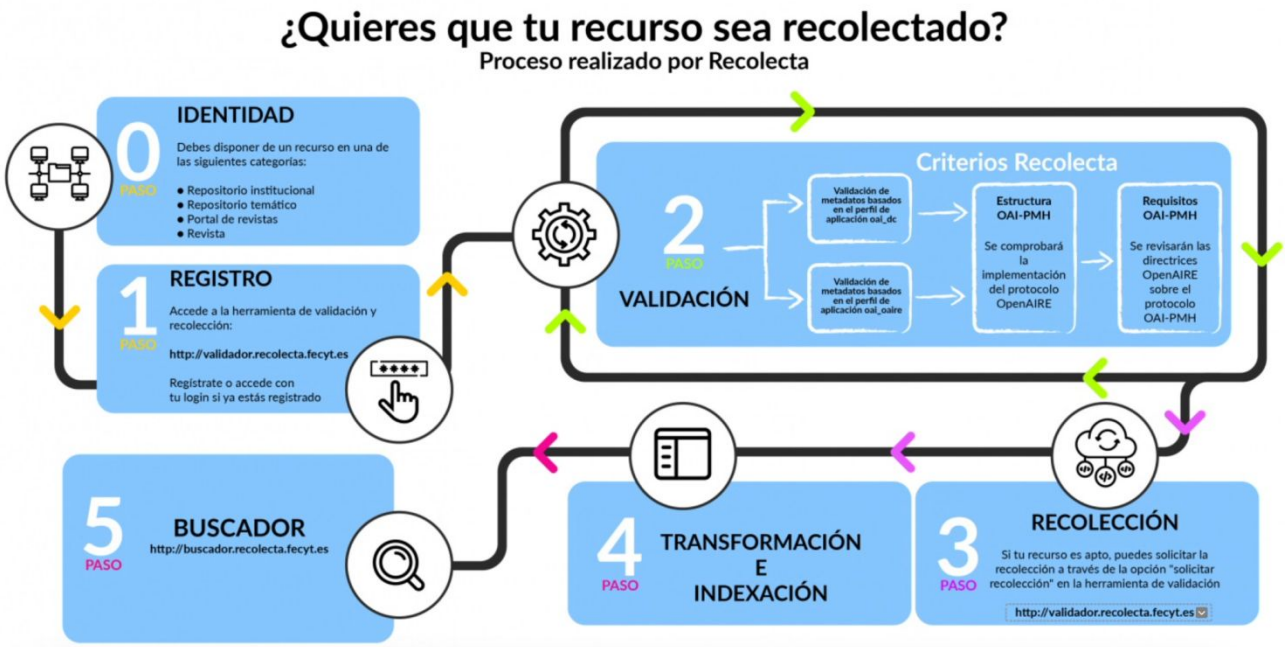

Figure 3. Recolecta harvesting and validator workflow (source: https://www.recolecta.fecyt.es)

If we analyse the selection of documents offered by Recolecta, the most significant document for the present study is the "Guide of research repositories assessment", whose second edition was published in 2021 (Grupo de trabajo de evaluación de repositorios, 2021). In this guide, Recolecta includes three elements that mention knowledge organisation questions: if there is an indexing policy known by authors, if a standardised classification system is applied, and if the repository uses controlled vocabularies or ontologies whose concepts are endowed with persistent identifiers. Regarding the indexing policies, they remark on the existence of a document stating this policy and their requirements, such as the use of controlled vocabularies if the repository applies a controlled indexing model. Regarding the use of standardised classification systems, Recolecta recommends the use of one or more standardised classification systems, such as CDU, JEL, UNESCO, and so forth. The importance of these systems is emphasised since they are of great help in performing selective collections by aggregators and can greatly facilitate the creation of value-added services.

The other document that includes a mention to the organisation of knowledge is the *COAR Community Framework for Good Practices in Repositories*. Although the document listed in their webpage is version 1 (2020), Recolecta also accepted version 2, dated July 2022. Version 1 of the framework includes as an essential characteristic in the objective of Discoverability, the fact that the repository supports quality metadata and controlled vocabularies. However, as it has been stated above, version 2 of the framework has reworded this first characteristic excluding the mention to controlled vocabularies, which has been moved to the column of desired characteristics.

As an aggregator, the search of contents from the 111 institutional repositories is very limited. The simple query just offers the option to introduce terms and search publications or projects, by their titles and authors. The advanced search just includes the option of querying by author's ORCID, name of repository, and DOI or Handle, and limit to or exclude from the query the publications, research data, projects, and open access products. At the results page, the filters by repository, author, language, document type, and year are available. No subject metadata is used for simple or advanced searches, result filters, nor is it shown in the metadata of harvested records. Though the harvester demands the repositories to be compliant with the OAI-PMH protocol, using the application profiles oai_dc and oai_oaire, they do not offer the option to visualise the harvested metadata records through their system.

3.3. International networks and aggregators: LA Referencia, OpenAire, and COAR

In this section are reviewed the approach of international networks and aggregators in terms of knowledge organisation recommendations, especially those on subject metadata coding.

Starting with the LA Referencia network, in 2015 it published its own policies through its document *Metadata and Harvesting Policies* (LA Referencia, 2015). They establish a series of interoperability guidelines, whose compliance must be guaranteed by the national nodes while recommending their adoption by the repositories that make up the network. The guidelines, agreed at the regional level, are based on DRIVER 2.0. and the *OpenAIRE Guidelines for Literature Repository Managers 3*, adopted by the European Union. Their compliance or non-compliance determines whether a record is accepted or rejected by LA Referencia in the collection phase.

The LA Referencia technical team collaborated on the new version of the Guidelines 4.0, published in November 2018, and approved by the LA Referencia Board of Directors at the end of 2018. Their implementation will be gradual starting in 2020, though it is not explicitly stated when the process of implementation is to be completed.

Following the abovementioned guidelines, LA Referencia policies include a section for the dc:subject metadata element. Regarding coding schemes, the policies only indicate that for controlled terms, thesauri, subject heading lists, and so on, will be used, and for classification numbers: CDD, CDU, LCC, and so forth. The policies even include a comment declaring that in the case of LA Referencia, there will not be any rule in this regard, as a national decision. Besides that, LA Referencia does not make any restriction or more specific recommendation.

Continuing the line of recommendations of international aggregators and networks, we analysed the guidelines established by OpenAIRE, which, as we have already seen, are followed nationally and internationally. OpenAIRE 4.0 (OpenAIRE, 2022b), includes a more detailed section dedicated to the dc:subject metadata element. Starting by mentioning that this element is used for subject, keyword, classification code, or key phrase describing the resource. The guidelines continue giving specific technical advice for encoding terms using additional attributes of the subject property, when terms are taken from a standard classification schema. It recommends using URI for terms taken from classification schemes, such as DDC or UDC. Moreover, if no specific classification scheme is used, OpenAIRE guidelines recommend the Dewey Decimal Classification (DDC).

The OpenAIRE Explore portal is a comprehensive and open dataset of research information that offers two options for subject searching: by vocabulary, and by Fields of Science. For subject search, at least the following subject vocabularies are used: ACM Computing Classification System and Microsoft Academic Graph classification. The Microsoft Academic project was active from 2015 to the end of 2021, when it was retired.

OpenAIRE Explore offers the indexing result of the OpenAIRE Graph production workflow (OpenAIRE, 2023). OpenAIRE Graph also performs document classification with an algorithm that employs analysis of free text stemming from the abstracts of the publications. The algorithm classifies the publication's full texts using a Bayesian classifier and weighted terms according to an offline training phase that used the following taxonomies: arXiv, MeSH (Medical Subject Headings), ACM, and DDC (Dewey Decimal Classification, or Dewey Decimal System). Moreover, the records included in the Graph are enriched with subjects based on the Microsoft Academic Graph (MAG) FieldsOfStudy.

The second option that applies knowledge organisation methods is the integration of a Field-of-Science (FoS) taxonomy into the OpenAIRE dataset, aiming to organise and discover research more effectively. Selecting a field, subfield, or topic from the taxonomy the system takes the user directly to the results page applying these criteria. It makes use of the full capabilities of the OpenAIRE Research Graph (full-texts, citations, references, venues) applying AI and bringing forward multidisciplinary potential.

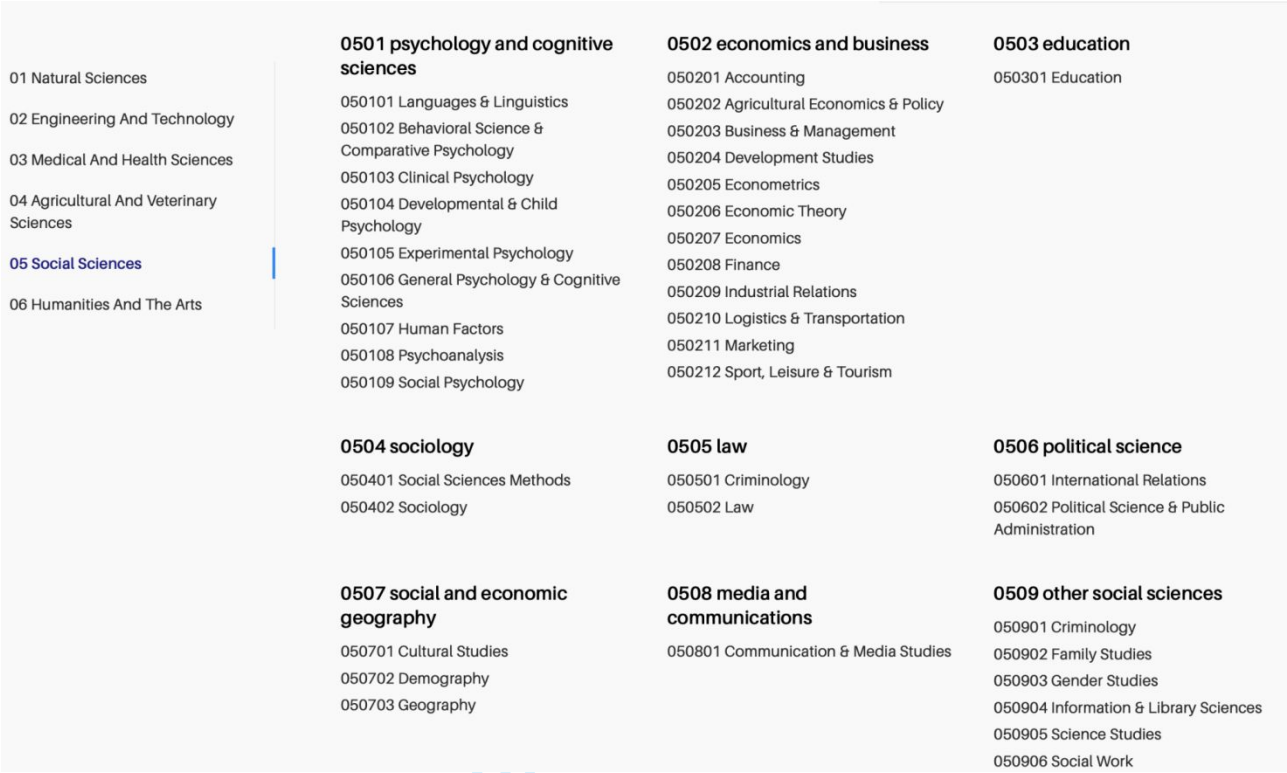

Figure 4. OpenAIRE Field-of-Science (FoS) taxonomy. Field 05, Social Sciences (source: https://explore.openaire.eu/fields-of-science)

FoS taxonomy (https://explore.openaire.eu/fields-of-science) is organised in six areas, and has three levels of hierarchical depth. It is based on previous work of the OpenAIRE partner Athena Research Center, named SciNoBo, a Hierarchical Multi-Label Classifier of Scientific Publications (Gialitsis *et al*., 2022). For the development of the taxonomy of knowledge domains, the authors relied on the OECD/FORD, Fields Of Research and Development classification, following the Frascati manual (OECD, 2015), and completed it by manually linking labels to Fields of Knowledge from the Science-Metrix classifications scheme (https://science-metrix.com/classification/). Thus, the classification is extended from two to three hierarchical levels. The network developed by SciNoBO was populated by exploiting Crossref and Microsoft Academic Graph (MAG).

In the case of COAR, the attention paid to metadata and vocabularies stands out, with a Controlled Vocabularies Editorial Board, which has established the COAR Metadata Working group (https://www.coar-repositories.org/news-updates/what-we-do/controlled-vocabularies). The group highlights the importance of standardisation and unification of metadata and controlled vocabularies to achieve a unified body of scholarly materials. These initiatives were defined in the roadmap for the internationalisation of metadata guidelines and vocabularies, together with other organisations (Shearer *et al*., 2019).

Their website includes a FAQ section, which specifies the benefits of controlled vocabularies, and stating that the controlled vocabularies more relevant to repositories are: subject heading lists, authority files, taxonomies, alphanumeric classification schemes, and ontologies.

The COAR Metadata Working Group has developed three controlled vocabularies, for resource type, access rights, and version type (https://vocabularies.coar-repositories.org). The thematic description has not been taken into account when making recommendations on the controlled vocabularies to be used. One of the issues highlighted on their website is a set of user stories and the metadata requirements they set out. For example, the first story focuses on discovery of content, and the metadata requirements include abstract and domain subject headings or keywords.

COAR also has a framework of good practices in repositories, grouped into several facets of their operations, such as discovery, access, reuse, integrity, quality assurance, preservation, privacy,

and sustainability. The aim is to offer a set of criteria to repositories to evaluate and improve their current operations based on a set of applicable and achievable good practices. In any case, COAR limits itself to indicating how to add terms from controlled vocabularies in the metadata records. They require the inclusion of the concept-URI when referring to a concept from the controlled vocabulary, and optionally one or more labels associated with the concept.

In the second version of the COAR Community Framework, we can find relevant criteria for discovery related to knowledge organisation principles. The first essential criteria points to the support of basic Dublin Core metadata to its records, as well as more granular elements, like discipline-based metadata. As for desired characteristics, the framework highlights the fact that the repository facilitates the use of controlled vocabularies in its metadata records. The mention of controlled vocabularies was included as the first essential criterion in the first version of the framework (COAR, 2020), and not in the desired ones as in its second version (COAR, 2022).

COAR Community Framework for Good Practices in Repositories

Version 2 – updated on August 29, 2022

Essential Characteristic	Desired Characteristic
1. Discovery	
1.1 The repository enables users to apply basic Dublin Core metadata to its records, as well as more granular elements (e.g. to support multilingualism, FAIR-compliance, discipline-based, and regional metadata schemas)	**1.9** The repository facilitates linking in the metadata record between related contents such as preprints, published articles, data, and software (e.g. including PIDs for related resources held elsewhere)
1.2 The repository supports harvesting of metadata using OAI-PMH	**1.10** The repository supports PIDs for authors, funders, institutions, funding programmes and grants, and other relevant entities
1.3 In cases where the resource has been withdrawn, the repository provides a tombstone page and the metadata record remains publicly available	**1.11** The metadata in the repository are made available under a Creative Commons public domain dedication / waiver (CC0)
1.4 The repository assigns persistent identifiers (PIDs) that point to the landing page of the resource	**1.12** In the case of research data, the repository supports identifiers for data at multiple levels of granularity, where appropriate (for example, if there there is research using a subset of the full dataset and a citation of the data subset is needed)
1.5 The repository offers a search facility	**1.13** The repository facilitates the use of controlled vocabularies in its metadata records
1.6 The metadata in the repository are indexed by external academic discovery services and aggregators	**1.14** The metadata in the repository are available for download in a standard bibliographic format at no cost to the user
1.7 The repository is included in one or more disciplinary or general registry of repositories	.

Figure 5. Discovery essential and desired characteristics. COAR Community Framework for Good Practices in Repositories, version 2.0 (source: https://www.coar-repositories.org/coar-community-framework-for-good-practices-in-repositories/)

At the international level, it is observed that guidelines and policies that include recommendations on subject metadata encoding have been adopted, always recognising the autonomy of repositories to implement them according to their needs.

4. Conclusion

The study findings indicate that at the local level, the institutional repositories, the following barriers were raised: manuals and guides are difficult to locate and access, identify their authorship, and trace their update; an evident lack of tools for vocabulary control; a low level of adoption of content schemes for subject metadata, and a very high number of keywords without vocabulary control.

At the institutional repositories studied, the use of controlled vocabularies is not regulated, leaving the assignment of thematic metadata to the authors' discretion. It results in alphabetical listings of keywords and uncontrolled natural language terms, mixing thematic descriptors with other authorities of persons, institutions, or places.

National aggregators do not regulate these issues either and limit themselves to pointing to international guidelines and policies, which simply recommend the use of controlled vocabularies, using URIs to facilitate interoperability. These guidelines establish strategic lines to converge on in the future, but do not set time milestones. The only national aggregator that includes subject controlled terms is Oasisbr, using the areas of knowledge CNPq classification. This vocabulary is applied as a field of the query form, and to offer the indicator of the number of documents associated with each area of knowledge, as it also does the keywords tag cloud.

At the international level, it is observed that guidelines and policies of international networks and aggregators include recommendations on subject metadata encoding, always recognising the autonomy of repositories to implement them according to their needs.

Thus, LA Referencia, whose metadata collection policies were initially founded on the DRIVER 2.0. guidelines, is currently in the process of adapting to the OpenAIRE 4.0 (OpenAIRE 2022a). OpenAIRE guidelines (2022a) just suggest the use of classification schemes or controlled vocabularies and, if no specific classification scheme is used, they recommend the Dewey Decimal Classification (DDC). OpenAIRE maintains a list of supported subject classification schemes, to code the qualifier of dc:subject element (OpenAIRE, 2022b).

For its part, the COAR association has an editorial board on controlled vocabularies, which has developed vocabularies on resources types, access rights, and version types, but not for subject elements (COAR, 2021).

The results of our research could serve as a call to action for repositories and networks to pay more detailed attention to knowledge organisation techniques that could be applied, standardised, and harmonised, to create a more meaningful, quality, and interoperable corpus of scholarly publications. To support a search and discovery of contents by their topics and subjects treated in a more precise and integrative way. By not generalizing the use of controlled vocabularies using URI-based terms, the possibility of creating linked datasets, which would allow thematic searching in multiple academic repositories, is hindered.

Finally, further research is needed on the effective use of specific content schemes for the encoding of controlled vocabularies and classification schemes. For this purpose, it would be necessary to perform a thorough harvesting of the sample and extend it to cover a larger number of repositories.

References

Araújo, D.O. and Silva, M.B. (2021), "Repositório digital: a delimitação de um conceito por meio de mapa conceitual", *Informação@Profissões*, Vol. 10 No. 3, pp. 18-33, available at: https://www.uel.br/revistas/uel/index.php/infoprof/article/view/44878

Assumpção, F.S., Silva, R.E., Ferreira, J.A. and Bastos, F.M. (2014), "A conversão de registros na implantação de repositórios institucionais: o caso do Repositório Institucional UNESP", in *XVIII Seminário Nacional de Bibliotecas Universitárias*. *Anais do XVII Seminário Nacional*

de Bibliotecas Universitárias, UFMG, Belo Horizonte, available at: https://www.bu.ufmg.br/snbu2014/wp-content/uploads/trabalhos/403-2123.pdf

COAR (2020), *COAR Community Framework for Best Practices in Repositories* (Version 1), Confederation of Open Access Repositories, Zenodo, available at: https://doi.org/10.5281/zenodo.4110829

COAR (2021), *COAR Repository Toolkit*, Confederation of Open Access Repositories, available at: https://www.coar-repositories.org/news-updates/coar-repository-toolkit

COAR (2022), *COAR Community Framework for Best Practices in Repositories* (Version 2), Confederation of Open Access Repositories, available at: https://www.coar-repositories.org/coar-community-framework-for-good-practices-in-repositories/

Ferreira, B.B., Neves, B., Miguéis, A.E. and Borges, M.M. (2021), "Competências para a gestão de um repositório institucional: o caso do repositório institucional da Universidade de Coimbra", *Revista Eletrônica de Comunicação, Informação e Inovação em Saúde*, Vol. 15 No. 4, pp. 974-986, available at: https://www.reciis.icict.fiocruz.br/index.php/reciis/article/view/2272/2484

Fujita, M.S.L. and Tolare, J.B. (2019), "Vocabulários controlados na representação e recuperação da informação em repositórios brasileiros", *Informação and Informação*, Vol. 24 No. 2, pp. 93-125, available at: http://dx.doi.org/10.5433/1981-8920.2019v24n2p93

Fujita, M.S.L., Agustín-Lacruz, M.C., Tolare, J.B., Terra, A.L. and Bueno-de-la-Fuente, G. (2021), "A organização do conhecimento em repositórios institucionais: Uma análise da literatura recente publicada em periódicos de biblioteconomia e ciência da informação", in Silva, C.G.D., Revez, J. and Corujo, L. (Eds.), *Organização do Conhecimento no Horizonte 2030: Desenvolvimento Sustentável e Saúde*, pp. 703-716, available at: https://doi.org/10.51427/10451/50067

Fujita, M.S.L., Agustín-Lacruz, M.C., Tolare, J.B., Terra, A.L. and Bueno-de-la-Fuente, G. (2023), "Institutional repositories and knowledge organisation: A bibliographic study from library and information science", *Education for Information*, Vol. 39 No. 1, pp. 51-66, available at: https://content.iospress.com/articles/education-for-information/efi220015.

Gialitsis, N., Kotitsas, S. and Papageorgiou, H. (2022), "SciNoBo: A hierarchical multi-label classifier of scientific publications", in Laforest, F. and Trocy, R. (Eds.), *Companion Proceedings of the Web Conference (WWW '22)*, Association for Computing Machinery, New York, NY, pp. 800-809, available at: https://doi.org/10.1145/3487553.3524677

Gollub, K., Kamal, A.M. and Vekselius, J. (2021), "Knowledge organisation for digital humanities: An introduction", in Gollub, K. and Liu, Y-H. (Eds.), *Information and Knowledge Organisation in Digital Humanities*, Routledge, London, pp. 1-22, available at: http://dx.doi.org/10.4324/9781003131816-1

Grupo de Trabajo de Evaluación de Repositorios (2021), *Guía para la Evaluación de Repositorios Institucionales de Investigación* (4th ed.), Fundación Española para la Ciencia y la Tecnología, Madrid, available at: https://www.recolecta.fecyt.es/sites/default/files/documents/2022GuiaEvaluacionRecolecta.pdf

Hjørland, B. (2021), "Information retrieval and knowledge organization: A perspective from the philosophy of science", *Information*, Vol. 12 No. 3, p. 135, available at: https://doi.org/10.3390/info12030135

JISC (2022), *OpenDOAR: OpenDOAR Statistics. Growth of OpenDOAR*, available at: https://v2.sherpa.ac.uk/view/repository_visualisations/1.html

Kounoudes, A.D. and Zervas, M. (2011), "Best practices and policies in institutional repositories development: The Ktisis case", in *3rd International Conference on Qualitative and Quantitative Methods in Libraries, 24 - 27 May 2011 Athens Greece*. , available at: https://hdl.handle.net/20.500.14279/4837

LA Referencia (2015), *Metadatos y Políticas de Cosecha de LA Referencia*, available at: https://www.lareferencia.info/es/recursos/directrices-metadatos/4-metadatos-y-politicas-de-cosecha-interoperables-para-los-nodos-nacionales

Miguéis, A.E. (2021), "Repositório institucional académico da UC e políticas de acesso aberto", in Sequeiros, P., Carvalho, M.J. and Capinha, G. (Eds.), *Investigação e Escrita: Publicar Sem Perecer*, pp. 41-66, Imprensa da Universidade de Coimbra, Coimbra, available at: https://doi.org/10.14195/978-989-26-2156-2_2

Miguéis, A.E. and Neves, B. (2021), "A visão dos gestores de repositórios. O caso da Universidade de Coimbra", in Borges, M.M. and Sanz Casado, E. (Eds.), *Sob a lente da Ciência Aberta: Olhares de Portugal, Espanha e Brasil*, Imprensa da Universidade de Coimbra, Coimbra, pp. 273-294, available at: https://estudogeral.uc.pt/bitstream/10316/93276/1/184-Book%20Manuscript-679-1-10-20210202.pdf

OECD (2015), *Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development*, OECD Publishing, Paris, available at: https://doi.org/10.1787/9789264239012-en

OpenAIRE (2022a), *OpenAIRE Guidelines for Literature Repositories v. 4.0*, available at: https://openaire-guidelines-for-literature-repository-managers.readthedocs.io/en/v4.0.0/

OpenAIRE (2022b), *OpenAIRE API. Vocabularies: dnet:subject_classification_typologies*:, available at: http://api.openaire.eu/vocabularies/dnet:subject_classification_typologies

OpenAIRE (2023), *Graph Production Workflow: OpenAIRE Graph Documentation* (Version: 5.1.2.), available at: https://graph.openaire.eu/docs/graph-production-workflow/

Phillips, M.E., Tarver, H. and Zavalina, O.L. (2020), "Using metadata record graphs to understand controlled vocabulary and keyword usage for subject representation in the UNT theses and dissertations collection", *Cadernos BAD*, Vol. 1, pp. 61-76, available at: https://doi.org/10.48798/cadernosbad.2024

Shearer, K., Holt, I. and Walk, P. (2019), "*D5.2 – Roadmap for Internationalisation of Metadata Guidelines and Vocabularies*", OpenAIRE Advance: WP5 - Global OA Scholarly Communication Infrastructure, available at: https://www.coar-repositories.org/files/Metadata-Roadmap.pdf

Sousa, B.A.D. (2012), "Proposta de Criação de um Repositório Institucional para o IFPB", *Revista Brasileira de Biblioteconomia e Documentação*, Vol. 8 No. 1, pp. 66-84, available at: http://rbbd.febab.org.br/rbbd/article/view/196/228

Terra, A.L., Agustín-Lacruz, C., Bernardes, O., Fujita, M.S.L. and Bueno-de-la-Fuente, G. (2021), "Subject-access metadata on ETD supplied by authors: A case study about keywords, titles and abstracts in a Brazilian academic repository", *Journal of Academic Librarianship*, Vol. 47 No. 1, available at: https://doi.org/10.1016/j.acalib.2020.102268

Vidotti, S.A.B.G., Bastos, F.M., Ferreira, J.B., Grisoto, A.P., Assumpção, F.S., Silva, R.E., Rodrigues, V.S. and May, O.L. (2015), "Reutilização de Metadados para o Povoamento de um Repositório Institucional: Procedimentos Aplicados no Repositório Institucional UNESP", in *Proceedings of the International Conference on Dublin Core and Metadata Applications*, pp. 234-235, available at: https://dcevents.dublincore.org/IntConf/dc-2015/paper/view/379/396.html

Vidotti, S.A.B.G., Bastos, F.M., Grisoto, A.P., Arakaki, F.A. and Ferreira, J.B. (2016), "Coleta automática para povoamento de repositórios digitais: conversão de registros utilizando XSLT", *Tendências da Pesquisa Brasileira em Ciência da Informação,* Vol. 9 No. 2, pp. 1-21, available at: https://revistas.ancib.org/index.php/tpbci/article/view/390/390

Villalobos, A.P. de O. and Gomes, F.A. (2018), "Análise dos repositórios das universidades federais brasileiras", *PontodeAcesso*, Vol. 12 No. 3, pp. 126-144, available at: https://periodicos.ufba.br/index.php/revistaici/article/view/27929

Appendix 1: Codebook

I. Description

a. Repository name
b. URL
c. Regulatory framework
d. Repository software
e. Institutional organization and organic responsibility
f. Management units
g. Number of records
h. Institutional rules openly available
i. Update data (Institutional rules)

II. Collection management

j. Deposit (of research data) mode (self-archiving, delegated deposit, etc.)
k. Assistance instruments: FAQ; tutorials; procedure manuals; support mail or chat
l. Assistance instruments regularly updated
m. Authorship of instruments (clearly stated)
n. Digital preservation policies
o. Use of persistent identifiers (DOI, Handle, URN, ORCID, etc.)
p. Collections
q. Organizing systems

III. Metadata

r. Metadata curation (authority control)
s. Recommendations on subject metadata elements
t. Recommendations to use controlled vocabularies
u. Indexing policies known by authors
v. Use of standardized classification system (e.g. UDC, DDC)
w. use of controlled vocabularies or ontologies with concepts endowed with persistent identifiers (e.g. LCSH LD, UNESKOS)
x. Number of topic/subject
y. Percentage of subjects/records