Elements of Resource Representation in Institutional Repositories: a Bibliographic Review

Marisa De Giusti
Universida Nacional de La Plata (UNLP)
Comisión de Investigaciones Científicas de la provincia de Buenos Aires (CIC)

José Daniel Texier
Universidad Nacional Experimental del Táchira (UNET)

Abstract: This review focuses on identifying how the literature studies the existing problems in the Resource Representation (RR) of Institutional Repositories (IR). RR is a process of recording in a persistent manner a set of data (metadata) as a synthesis and replacement of the "real" object, to allow its identification, retrieval and dissemination. RR is defined by certain elements: resources, metadata schemata, storage and cataloging. On the other hand, IRs are based on functional processes according to the material that is deposited and the ISO 14721 standard: *ingest*, *storage*, *cataloging*, *indexing*, *search engine* and *browsing*. The results of this review show that identifying the problems found in these elements and functional processes is not a subject of study for the researchers, which leads to a vacant area in this field.

1. Introduction

In the last decade, there has been a growth [1] in institutional repositories (IR), which represent a source of digital information which is specialized, organized and accessible for users of diverse fields. IRs are computer systems which manage scientific and academic works for different institutions, without restriction and free of charge [2]. They are also in line with the ideals and aims of Open Access [3], [4], and help in rethinking the process of publishing scientific papers [5]. Likewise, digital libraries started their revolution in 1990 [2], and with the years they begun consolidating their presence in the scientific world, until they became intertwined with the concept and functionality of the IR. Therefore, in the context of this work, an Institutional Repository is a Digital Library, and a Digital Library is an Institutional Repository, due to the fact that both offer similar services and the use of each term depends on the context of use and, consequently, of the resources wanted for working [2].

According to the different reviewed works on the operation of an IR [6]-[11] and the recommendations of the ISO 14721 standard (also known as the OAIS model), every repository must follow these functional processes regarding the deposited materials: *ingest*, *storage*, *cataloging*, *indexing*, *search engine* and *browsing*. In respect to how the deposit works, the functional processes are: preservation and management. Therefore, Resource Representation (RR) in an IR is defined by the functional processes related to the stored material and to the process of registration in a persistent manner of a set of data acting as a synthesis and replacement of the "real" object, in order to allow users [7] to identify, retrieve and disseminate it. When we mention resources, we are referring to physical or digital objects which are described by listing a set of specific data (called metadata) that distinguish them from other objects [7].

The concept of metadata is not something new, as they were already in use before the arrival of Internet as a way to catalogue books and journals through a normalization of data to allow for organized retrieval. In Information Science, metadata are used to refer to available records of information resources [13]. In other words, metadata are data that stand in description of other data, that is, they are a form of structured information that describes, explains and/or locates an information resource in order to identify, retrieve, use, manage o preserve it in a more systematic and transparent manner. Several

models, schemata, formats and standards have been developed for the representation of metadata, which, although sharing syntax and an XML information structure, differ in respect to the information they describe.

Based on the previous explanation, four key elements are defined in the representation of resources in institutional repositories [7], [12], relevant to this bibliographic review:

- resource typology,
- metadata schemata,
- storage, and
- cataloging, represented by controlled vocabularies, thesauruses and abstract entities (i.e., elements with their own descriptive information) such as authors, institutions and journals.

Thus, the objective of this bibliographic review is to learn how scientific literature studies the problem of resource representation in institutional repositories as a whole. In other words, to find solutions which involve the six functional processes of an IR, (depending on the deposited material) and the four elements of a RR. With this aim, the review was organized as follows: the second section shows the methodology used in the review; the third section describes the results; the fourth section contains an analysis, a discussion on results and the contribution of the presented work; and finally, the fifth section consists of some conclusions.

2. Methodology

A systematic bibliographic review consists in the identification, assessment and interpretation of every possible relevant research in a rigorous manner in order to answer a question, a particular area of research or a phenomenon of interest [15]. As a basis for the development of this review, some guidelines were taken from medical literature [16] and from criteria defined by Kitchenham in 2004 [15]:

- ***Research question***: The question that guided this bibliographic review was determining how the literature approaches the subject of resource representation (elements) inside institutional repositories (functional processes) taking into consideration the recommendations of the ISO 14721 standard.
- ***Assessment of the search strategy***: Such assessment was organized following the general PICOC [18] guidelines, which analyze effectiveness from five perspectives:
  - Population: the representation of resources in the LIS dominion.
  - Intervention: the elements in a resource representation and the functional processes of institutional repositories.
  - Comparison: according to the recommendations of the ISO 14721 standard, related problems are analyzed and compared with the elements and functional processes.
  - Result: result types are not limited in searches according to given criteria, since all the information available in the dominion of the study was needed.
  - Context: no restriction was applied.
- ***Search strategy and criteria***: Two search groups were defined for the Scopus bibliographical database, according to the guidelines of the ISO 14721 standard. The first search group, called "Problems with the Elements" (PE), focused on existing problems in each of the four elements (including sub-elements) in a resource representation: resources, metadata, storage and cataloging, always in the context of the LIS (Library & Information Science) dominion. The second search group, known as "Problems with the Processes" (PP), was based on existing problems in the six functional processes of repositories and their relation with the elements in

the resource representation. The search criteria in Scopus for both groups (PE and PP) are the same: types of documents to take into consideration, *article* and *review*; the selection fields in the database are: title, abstract and keywords; and, thematic area of the articles, *"Computer Science"* and *"Social Science"*. There's no restriction regarding the year of publication, and all searches were done on the 5th November 2013. In the PE group, the 10 searches are based on each of the elements and sub-elements in the resource representation. The results were restricted according to the presence of the *"digital library"* string as a keyword in the journals. The following truncated descriptor terms were used:

1. (types* AND resource*) AND problem*
2. metadata AND problem*
3. storage AND problem*
4. catalog* AND problem*
5. "controlled vocabular*" AND problem*
6. thesaurus AND problem*
7. "abstract entiti*" AND problem*
8. author AND problem*
9. institution AND problem*
10. journal AND problem*

The second group (PP) consisted in 24 searches to identify existing problems among the six defined functional processes and the four elements in a resource representation in the LIS dominion. The following six truncated descriptor terms were the basis of the searches, and each one related to the four elements in the RR:

1. ingest*
2. storage
3. cataloging
4. indexing
5. "search engine"
6. browsing

- ***Data extraction and synthesis:*** The results from the 34 searches (10 from the PE group and 24 from the PP group), were exported from Scopus as CSV (*comma-separated value*) files, which were later imported into Google Refine [19], a tool that offers additional functions to spreadsheet managers such as LibreOffice Calc or Excel. These result files and the detailed process of the performed searches can be found in a GitHub [20] project.

3. Results

3.1. Problems with the Elements group (PE)

The following Table 1 shows the found articles which correspond to the four elements and six sub-elements in resource representation. The third column (*"Arts. with DL & IR terms"*) shows results with the presence of the terms *"Digital Library"* or *"Institutional Repository"*. The last column (*"Arts. restricted by CS and SocS areas"*) represents those articles found when the search was limited to the *"Computer Science"* and *"Social Science"* areas. This last column represents a first result according to the established criteria (248 articles), to select the corpus to be analyzed in the fourth section.

#	Elements - Terms which expose problems	Arts. with DL & IR terms	Arts. restricted by CS & SocS areas
1	types OR resources	507	130
2	metadata	221	37
3	storage	181	45
4	cataloging	69	28
5	"controlled vocabular*"	11	2

6	thesaurus	20	9
7	“abstract entities”	0	0
8	author	164	40
9	institution	100	32
10	journal	77	26
	TOTAL - combination	979	248

Table 1 - Problems in the elements in a RR

3.2. Problems with the Processes group (PP)

Table 2 shows the amount of found articles which identify the problems that are present in the IR according to the six functional processes of a repository in relation to the four elements in a resource representation. The last column (*"intersection"*) shows whether there is any article which deals with the problem in the four elements in a resource representation for a specific functional process in an IR.

Functional processes	Resources	Metadata	Storage	Cataloging	Intersection
*ingest**	1	3	3	3	0
storage	44	31	-	29	2
cataloging	7	6	3	6	0
indexing	20	23	23	24	0
“search engine”	35	22	4	24	0
browsing	17	9	8	6	0

Table 2 - Relation between processes and elements

The two articles that result form the intersection of the elements with the *storage* process are:

- Integrating chemistry scholarship with web architectures, grid computing and semantic web [21].
- Data for the future The German project "Co-operative development of a long-term digital information archive" [22].

The first one was presented in a lecture, which is the reason why in order to respect the search criteria for the PE group (only in journals and bibliographical reviews) it was discarded. On the other hand, the second article is already included in the PE group.

3.3. General considerations

Some general results can be drawn out taking into account the 248 articles obtained in the searches regarding: authors, year of publication, and journals.

- **Authors:** Table 3 shows 5 authors (from the 558 found in 248 articles) which have the highest number of publications, that is, between 6 and 3 articles each. We infer that this information reflects the existence of lines of research such as the 5S model by Gonçalves, Fox and Laender [23], the digital libraries in university education by Fox [24], [25], author deduplication by Gonçalves, Ferreira and Laender [26], and retrieval systems by Herrera-Viedma [27].

Authors	Amount of articles
Gonçalves Marcos A.	6

Laender Alberto H.F	4
Ferreira Anderson A.	3
Fox Edward A.	3
Herrera-Viedma Enrique	3

Table 3 - Authors most present in the results

- **Year of publication:** Figure 1 shows a timeline with the time and amount of publications reported regarding the topic. The years 2007 and 2012 stand out with 26 and 25 articles, as a clear evidence of how new this field of research is.

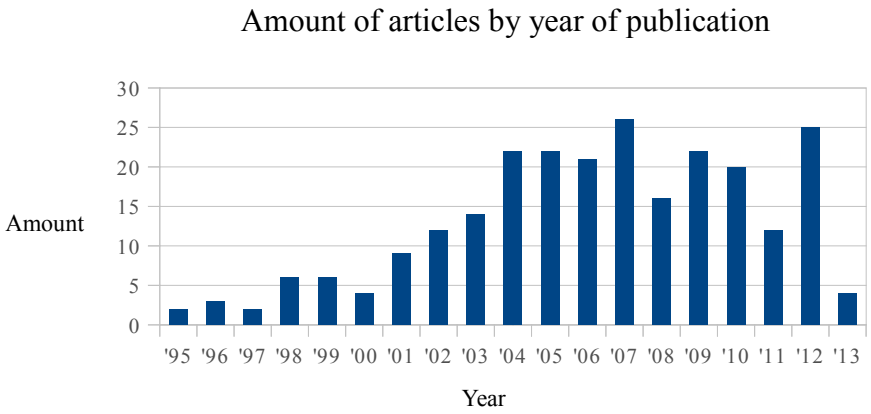

Figure 1 - Year of publication of the results

- **Journals:** Table 4 shows the first 13 journals of a total of 110, sorted in a descending order with the amount of published articles from the total 248 articles found.

Journals:	**Amount of articles**
Electronic Library	44
Program	17
OCLC Systems and Services	15
Library Hi Tech	14
Online Information Review	14
Journal of the American Society for Information Science and Technology	13
Science and Technology Libraries	11
Computer Physics Communications	9
International Journal on Digital Libraries	8
Aslib Proceedings: New Information Perspectives	7
IEEE Transactions on Knowledge and Data Engineering	6
Journal of Digital Information	6
Journal of Library Metadata	6

Table 4 - Journals most present in the results

3.4. Corpus

The aim of this work is to analyze the literature that was found and which exposes the different existing problems in resource representation in an institutional repository. For this goal, 248 publications were found which deal with at least one element in the RR. However, after calculating an intersection between elements and sub-elements, through the Google Refine project, an article was found in which the four main elements of the representation are present (see Table 5).

Groups	Amount of articles
A single element	194
Two elements	45
Three elements	8
Four elements	1
TOTAL	248

Table 5 - Relation between elements

Our interest resides in studying resource representation as a whole, therefore, in Table 5 a vacant area in the LIS dominion becomes evident, seeing that only one such publication is shown [22]. This way, Table 6 shows the final corpus (9 articles) chosen for our study from a selection of the publications focused in three and four elements.

#	Article or review	Element or sub-elements
1	Data for the future The German project "Co-operative development of a long-term digital information archive" [22]	4 elements: resources, metadata, storage, cataloging
2	Towards accessibility to digital cultural materials: An FRBRized approach [28]	3 elements: resources, metadata, cataloging
3	The growth of electronic journals in libraries: Access and management issues and solutions [29]	3 elements: resources, metadata, cataloging
4	BibPro: A citation parser based on sequence alignment [30]	3 elements: resources, metadata, cataloging
5	Digital library development: Identifying sources of content for developing countries with special reference to India [31]	3 elements: resources, metadata, cataloging
6	From digital library to institutional repository: A brief look at one library's path [32]	3 elements: metadata, storage, cataloging
7	Help features in digital libraries: Types, formats, presentation styles, and problems [33]	3 elements: resources, metadata, cataloging
8	Provision of digital preservation metadata: A role for ONIX? [34]	3 elements: resources, metadata, storage
9	Subject Access: Conceptual Models, Functional Requirements, and Empirical Data [35]	3 elements: resources, metadata, cataloging

Table 6 - Resulting corpus

4. Analysis and Discussion

The analyzed corpus consisted of 9 of the 248 articles found. This first discovery demonstrates a neglected study area in the LIS dominion, seeing that both in research articles and in bibliographic reviews the topic of resource representation is not dealt with inside institutional repositories as a whole

in which the six functional processes of an IR and the four elements of a RR come together.

Of the articles obtained, three were discarded for reasons that are explained for each case. The first of them was the work by Xie [33], due to the fact that it is not concerned with any problem related to the purposes of this review. The works by Chen *et al.* [30] and Buelhler *et al.* [32] were discarded because they focus on the integration of an institutional repository of article quotes and digital libraries respectively, coming from different established software platforms. These articles highlight among their problems and solutions some which are related to the elements in a RR and functional processes of IRs, but they not meet the criterion of studying at least 3 of the elements in a RR.

The 6 articles that were part of the final corpus were those which expose problems related to the purpose defined for this study, for example: diversity of technology solutions, treatment of diverse resource typologies, metadata schemata, recommendations in resource storage, resource preservation, OAIS model recommendations and the application of conceptual models to solve the problem of RR as a whole, such as FRBR (*Functional Requirements for Bibliographic Records*), FRAD (*Functional Requirements for Authority Data*) and FRSAD (*Functional Requirements for Subject Authority Data*) [22], [28], [29], [31], [34], [35].)

The three pillars of analysis that served to group the approaches in the publications that were found are centered in:

1. *Resources:* the work by Altenhöner [22] mentions that digital objects should be seen and treated as bitstreams, according to the ISO standard. On the other hand, the works of Weng et al. [28] and Mischo et al. [29] study the diversity of cultural and electronic resources, adapted to metadata schemata known as MARC or METS. Likewise, Jeevant [31] explains the treatment of digitized academic and scientific resources, and Brindley et al. [34] are focused in the migration of the book resource. By contrast, Zavalina studies the resource in general terms and adapted to the FRBR family of models [35]. In short, a great amount of resources are identified that should be part of an IR, such as: articles, reviews, proceedings, papers, theses, datasets, administrative documents, government documents, technical reports, etc. Thus, an IR has to adapt to existing typologies and to new types of resources that are yet to come in the future.
2. *Metadata schemata:* in the corpus there is evidence of several traditional schemata as solutions to IR. Among the most quoted we find the METS schema, as studied by Altenhöner [22] and Brindley et al. [34]; and MARC by Weng et al. [28] and Mischo et al. [29]. Additionally, authors such as Altenhöner [22], Mischo et al. [29], Jeevant [31] and Brindley et al. [34] use general purpose schemata of their own development, for example, Dublin Core. The work by Brindley et al. deserves a special mention [31] because it deals with the topic of digital preservation and recommends the use of the PREMIS schema. All of these schemata must be allowed by the IRs in order to avoid information loss and to enable interoperation with other IRs. Repositories must also be capable of adapting to metadata schemata that may arise in the future and are established by means of recommendation or imposition as a de facto standard.
3. *Storage:* the work Jeevant [31] recommends a process to manage the persistence of metadata and the digital object, and Brindley et al. [34] works directly with the relational database paradigm. It should be noted that the works by Altenhöner [22] and Brindley et al. [34] mention the recommendations of the ISO 14721 standard. Therefore, these works are focused in solving the persistence of information through the model and guaranteeing the retrieval of the information in a regular way (by means of queries), and in case unforeseen accidents happen (by the use of mirror or distributed copies). Moreover, they recommend the use of persistent resource identifiers and the performance of resource review and modification tasks in order to improve the integrity and quality of the information (given that deposits originate from diverse

sources and means) and the correct use of bibliographic controls.

4. *Cataloging:* the works that were found recommend different ways of guaranteeing the normalization of the information entered and stored inside an IR. Weng et al. [28] focus in the recommendation of the FRBR model to catalogue the different types of resources, Mischo et al. [29] suggest the strict use of bibliographic controls such as a analyzing documents both in form and content, and the work by Jeevant [31] recommends the use of a cataloging guide built on the basic principles of the discipline. Additionally, Zavalina [35] recommends the use of the models of the FRBR family (FRBR, FRAD and FRSAD) and the implementation of the RDA (*Resource Description and Access*) cataloging code, in order to avoid using the AACR2 (second edition of the *Anglo-American-Cataloguing* rules) or their predecessors, with Zavalina approaching the cataloging problem through an access by topic using two actors: users who search for information in the repository systems and information professionals who analyze and create resource metadata. In consequence, the goal of these works is to make the points of access easier so users can locate and retrieve the indexed information, which generally is centered in the title, author and subject fields, and in some cases allow for a full text search.
5. *Resource Incorporation:* based on the functional process of adding items into a repository, the work by Altenhöner [22] notes that deposits come from different sources and are achieved through different means, but he is not focused in any solution that is based on any standard, as do Brindley et al. [34], who recommend keeping the criteria of the ISO standard and the use of information packages, which in this case would be the SIPs. A similar situation can be seen in the work by Jeevant [31], which only explains the importance of the incorporation process but does not mention it as either a problem or an advantage, simply as one of the steps to follow to make resources available to a community.
6. *Data Management:* the works that were found on how to populate, maintain and access the information refer to the case of Weng et al. [28] in the application of the FRBR model and the works of Altenhöner [22] and Mischo et al. [29] in the use of cataloging guides for controlling information. Again, Zavalina [35] recommends data management to be focused on the FRBR model family. To sum up, these works are focused on being able to create points of access and normalizing the information (metadata) related to digital objects as an answer to the growth of repositories, migration processes, or simply its everyday use.
7. *Access:* it is of foremost importance to the user, as it represents the entry way to the repository. The works that were found study the problem of access from the perspective of functional processes *Indexing, Search* and *Browsing*. In the works by Weng et al. [28] and Zavalina [35] a solution to access through the application of the FRBR model is analyzed, this being a model which would help create a solid cataloging effort, allowing the proper indexing of resources and guaranteeing the right results for searches, and navigation through all the defined points of access. However, Zavalina incorporates the importance of topic management for a proper access and incorporates the FRAD and FRSAD models [35]. The works by Mischo et al. [29] and Jeevant [31] offer broad guidelines to web access, but do not bring any actual solutions to the problem.

The seven reported areas of discussion stem from the four elements in a RR and the four functional modules explicitly mentioned in the ISO model, according to the material deposited in the IR. The authors present their problems in these areas and report a diversity of solutions for each analysis which raise doubts in regards to the existence of solutions that can encompass the best in the proposals of each of the authors.

5. Conclusions

In the second section the applied methodology for this review was presented, as were the different searches in Scopus and the use of Google Refine, all of which can be found in a GitHub project [20], so that any researcher can replicate the results of this review. Therefore, this review led to the following conclusions:

- In the groups for the searches that were performed (PE and PP groups), a vacant area appears in the LIS dominion, more specifically in the resource representation in an institutional repository according to the works analyzed and the ISO 14721 standard. Those few articles with element relations (Table 5) and the resulting corpus (Table 6) help to focus on the different problems present in IRs as a whole, and to begin creating solutions in the same direction.
- Table 1 shows the existence of resource representation in a broad manner, but the topic is considerably reduced, a 25.33% (from 979 to 248 articles) if the RR problem is focused into the field of Computer Science, Information Science and Documentary Sciences. Thus, it follows that resource representation is not an exclusive topic to institutional repositories or digital libraries.
- Figure 1 makes apparent just how innovative and dynamic the topic of this review is. Though publications started in 1995, it shows a steady growing presence since 2001.
- Table 4 shows the journals with the higher amount of articles found, where those 13 journals from a total of 110 have 170 articles of the 248 found, which represents a 68.55%. It is considered to be a very high proportion that could draw enough attention to have a study made based on said results, and the resource representation area in institutional repositories.
- Finally, a study that relates the problem of resource representation with the diverse conceptual models of digital libraries and institutional repositories mentioned in these articles is recommended. These studies are as follows: formal model proposed by Gonçalves et al. [23], OAIS reference model [12], FRBR conceptual model [36], among others. Thus, repositories can be analized in detail in the light of a general model, which can extract the best of each of them on the basis of functional elements and processes studied in this revision.