# SOCIAL SCIENCES AND HUMANITIES RESEARCH DATA AND METADATA: A PERSPECTIVE FROM THEMATIC DATA REPOSITORIES

## Datos y metadatos de investigación en ciencias sociales y humanidades: una aproximación desde los repositorios temáticos de datos

Nancy-Diana Gómez, Eva Méndez and Tony Hernández-Pérez

**Nota**: Este artículo se puede leer en español en:
*http://www.elprofesionaldelainformacion.com/contenidos/2016/jul/04_esp.pdf*

**Nancy-Diana Gómez**, librarian and graduated in arts from the *University of Buenos Aires*, currently is a PhD student with the *Archives and libraries in the digital environment* program at the *Carlos III University of Madrid* (*UC3M*). She has taught in the *Department of Library and Information Science* at the *UC3M* (2009-2013) and at the *University of Buenos Aires*, where she was also director of the *Central Library* of the *Faculty of Natural Sciences* (1994-2005). She is co-coordinator of the *Latin American list on Open Access Repositories* (*Llaar*), and participates in national and international research projects.
*http://orcid.org/0000-0002-6218-6248*

*ndgomez@bib.uc3m.es*

**Eva Méndez** is an associate professor with the *Department of Library and Information Science* at the *Carlos III University of Madrid*, where she is currently vice provost for *Strategy and digital education*. Doctor of documentation, her teaching and research deal with metadata, semantic web, digital libraries, open access, information policies, and social web. She is a member of the *Dublin Core* (*DCMI*) advisory board. Since 2015 she has also belonged to the *OpenAIRE* advisory committee and the *Rebiun* executive committee. She has participated as an independent expert for the *European Commission* on digital libraries and open science.
*http://orcid.org/0000-0002-5337-4722*

*emendez@bib.uc3m.es*

Manuscript received on 18-04-2016
Approved on: 08-06-2016

**Tony Hernández-Pérez** holds a PhD in information science and is a professor with the *Department of Library and Information Science* at the *Carlos III University of Madrid* where he is the director of the doctoral program in documentation. His teaching and research are linked to the *TecnoDoc* group matters: social web, web content management, metadata, information retrieval, e-learning, and journalistic and audiovisual documentation.
*http://orcid.org/0000-0001-8404-9247*

*tony@bib.uc3m.es*

*Universidad Carlos III de Madrid*
*Facultad de Humanidades, Comunicación y Documentación*
C/ Madrid, 128. 28903 Getafe (Madrid), Spain

## Abstract

This paper studies research data repositories in the social sciences and humanities (SSH), from the *Registry of Research Data Repositories* (*re3data*), paying particular attention to metadata models used to describe the datasets included in them. 397 repositories are reviewed at the general level, including those of a multidisciplinary nature. We discuss and reflect on the special features of research data in these disciplines, and on coverage and information collected by *re3data*. The metadata schemas and standards most commonly used in SSH repositories are analyzed, with special emphasis on the six main repositories.

## Keywords

Repositories; Research data; Metadata; Social sciences; Humanities; *Re3data*.

## Resumen

Se estudian los repositorios de datos de investigación en ciencias sociales y humanidades (CSH), recogidos en el *Registro de repositorios de datos de investigación* (*re3data*), prestando especial atención a los modelos de metadatos que utilizan para describir los datasets incluidos en ellos. Se revisan a nivel global los 397 repositorios que, según *re3data*, recogen datos de investigación sobre esas disciplinas, incluidos, los de carácter multidisciplinar. Se discute y reflexiona sobre las particularidades de los datos de investigación en estas disciplinas y sobre la cobertura e información que recoge *re3data*. Se analizan los esquemas y estándares de metadatos más utilizados en los repositorios de CSH, con un análisis más pormenorizado de los seis repositorios de datos especializados más importantes.

## Palabras clave

Repositorios; Datos de investigación; Metadatos; Ciencias sociales; Humanidades; *Re3data*.

**Gómez, Nancy-Diana**; **Méndez, Eva**; **Hernández-Pérez, Tony** (2016). "Social sciences and humanities research data and metadata: A perspective from thematic data repositories". *El profesional de la información*, v. 25, n. 4, pp. 545-555.

*http://dx.doi.org/10.3145/epi.2016.jul.04*

## 1. Introduction

Research data management is becoming increasingly important in all scientific fields. A logical and necessary evolution due, on the one hand, to the technological development that increasingly allows a science based on data, and secondly, to the political impetus of the idea of open science that includes, besides open access to publications, the opening of the data used in the research process.

Sharing research data has become standard practice in disciplines where there is a collaborative scientific culture, such as physics, astronomy (**Pepe** *et al.*, 2014), and genetics (**Paltoo** *et al.*, 2014). This disciplinary culture is further compounded by the fact that publicly funded research institutions are beginning to require researchers to publish the results, not only in the form of publications, but also by opening the underlying data used. Opening research data is recommended by *OECD* (2015) and required by the US government and various funding agencies such as the *National Science Foundation* (*NSF*, 2014) and *National Institutes of Health* (*NIH*, 2015). In Europe, open access to research data has been, so far, only a pilot (*ORD Pilot*) for nine areas of projects funded under *Horizon 2020* with other areas and programs invited to voluntarily participate (*European Commission*, 2016). However, on April 19, 2016, the *Commission* stated that by 2017 research data will be open by default for all new *H2020* funded projects (COM 2016, p. 8).

The trend towards open data is growing within all institutions involved in research, both by the agencies that fund, and the organizations that carry out research (e.g. *League of European Research Universities*, *LERU*, 2013), and by journal editors who publish research results (e.g. *PLoS*, 2014). Although this trend varies from one discipline to another and between individual researchers, there are many motivations for sharing data (**Kim**; **Stanton**, 2016) and benefits that transcend trends or mandates (**Lyon**, 2016):

- increases the possibility of research having more impact and visibility;
- favors the reproducibility of science;

- saves costs when creating data;
- promotes collaboration;
- contributes to increased credibility in the system.

Of course, there are also many researchers reluctant to share "their" data. A study carried out by *Wiley* surveyed 2,886 researchers (**Ferguson**, 2014) and revealed some of their concerns:

- afraid of the negative consequences of sharing data (misuse, legal or commercial consequences, etc.);
- lack of recognition;
- amount of work involved in preparing the data for publication;
- lack of knowledge about how and where to share data.

> For each discipline or scientific domain there is a unique interpretation of datasets or datasets' research, their nature, data collection, and metadata description

## 1.1. Research data: a discipline problem as seen from the social sciences and humanities (SSH)

For each discipline or scientific domain there is a particular interpretation of datasets and research data, their nature, and collection procedures. And of course, variations in the way that data are described with metadata and the problems associated with sharing. **Christine Borgman**, who has extensively dealt with this (**Borgman**, 2008; **Borgman**; **Wallis**; **Mayernik**, 2012) refers to the concept of data as:

> "Facts, numbers, letters and symbols that describe an object, idea, condition, situation or other factors" and also "digital manifestations of literature (including text, sound, still images, moving images, models, games or simulations)".

Moreover, the *NSF* in the USA distinguishes between observational data, computer data, and experimental data, but all are considered digital (**Borgman**; **Wallis**; **Mayernik**, 2012).

However, in the social sciences and humanities (SSH) not all data are collected digitally and data may take many other forms and formats. For example, in sociology the data from surveys and interviews can easily be captured digitally; however, in archeology the results of observational data can be more closely linked to the object and to the background information about the object [geographical coordinates, samples and drawings of the object (on paper), photographs, or videos (digital)] (**Frank**; **Yakel**; **Faniel**, 2015).

Another key issue in SSH is the source of the data, because many investigations are based on data that were not originally produced by or for the researchers. For example, government data and corporate documents which are used to generate new data, that is, data used "for" research to generate other data "from" research. Humanities scholars are much more dependent on external data sources than researchers from other disciplines. Almost every record of human activity can be considered "data" (**Borgman**, 2008). Compared with those of pure sciences, SSH researchers generate much less data through observations, since generally they tend to use data from all kinds of sources, which may include sounds for linguistic studies and films for object, dress, or speech analysis. They also use historic materials, such as books, maps, newspapers, journals, photographs, and administrative records —as a result research data and publications may be confused or intermingled.

The *National Endowment for the Humanities (NEH)* in the USA defines data as materials generated or collected in the course of an investigation, for example, citations, software code, databases, geospatial coordinates, reports, and articles. However, the *NEH* expressly excludes article drafts and communications with colleagues (*NEH*, 2015). Furthermore, within the broad spectrum of subjects and disciplines covering the humanities, there can be different definitions of data, which can further complicate the outlook for their management and recovery.

Perhaps the unique characteristics of SSH researchers helps explain why only 46% of them share data in repositories (**Meadows**, 2014). Or, perhaps it is a lack of knowledge about where and how to share, the fuzzy boundaries between data and publications, and between data "from" research and "for" research.

## 1.2. Metadata or how to make data useful for research

Unlike what happens with publications, where despite different disciplinary styles there is a common core of formal properties, scientific data show a heterogeneity that varies radically across disciplines, thematic areas, and even between research groups and researchers.

The *NSF* in the US requests that the *data management plan* includes the metadata standards that are used (**Bischoff**; **Johnston**, 2015). The pilot open data (*ORD Pilot*) of the *European Commission* (2016) further requests that the metadata associated with data –ultimately what makes the data useful - be included. In the world of digital libraries, metadata have always contributed to making data useful by describing publications and other digital or digitized objects or assets. And in the world of data, metadata makes data useful: describing, dimensioning, and contextualizing so that they can be found, regardless of the silo discipline in which are situated, enabling reuse across other domains. Without metadata and descriptions of research methods and context, data are just collections of numbers, codebooks, pretty pictures, or boxes of stones (**Borgman**, 2008).

Funding agencies are raising awareness and putting pressure on researchers to manage their data, share data in a reusable way, facilitate the recovery and preservation of data, and ensure that data are FAIR (findable, accessible, interoperable, and reusable). The creation of FAIR data and science highlights the need to improve the e-infrastructure for scientific information reuse (**Wilkinson** *et al.*, 2016), but also the need to promote interoperability from the metadata.

When researchers share their data and metadata in a data repository, they should translate the meta-information they use in their VREs (virtual research environments), on their servers, and on their personal computers –what **Tenopir** *et al.* (2015) called *laboratory metadata* or *institution specific metadata*- into the standard metadata schema used in the repository. **Tenopir** and his research team surveyed more than 1,000 researchers in each of its two studies, conducted in 2011 and 2015, on how to manage their data (**Tenopir** *et al.*, 2011); more than 50% said they did not use any metadata standard, 14% said they used some standard within their institution, and 20% used their laboratory standard (in the 2011 study); the 2015 study found similar results (47.9% none and 16.7% laboratory standard). In our study we analyzed the metadata schemes used by repositories, or at least those schemes that repository administrators claim to use to describe the data deposited in *re3data* by SSH researchers.

> The creation of data and science FAIR underscores the need to improve the e-infrastructure for the reuse of scientific information and the need to promote interoperability from metadata

## 2. Objectives and methodology

According to the context that we provided in the previous section, this article focuses on two domains (social sciences and humanities) where there has not been a historic tradition of collaboration, managing research data, standardized metadata schemes (with some exceptions), virtual research environments, or other e-infrastructures that require the use of metadata. We address the problem of scientific data management in SSH, through a study of data repositories of those disciplines, included in *re3data* (a repository funded by the *German Research Foundation*), to answer the following research questions:

- What kind of data are stored and managed by specific SSH repositories?
- How is the distribution of research data repositories among the various areas of knowledge within SSH?
- What thematic areas are most represented?
- What metadata schemes are used in these repositories to identify and describe the different types of data?
- Is there a predominant scheme or model in each case?

### 2.1. Objectives

- To identify SSH research data repositories.
- To study what types of data result from research in these disciplines by analyzing data stored in major repositories.
- To present the metadata schemes most used in these repositories.

This is an exploratory study to identify the most representative SSH specialized repositories, to investigate their practices, and to verify the type of stored data and metadata schemes that they use or claim to use.

### 2.2. Methodology

For the analysis we used the aforementioned *re3data* (*Registry of Research Data Repositories*) as a source, because it is a reference registry for data repositories recommended by both the *European Commission* (2016), and various publishers (*PeerJ*, *Springer*, *Nature's Scientific Data*, etc.). This registry enables easy identification of data repositories by subject or discipline.

Initially a quantitative and analytical methodology to analyze the 397 repositories included in the SSH category of *re3data* was considered. However, during the initial phase of the research the course was changed to carry out a detailed study of a small sample of the most representative SSH specialized repositories, three in social sciences and three in humanities, to verify the declared metadata schemes. Thus, the work was carried out in three phases:

**a) Extraction and treatment of *re3data* records**

In this phase, several tasks were carried out:

a.1. Retrieval and extraction, through the API provided by *re3data*, of a total of 1,457 registered repositories (at the time of data collection, February, 18th 2016). Please note that in April 2016, *re3data* announced that it has already reached 1,500 data repository records. Although the latest version of the descriptive scheme (metadata) of *re3data* is 3.0 (**Rücknagel** *et al.*, 2015), the API responds to the first version of the scheme, which is much more limited than the latest version.

a.2. From the repositories list, 1,457 records describing them were downloaded in xml format using scraping techniques with *R*[1].

a.3. The records were treated through xslt[2] to process the information for this study, mainly: data and metadata types used by the repositories and their classification and identification schemes.

**b) Sample selection and quantitative analysis of the extracted data**

The objective of this phase was to filter the data repositories to which we wanted to focus the study, those with SSH content. We selected those that contained some thematic classification scheme on humanities or social sciences in their description, according to the classification used by *re3data* that can be seen in table 1.

In the case of thematic classification, it should be noted that to classify a repository according to its metadata schema, *re3data* provides the property *SubjectScheme* as a mandatory attribute that allows researchers to enter an unlimited number of values, always bearing in mind that the only allowed values are those from the thematic classification of the *German Research Foundation* (*DFG Classification of subject area*). This classification covers four broad areas:

- humanities and social sciences
- life sciences

- natural sciences
- engineering sciences.

It should be taken into account that each repository can be described with as many subjects as it covers, so that a single repository may appear in more than one subject area and even simultaneously in the four thematic areas, as happens in multidisciplinary cases.

After filtering, we obtained 397 records related to SSH, which constituted our sample size to analyze the types of data and metadata schemas of each repository.

To identify the type of data, the *ContentType* property (not mandatory) of the *re3data* scheme was used, which allows specification of all types of content available in a repository. The values allowed in this field are restricted to the types of content recognized and identified in the *Parse.insight* (*Permanent Access to the Records of Science in Europe*) project. The *Parse* classification has 15 options:

-Archived data
-Audiovisual data
-Configuration data
-Databases
-Images
-Network based data
-Plain text
-Raw data
-Scientific and statistical data formats
-Software applications
-Source code
-Standard office documents
-Structured graphics
-Structured text
-Other.

Table 1. Number of SSH repositories, including multidisciplinary repositories, according to the *DFG* classification

| ***SubjectScheme* de *DFG*** | **Nunber of repositories** |
|---|---|
| 1 Humanities and social sciences | 397 |
| 101 Ancient cultures | 15 |
| 102 History | 34 |
| 103 Fine arts, music, theatre and media studies | 26 |
| 104 Linguistics | 47 |
| 105 Literary studies | 10 |
| 106 Non-European languages and cultures, social and cultural anthropology, Jewish studies and religious studies | 18 |
| 107 Theology | 4 |
| 108 Philosophy | 3 |
| 109 Education sciences | 146 |
| 110 Psychology | 14 |
| 111 Social sciences | 155 |
| 112 Economics | 114 |
| 113 Jurisprudence | 27 |

However, it is important to note that it is not mandatory to select one of them when completing the registration on the repository.

Finally, to identify the metadata schema, we used the *MetadataStandardName* scheme property, which again is not mandatory.

**c) Identification of a subset of key data repositories in SSH**

Once we identified the subset, a qualitative and individual analysis of the metadata schemes was conducted.

This last phase of the methodology was included because we identified two limitations of *re3data*:

- to complete / declare the metadata schema used by a repository is not mandatory;
- the information about the standard used is the one at the time when the registration was completed and it might have changed over time.

> The idiosyncrasies of social sciences and humanities researchers may lead many of them to withhold their data, but this withholding may also be the result of ignorance about where and how to share

When accessing the repositories other difficulties were revealed:

- corroborating the metadata schemes declared in *re3data*;
- restricting access to authorized users, in some cases;
- missing manual or bibliography, etc., in some cases.

So, to continue the study, three repositories in social sciences and three in the humanities were selected based on:

- coverage or number of datasets stored;
- level of use made by their respective communities;
- representation for this study, covering various topics and countries.

In the case of humanities, a repository of linguistics (*Clarin*), one of archeology, and another of history and art (*Prometheus*) were selected, because these subdisciplines were represented by the data repositories in *re3data*. Selected repositories are shown in table 2.

## 3. Results and discussion

### 3.1. Research data repositories in SSH

A first overview of the existing repositories registered in *re3data* SSH, can be seen in figure 1: A *treemap* representing the number / volume of repositories of the areas studied, according to the sub-classification of SSH in table 1.

In order to give a more accurate picture, a table with the number of repositories according to the *DFG* classification and the *SubjectScheme* used by *re3data*, which includes four levels, is provided. In table 1 it is indicated to the third level.

It is to be noted that a multidisciplinary repository may be in more than one category, so the sum of the parts exceeds the total. According to the thematic classification of the *DFG*, social sciences (codes 109 to 113) have a higher representation: 456 versus 157 in the humanities.

Table 2. Selection of representative repositories (SSH)

| **Social sciences** | |
|---|---|
| *Inter university Consortium for Political and Social Research* (*ICPSR*, EUA) | *http://www.icpsr.umich.edu* |
| *UK Data Service* (Reino Unido) | *https://www.ukdataservice.ac.uk* |
| *Gesis Zacat* (Alemania) | *http://zacat.gesis.org/webview* |
| **Humanities** | |
| *Common Language Resources and Technology Infrastructure* (*Clarin*, EU): | *http://www.clarin.eu* |
| *Archaeology Data Service* (Reino Unido) | *http://archaeologydataservice.ac.uk* |
| *Prometheus* (Alemania) | *http://www.prometheus-bildarchiv.de* |

### 3.2. SSH research data

The types of content available in *re3data* were represented according to the types recognized and identified in the *Parse.insight* project. And, as noted, the type of scientific and statistical data (formats such as spss, fits, gis, etc.) along with documents (*Word*, *Excel* or similar *OpenOffice* formats) and images (jpeg, jpeg2000, gif, tif, png, svg, etc.) are the most commonly used in digitization projects in the humanities. Figure 2 shows that the proportion of content types is relatively balanced in all areas of science, in contrast with the use made in humanities and social sciences. For each type of data (scientific data, images, plain text, raw data, etc.) the use that is done in SSH is usually about 27% (a minimum of 20% and a maximum of 32%). It is surprising that there was not a higher percentage of "standard office documents" as compared to "scientific and statistical data formats" or "raw data" type. All document types are present with more or less the same proportions, (73% in other disciplines and 27% in SSH). Even in "audiovisual data", near the end of the graph, there were fewer "standard office documents", and the proportion (69.3% in other disciplines and 30.7% in SSH) was maintained.

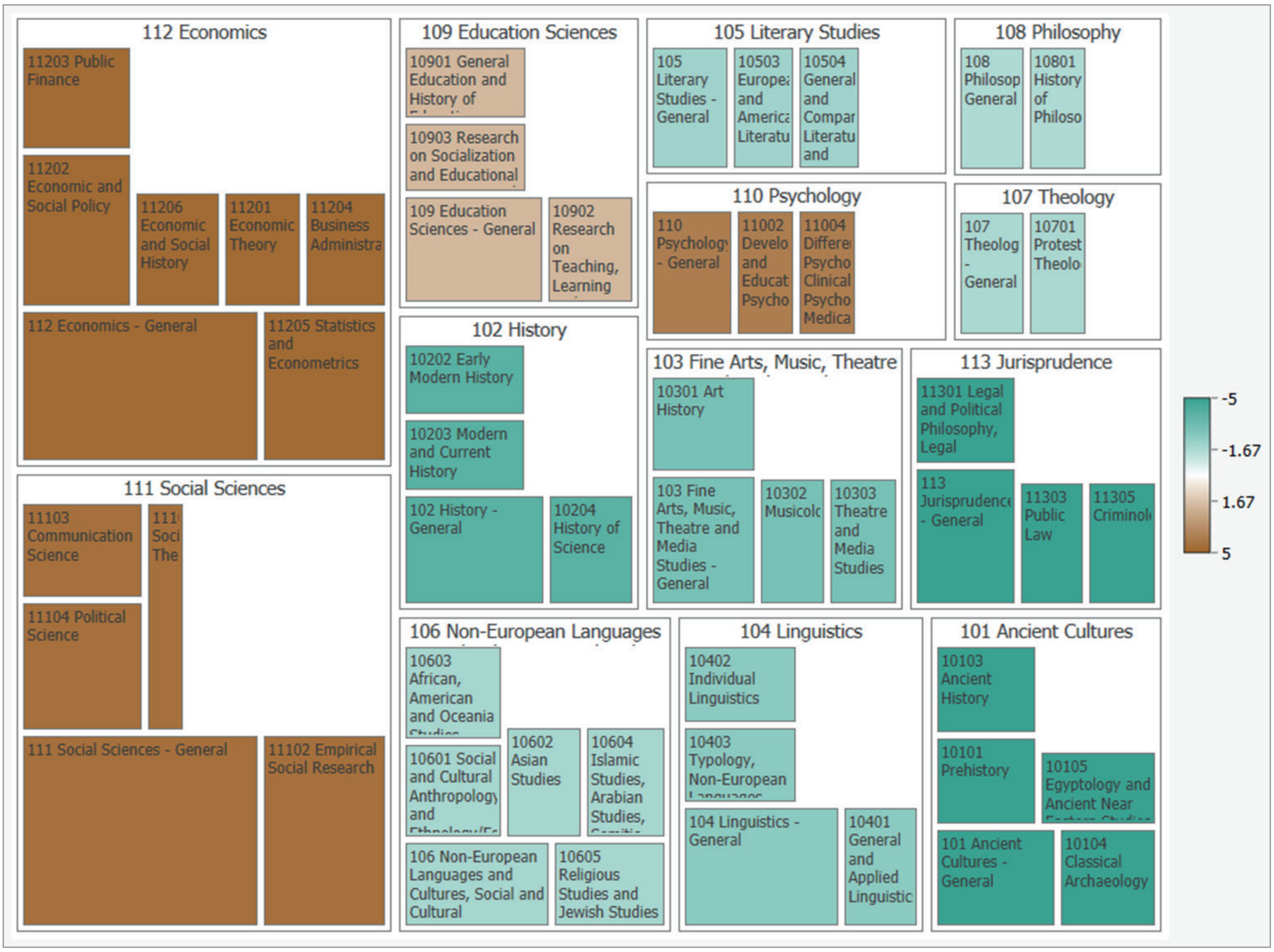


Figure 1. Proportional representation of the SSH repositories in *re3data*, including multidisciplinary repositories

Table 3. Metadata schemas in representative repositories of SSH

| Repository | Metadata schema |
|---|---|
| **Social sciences** | |
| *Inter university Consortium for Political and Social Research ICPSR* (EUA)<br>*http://www.icpsr.umich.edu* | DDI<br>DC |
| *UK Data Service* (Reino Unido)<br>*https://www.ukdataservice.ac.uk* | DDI, DC, ISO 19115, METS (*Metadata encoding and transmission standard*), ISAD (*International standard archival description*) |
| *Gesis Zacat* (Alemania)<br>*http://zacat.gesis.org/webview* | DDI<br>DC |
| **Humanities** | |
| *Common Language Resources and Technology Infrastructure* (*Clarin*, EU)<br>*http://www.clarin.eu* | IMDI (*ISLE meta data initiative*), TEI headers, DC, DCTerms, DC-OLAC (*Open language archive community*) (**Van-Uytvanck**; **Stehouwer**; **Lampen**, 2012) |
| *Archaeology Data Service* (Reino Unido)<br>*http://archaeologydataservice.ac.uk* | ADS Schema<br>DC<br>MIDAS |
| *Prometheus* (Alemania)<br>*http://www.prometheus-bildarchiv.de* | EDM (*Europeana data model*)<br>METS<br>DC |

## 3.3. Metadata schemes used in SSH research data repositories

22.8% (332) of all *re3data* repositories specify the metadata scheme/s they use. As seen in figure 3, in the field of SSH (in lighter color) *Dublin Core* and DDI (*Data documentation initiative*) are by far the most used. The reason for "other" being the highest value is that it is a non-mandatory field in all versions of *re3data* scheme, and indicates the wide variety of metadata used in all disciplines, with a few dominant schemes in certain areas, and many specific variations in those disciplines in which no scheme stands out as dominant.

Moreover, 25.2% of SSH repositories declare some type of metadata schema. Of these, 45% use *Dublin Core*, the most common metadata model in 45 repositories. Both DDI and "other", are second with 37% each, used in 37 repositories. It should be noted that "other" refers to homegrown metadata schemes (of the institution or of the laboratory). Both the graphic representation of the situation and the metadata schema name and number repositories that use them can be seen in figure 4. It is noteworthy that 74.8% of the repositories do not provide this information, and that multidisciplinary repositories may use more than one scheme, so it is possible to find schemes from other scientific areas.

In order to review the metadata schemes used in SSH, six representative repositories (table 2) were selected. We studied them identifying the metadata schema used, either by analyzing the repository, or looking at the repository site guidance or instructions for the deposit, or looking for papers on the repositories studied where this information was declared.

The detailed analysis of the selected repositories confirms

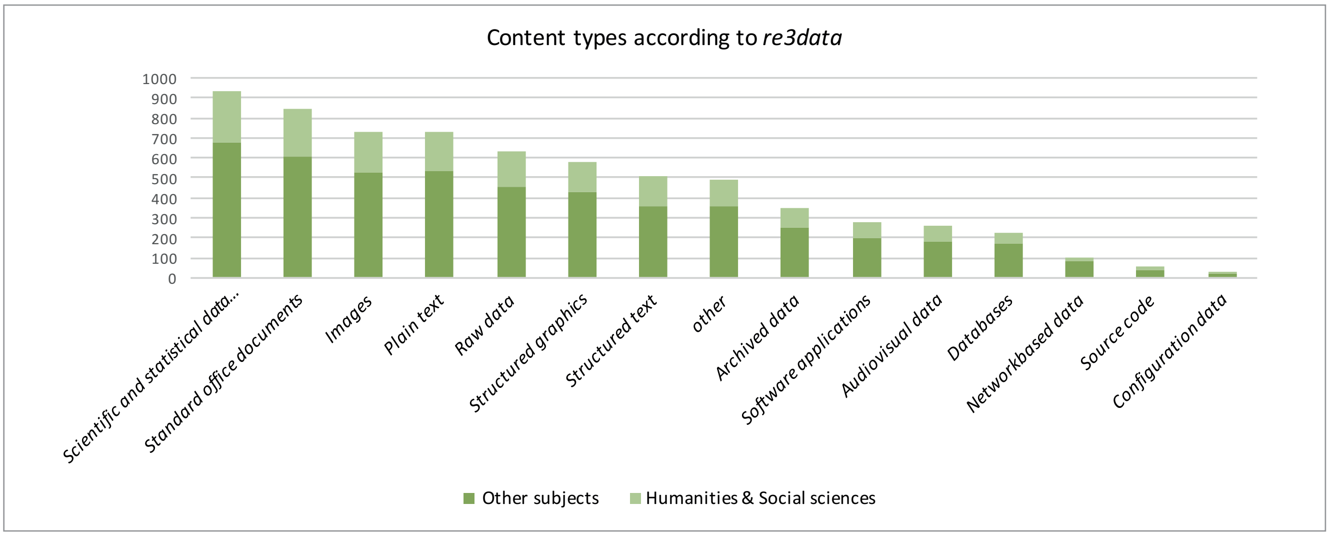


Figure 2. Types of content declared in *re3data*. In the vertical axis there is the number of repositories where each type of content was found.

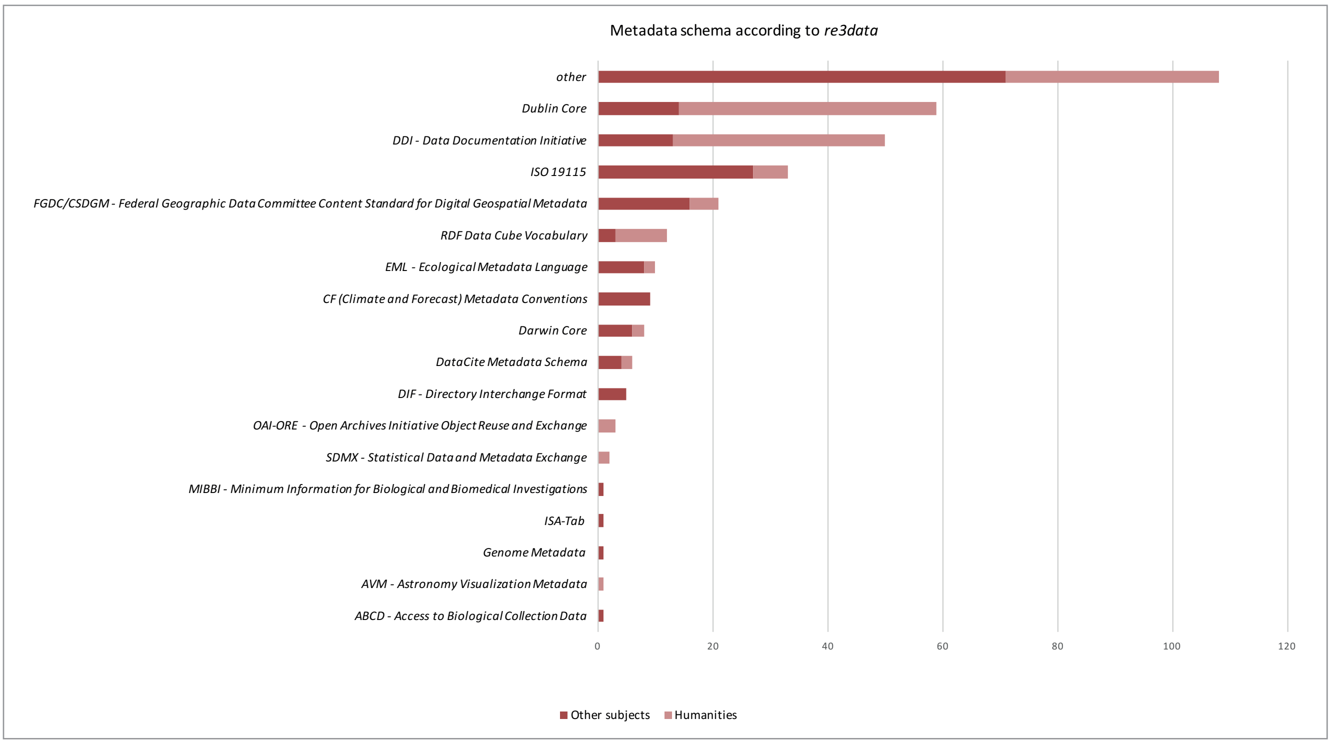


Figure 3. Metadata schemes declared by the data repositories included in *re3data*

the trend that *re3data* offers in social science: the dominant metadata scheme is DDI (*Data documentation initiative*), an international standard for describing statistical data and social science data with great tradition. DDI describes the data resulting from observation methods in social, behavioral, economic, and health sciences. It takes into account the data collection processes, the varying levels of description, and methods. It is a scheme that could be called classic, since it was originated in 1995, when the *Dublin Core* appeared within the social science community, and with the objective of describing data. Since then it has evolved steadily, maintained by the *DDI Alliance* (**Vardigan**, 2013).
*http://www.ddialliance.org*

> The heterogeneity and complexity of research data repositories is manifested in the metadata schemes that are chosen to describe them, which is even more evident in the humanities

In the case of the humanities, the metadata schemes used are more diverse and particular, as shown by the selected repositories analyzed in this article (table 3). However, most schemes are not found within the repositories registered in *re3data*. This explains the high percentage for the category "other" (37%) in the humanities repositories, because until version 3.0 of the *re3data* scheme, only metadata standards collected by the *Digital Curation Centre* were recognized as allowable values.
*http://www.dcc.ac.uk/resources/metadata-standards*

This diversity of metadata, or lack of common or regular humanities standards, is justified in the heterogeneity of data repositories and to what is considered as "data" in humanities, as discussed in the introduction.

Within SSH the use of *Dublin Core* (DC) is extensive (figure 4). This predominance is due to:

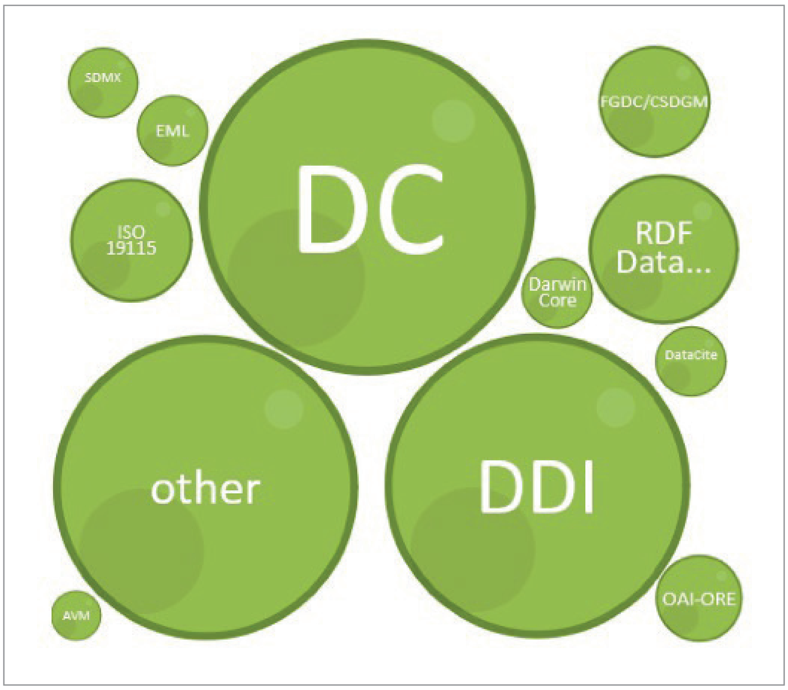
- a linkage with document / publications repositories; and a lack of distinction between these and data repositories, and
- the level of standardization that DC has attained and its interoperability OAI-PMH between repositories.

We agree with the argument given by **Willis**, **Greenberg** and **White** (2012) that creators of metadata schemes are more likely to change and adapt or enhance an existing scheme than to create a new one. Once the DC has been installed, it is easier to adapt than it is to adopt a new schema.

## 4. Conclusions

The main conclusion we draw from this study is the corroboration of the heterogeneity and complexity of research data repositories, which is glaring within the humanities. This heterogeneity is manifested in the metadata schemes that researchers choose for description. We have reached several conclusions in the course of this work:

1) Following the merger between *Databib* and *re3data* in the same registry at the end of 2015, *re3data* has become the registry par excellence for finding research data repositories in all disciplines; which we used to identify and analyzed 397 repositories in SSH. Its greatest weakness, for now, is that it lacks mechanisms to know when a data repository record has been modified or how to change the characteristics initially declared. The information about the repositories cannot be updated online. Since February 2016 this



| Scheme name | N. |
|---|---|
| *Dublin Core* | 45 |
| DDI | 37 |
| Other | 37 |
| *RDF Data cube vocabulary* | 9 |
| *ISO 19115* | 6 |
| *FGDC/CSDGM* | 5 |
| *OAI-ORE* | 3 |
| *EML* | 2 |
| *Darwin Core* | 2 |
| *DataCite* | 2 |
| *SDMX* | 2 |
| *AVM* | 1 |

Figure 4. Metadata schemes used in SSH according to *re3data*

problem has been alleviated by sending a form to *re3data* requesting the needed changes. A manual mechanism that is, hopefully, temporary.

The metadata scheme used by *re3data* in its current version (v. 3.0) describes repositories and incorporates some characteristics about reuse, metrics and policies. This model seems to be evolving in the right direction if it does not include a large increase the existing set of characteristics. Automation mechanisms and online editing, as now happens with publications repositories and aggregators, should be implemented.

> *Re3data* (*Registry of Research Data Repositories*) is the source of reference for identifying repositories to deposit research data classified by subject or discipline

The *DFG* thematic classification used by *re3data* is too generic. Therefore, it is not easy to narrow the theme of each repository because the vast majority are declared multidisciplinary, but often they are not, or they are multidisciplinary in a very small way.

2) Taking into account the limitations of *re3data* to describe the repositories, we can say that data and metadata schemas are less homogeneous in humanities than in social sciences. Despite the small number of data repositories that declare the metadata standard used, *re3data* confirms the trend of use of DDI metadata schema in social sciences. This may be due to the maturity of the standard, its amount of implementations, and that it was a scheme that was originally created to describe data, not documents. It is something similar to the case of digital geospatial information systems, where, since the mid-90s, *FGDC* (*Federal Geographic Data Committee*) and *ISO 19115* standards have been used to describe geospatial data infrastructures.

The adoption of DDI by some of the most important repositories such as *Icpsr*, *Gesis,* and the *Dataverse* network of data repositories bode well for the future of metadata standards in social sciences, where “from” (and “for”) research data support statistics, surveys, opinion polls, etc., to which the DDI standard has been addressed from its inception.

3) In humanities the situation is more complex and diverse. *Dublin Core* (DC) seems to be widely used according to the generic data extracted from *re3data*, but if we drill down to the details of data repositories on specific fields such as linguistics or archeology, we see that they are using their own schemes or adapting DC to a greater or lesser extent. It should also be noted that many humanities projects, especially on text digitization, use *TEI Header* linked to the *TEI* (*Text encoding initiative*) standard, while in other cases they lack description schemes. The exposure of their research data is done simply through content managers with little use of metadata. Also it is not unusual to see metadata schemas used to describe humanities data in their data repositories, standards used for library creation, or for the description of textual publications, images, or audiovisuals (not only DC, but also EDM, METS, and MIDAS). This happens because of the tenuous differentiation in some of these disciplines, between data and documents, and between data “from” and “for” research.

4) *Dublin Core* (DC) is the default standard for publications’ repositories, and this trend includes data repositories, at least in the first instance or approach. Although DC has well established mechanisms to create application profiles that fit the description of any type of information or private co-

llection, it is still too early to confirm whether this standard can be adapted to the idiosyncrasies of all disciplinary research data.

## Notes

1. Web scraping (web harvesting or web data extraction) is a computer software technique for extracting information from websites.

*R* is a programming language and software for statistical computing and graphics supported by the *R Foundation for Statistical Computing*. It is widely used among statisticians and data miners for developing statistical software and data analysis.

2. Xslt (extensible stylesheet language transformations) is a language for transforming xml documents into other xml documents, or other formats such as html for web pages, plain text or into xsl formatting objects, which may subsequently be converted to other formats, such as pdf, postscript and png.

## Acknowledgements

This work is part of the project *Curator-e: custody and digital management for the reuse of open data research in the humanities and social sciences*, funded by the Spanish *State Program of Research, Development and Innovation Facing Society Challenges*, *Ministerio de Economía y Competitividad* (*Mineco*), Spain (CSO2013-46754-R).

## 5. References

**Bishoff, Carolyn**; **Johnston, Lisa** (2015). "Approaches to data sharing: An analysis of NSF data management plans from a large research university". *Journal of librarianship and scholarly communication*, v. 3, n. 2, p. eP1231.
*http://dx.doi.org/10.7710/2162-3309.1231*

**Borgman, Christine L.** (2008). "Data, disciplines, and scholarly publishing". *Learned publishing*, v. 21, n. 1, pp. 29-38.
*http://dx.doi.org/10.1087/095315108X254476*

**Borgman, Christine L.**; **Wallis, Jillian C.**; **Mayernik, Matthew S.** (2012). "Who's got the data? Interdependencies in science and technology collaborations". *Computer supported cooperative work* (*CSCW*), v. 21, n. 6, pp. 485-523.
*http://nldr.library.ucar.edu/repository/assets/osgc/OSGC-000-000-012-014.pdf*
*http://dx.doi.org/10.1007/s10606-012-9169-z*

COM (2016) 178 final. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: European cloud initiative - Building a competitive data and knowledge economy in Europe.
*http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=15266*

*European Commission* (2016). *Guidelines on open access to scientific publications and research data in Horizon 2020, v. 2.1*. European Commission. Directorate General for Research and Innovation.
*http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf*

**Ferguson, Liz** (2014). "How and why researchers share data (and why they don't)". *Wiley Exchanges*. *Discover the future of research*, 3 November.
*https://hub.wiley.com/community/exchanges/discover/blog/2014/11/03/how-and-why-researchers-share-data-and-why-they-dont?referrer=exchanges*

**Frank, Rebecca D.**; **Yakel, Elizabeth**; **Faniel, Ixchel M.** (2015). "Destruction/reconstruction: preservation of archaeological and zoological research data". *Archival science*, v. 15, n. 2, pp. 141-167.
*http://dx.doi.org/10.1007/s10502-014-9238-9*

**Kim, Youngseek**; **Stanton, Jeffrey M.** (2016). "Institutional and individual factors affecting scientists' data-sharing behaviors: A multilevel analysis". *Journal of the Association for Information Science and Technology*, v. 67, n. 4, pp. 776-799.
*https://www.asis.org/asist2013/proceedings/submissions/papers/123paper.pdf*
*http://dx.doi.org/10.1002/asi.23424*

*LERU* (2013). *LERU roadmap for research data*. Advice paper n. 14.
*http://www.leru.org/files/publications/AP14_LERU_Roadmap_for_Research_data_final.pdf*

**Lyon, Liz** (2016). "Transparency: the emerging third dimension of open science and open data". *Liber quarterly*, v. 25, n. 4.
*http://dx.doi.org/10.18352/lq.10113*

**Meadows, Alice** (2014). "To share or not to share? That is the (research data) question…". *The scholarly kitchen*, 11 November.
*http://scholarlykitchen.sspnet.org/2014/11/11/to-share-or-not-to-share-that-is-the-research-data-question*

*NEH* (2015). *Data management plans for NEH Office of Digital Humanities. Proposals and awards*.
*http://www.neh.gov/files/grants/data_management_plans_2015.pdf*

*NIH* (2015). "NIH sharing policies and related guidance on NIH-funded research resources". *National Institutes of Health.*
*https://grants.nih.gov/policy/sharing.htm*

*NSF* (2014). "Chapter II. Proposal preparation instructions". *Grant proposal guide.* National Science Foundation. Where discoveries begin.
*http://www.nsf.gov/pubs/policydocs/pappguide/nsf15001/gpg_2.jsp#dmp*

*OECD* (2015). *Making open science a reality*. Organisation for Economic Co-operation and Development.
*https://www.innovationpolicyplatform.org/sites/default/files/DSTI-STP-TIP%282014%299-REV2_0_0_0_0.pdf*

**Paltoo, Dina N.**; **Rodriguez, Laura-Lyman**; **Feolo, Michael**; **Gillanders, Elizabeth**; **Ramos, Erin M.**; **Rutter, Joni L.**; **Sherry, Stephen**; **Wang, Vivian-Ota**; **Bailey, Alice**; **Baker, Rebecca**; **Caulder, Mark**; **Harris, Emily L.**; **Langlais, Kristofor**; **Leeds, Hilary**; **Luetkemeier, Erin**; **Paine, Taunton**; **Roomian, Tamar**; **Tryka, Kimberly**; **Patterson, Amy**; **Green, Eric D.** (2014). "Data use under the NIH GWAS data sharing policy and future directions". *Nature genetics*, v. 46, n. 9, pp. 934-938.
*http://dx.doi.org/10.1038/ng.3062*

**Pepe, Alberto**; **Goodman, Alyssa**; **Muench, August**; **Crosas, Merce**; **Erdmann, Christopher** (2014). "How do astronomers share data? Reliability and persistence of datasets linked in AAS publications and a qualitative study of data practices among US astronomers". *PLoS one*, v. 9, n. 8, p. e104798.
*http://dx.doi.org/10.1371/journal.pone.0104798*

*PLoS* (2014). "PLoS data policy prior to March 3, 2014". *PLoS*.
*http://goo.gl/QIRIab*

**Rücknagel, Jessika**; **Vierkant, Paul**; **Ulrich, Robert**; **Kloska, Gabriele**; **Schnepf, Edeltraud**; **Fichtmüller, David**; **Reuter, Evelyn**; **Semrau, Angelika**; **Kindling, Maxi**; **Pampel, H.**; **Witt, Michael**; **Fritze, Florian**; **Van-de-Sandt, Stephanie**; **Klump, Jens**; **Goebelbecker, Hans-Jürgen**; **Skarupianski, Michael**; **Bertelmann, Roland**; **Schirmbacher, Peter**; **Scholze, Frank**; **Kramer, Claudia**; **Fuchs, Claudio**; **Spier, Shaked**; **Kirchhoff, Agnes** (2015). *Metadata schema for the description of research data repositories*, v. 3.0.
*http://dx.doi.org/10.2312/re3.008*

**Tenopir, Carol**; **Allard, Suzie**; **Douglass, Kimberly**; **Aydinoglu, Arsev-Umur**; **Wu, Lei**; **Read, Eleanor**; **Manoff, Maribeth**; **Frame, Mike** (2011). "Data sharing by scientists: Practices and perceptions". *PLoS one*, v. 6, n. 6.
*http://dx.doi.org/10.1371/journal.pone.0021101*

**Tenopir, Carol**; **Dalton, Elizabeth D.**; **Allard, Suzie**; **Frame, Mike**; **Pjesivac, Ivanka**; **Birch, Ben**; **Pollock, Danielle**; **Dorsett, Kristina** (2015). "Changes in data sharing and data reuse practices and perceptions among scientists worldwide". *PLoS one*, v. 10, n. 8, p. e0134826.
*http://dx.doi.org/10.1371/journal.pone.0134826*

**Van-Uytvanck, Dieter**; **Stehouwer, Herman**; **Lampen, Lari** (2012). "Semantic metadata mapping in practice: the virtual language observatory". En: *LREC 2012: 8th Intl conf. on language resources and evaluation. European Language Resources Association* (*ELRA*), pp. 1029-1034.
*http://goo.gl/IMgP4c*

**Vardigan, Mary** (2013). "Timeline DDI". *Iassist quarterly*, v. 37, pp. 51-55.
*http://www.iassistdata.org/sites/default/files/iq/iqvol371_4_vardigan2.pdf*

**Wilkinson, Mark D.**; **Dumontier, Michel**; **Aalbersberg, Ijsbrand-Jan**; **Appleton, Gabrielle**; **Axton, Myles**; **Baak, Arie**; **Blomberg, Niklas**; **Boiten, Jan-Willem**; **Da-Silva-Santos, Luiz-Bonino**; **Bourne, Philip E.**; **Bouwman, Jildau**; **Brookes, Anthony J.**; **Clark, Tim**; **Crosas, Mercè**; **Dillo, Ingrid**; **Dumon, Olivier**; **Edmunds, Scott**; **Evelo, Chris T.**; **Finkers, Richard**; **González-Beltrán, Alejandra**; **Gray, Alasdair J.G.**; **Groth, Paul**; **Goble, Carole**; **Grethe, Jeffrey S.**; **Heringa, Jaap**; **Hoen, Peter A.C't**; **Hooft, Rob**; **Kuhn, Tobias**; **Kok, Ruben**; **Kok, Joost**; **Lusher, Scott J.**; **Martone, Maryann E.**; **Mons, Albert**; **Packer, Abel L.**; **Person, Bengt**; **Rocca-Serra, Philippe**; **Roos, Marco**; **Van-Schaik, Rene**; **Sansone, Susanna-Assunta**; **Schultes, Erik**; **Sengstag, Thierry**; **Slater, Ted**; **Strawn, George**; **Swertz, Morris A.**; **Thompson, Mark**; **Van-der-Lei, Johan**; **Van-Mulligen, Erik**; **Velterop, Jan**; **Waagmeester, Andra**; **Wittenburg, Peter**; **Wolstencroft, Katherine**; **Zhao, Jun**; **Mons, Barend** (2016). "The FAIR guiding principles for scientific data management and stewardship". *Scientific data*, v. 3, p. 160018.
*http://dx.doi.org/10.1038/sdata.2016.18*

**Willis, Craig**; **Greenberg, Jane**; **White, Hollie** (2012). "Analysis and synthesis of metadata goals for scientific data". *Journal of the American Society for Information Science and Technology*, v. 63, n. 8, pp. 1505-1520.
*http://dx.doi.org/10.1002/asi.22683*

DIGIBÍS transforma y enriquece las bibliotecas con aplicaciones de gestión digital de nueva generación.
El usuario accede fácilmente a los recursos de la biblioteca desde cualquier lugar, en cualquier momento, con cualquier dispositivo.
La biblioteca es recolectada automáticamente por Hispana, Europeana y otros agregadores internacionales.