State of the art report on open access publishing of research data in the humanities

Stefan Buddenbohm, Nathanael Cretin, Elly Dijk, Bertrand Gaiffe, Maaike de Jong, Jean-Luc Minel, Nathalie Le Tellier-Becquart

Humanities at Scale

D7.1 State of the art report on open access publishing of research data in the humanities

HaS-DARIAH
INFRADEV-3-2015-Individual implementation and operation of ESFRI projects
Grant Agreement no.: 675570

Date: 07-05-2018
Version: 1.1

Project funded under the Horizon 2020 Programme

Grant Agreement no.:	675570
Programme:	Horizon 2020
Project acronym:	HaS-DARIAH
Project full title:	Humanities at Scale: Evolving the DARIAH ERIC
Partners:	DIGITAL RESEARCH INFRASTRUCTURE FOR THE ARTS AND HUMANITIES CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE KONINKLIJKE NEDERLANDSE AKADEMIE VAN WETENSCHAPPEN – KNAW GEORG-AUGUST-UNIVERSITAET GOETTINGEN STIFTUNG OEFFENTLICHEN RECHTS
Topic:	INFRADEV-3-2015
Project Start Date:	01-09-2015
Project Duration:	28 months
Title of the document:	State of the art report on open access publishing of research data in the humanities
Work Package title:	Open Data Infrastructure
Estimated delivery date:	1 September 2016
Lead Beneficiary:	DANS-KNAW
Author(s):	Stefan Buddenbohm [buddenbohm@sub.unigoettingen.de] Nathanael Cretin [nathanael.cretin@openedition.org] Elly Dijk [elly.dijk@dans.knaw.nl] Bertrand Gaiffe [bertrand.gaiffe@atilf.fr] Maaike de Jong [maaike.de.jong@dans.knaw.nl] Jean-Luc Minel (jean-luc.minel@u-paris10.fr) Blandine Nouvel (blandine.nouvel@frantiq.fr)
Quality Assessor(s):	Claudia Engelhardt [Claudia.engelhardt@sub.uni-goettingen.de] Julius Peinelt [peinelt@cmb.hu-berlin.de]
Keywords:	Humanities, open access, research data, digital humanities

Revision History

Version	Date	Author	Beneficiary	Description
0.1	16 May 2016	Elly Dijk (DANS-KNAW) with contributions of WP 7 partners	WP7 & WP8 members	First draft
0.2	25 July 2016	Stefan Buddenbohm (UGOE) Nathanael Cretin (CNRS) Elly Dijk (DANS) Bertrand Gaiffe (CNRS) Maaike de Jong (DANS) Jean-Luc Minel (CNRS) Blandine Nouvel (CNRS) Contributed with comments and editing: Claudia Engelhardt (SUB) Julius Peinelt (DARIAH)	HaS - DARIAH project participants	Second draft
1.0	12 August 2016	Stefan Buddenbohm (UGOE) Nathanael Cretin (CNRS) Elly Dijk (DANS) Bertrand Gaiffe (CNRS) Maaike de Jong (DANS) Jean-Luc Minel (CNRS) Blandine Nouvel (CNRS)	HaS - DARIAH project	Final version
1.1	4 May 2018	Stefan Buddenbohm (UGOE)	HaS – DARIAH project	Revised edition considering comments by reviewers

Table of Contents

Table of Figures

Executive Summary

Publishing research data as open data is not yet a common practice for researchers in the arts and humanities, and lags behind other scientific fields, such as the natural sciences. Moreover, even when humanities researchers publish their data in repositories and archives, these data are often hard to find and use by other researchers in the field. The goal of Work Package 7 of the HaS (Humanities at Scale) DARIAH project is to develop an open humanities data platform for the humanities. Work in task 7.1 is a joint effort of Data Archiving and Networked Services (DANS), Centre National de la Recherche Scientifique (CNRS) and the University of Gottingen – State and University Library (UGOE-SUB).

This report gives an overview of the various aspects that are connected to open access publishing of research data in the humanities. After the introduction, where we give definitions of key concepts, we describe the research data life cycle. We present an overview of the different stakeholders involved and we look into advantages and obstacles for researchers to share research data. Furthermore, a description of the European data repositories is given, followed by certification standards of trusted digital data repositories. The possibility of data citation is important for sharing open data and is also described in this report. We also discuss the standards and use of metadata in the humanities. Finally, we discuss and highlight one best practice example of open access research data system in the humanities: the French open research data ecosystem.

With this report we provide information and guidance on open access publishing of humanities research data for researchers. The report is the result of a desk study towards the current state of open access research data and the specific challenges for humanities. It will serve as input for Task 7.2, which will deliver a design plan for an open humanities data platform, and for Task 7.3, which will deliver this platform.

	Nature of the deliverable	
✓	R	Document, report
	DEM	Demonstrator, pilot, prototype
	DEC	Websites, patent fillings, videos, etc.
	OTHER	
	Dissemination level	
✓	P	Public
	CO	Confidential only for members of the consortium (including the Commission Services)
	EU-RES	Classified Information: RESTREINT UE (Commission Decision 2005/444/EC)
	EU-CON	Classified Information: CONFIDENTIEL UE (Commission Decision 2005/444/EC)
	EU-SEC	Classified Information: SECRET UE (Commission Decision 2005/444/EC)

Disclaimer

The Humanities at Scale is project funded by the European Commission under the Horizon 2020 programme. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

1. Introduction

In recent years it has become more common for researchers to publish their research data as open data. Funding agencies, both at the European and national level, increasingly require the research data (and publications) resulting from funded research projects to be published open access. However, open access data publishing is not yet standard practice in most disciplines. In the arts and humanities in particular, there is no culture of data sharing and reusing among researchers. Even when researchers in these fields publish their data in the European repositories and archives, the data is usually difficult to find and to access.
With the project Humanities at Scale (HaS), DARIAH-EU[1] aims to connect with the open access movement in the European Union. At the moment, researchers funded by Horizon 2020 in nine dedicated research areas are obliged to publish their research data as open data. From 2017 onwards this will be the case for the researchers in all disciplinary areas[2].

Examples of European projects and organisations that promote open access of publications and data are OpenAIRE 2020[3], OAPEN[4], Knowledge Exchange[5], Open Data Institute[6] and the Open Knowledge Foundation[7].

Work package (WP) 7 'Open Data Infrastructure' of the HaS-DARIAH project will intensify the collaborations with these open access initiatives and will support the implementation of corresponding services within the arts and humanities. With this WP, HaS aims to develop an open humanities data platform to make information about existing data collections and open research data more accessible. WP 7 contains three tasks:

- Task 7.1. State-of-the-art open access research data for the humanities
- Task 7.2. Towards a sustainable Open Data platform for humanities
- Task 7.3. DARIAH open humanities platform

Task 7.1 is a joint effort of Data Archiving and Networked Services (DANS), Centre National de la Recherche Scientifique (CNRS) and the University of Gottingen – State and University Library (UGOE-SUB), and the deliverable of the task is this report. Here, we present an overview of the key aspects of open access publishing of research data in the humanities. In Chapter 1, we start with an introduction where concepts such as research data, open data and metadata are defined. This is followed by a description of the

[1] www.dariah.eu

[2] See the press release: http://europa.eu/rapid/press-release_IP-16-1408_en.htm

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

[4] http://www.oapen.org/home

[5] http://www.knowledge-exchange.info/

[6] https://theodi.org/

[7] https://okfn.org/

research data life cycle in Chapter 2. Chapter 3 presents the different stakeholders involved in the publishing of open data, followed by Chapter 4 where we discuss the advantages and obstacles for researchers to share research data. Chapter 5 presents the different aspects of data citation. Chapter 6 gives an overview of the European data repositories, the certification standards of trusted digital data repositories, and the use of metadata. Finally, in Chapter 9 we highlight in detail the French open research data ecosystem as a current best practice example of an integrated open access research data system in the humanities. With regard to the allocated resources in the work package only the case of the French open research data ecosystem will be described in detail, whereas for other exemplary infrastructures chapter 5 will provide only an overview and a matrix as supplement to this report will be provided.
This is a state of the art report about open access publishing of research data and, in addition to the provision and guidance on open access publishing of research data to humanities researchers, is a desk study in preparation of HaS-DARIAH Task 7.2, which will deliver a design and sustainability plan for the DARIAH open humanities data platform. This platform will be implemented with Task 7.3.

1.1 Disciplinary scope

DARIAH is an acronym for the Digital Research Infrastructure for the Arts and Humanities. It is a European infrastructure for arts and humanities scholars working with computational methods. It supports digital research as well as the teaching of digital research methods. Humanities at Scale (HaS) is a project of DARIAH and focuses on fostering new and sustaining existing knowledge in digitally enabled research in the arts and humanities. The arts and humanities contain many disciplines[8], including classical studies, history, languages, law, performing arts, archaeology et cetera. Digital open data storage, analysis and publishing in the natural sciences, for example astronomy or human genetics, are done on a much larger scale than in the humanities (and social sciences). DARIAH aims to promote open data publishing and the reuse of research data in the arts and humanities, and WP 7 of HaS DARIAH will support this aim.

1.2 Definitions

Before further discussion we present definitions of several key subjects that are central to open access data publishing: research data, open data, metadata as well as data archives.

Research data

Horizon 2020[9], the EU Framework Programme for Research and Innovation, uses the following definition: 'Research data refers to information, in particular facts or numbers,

[8] https://en.wikipedia.org/wiki/Humanities

[9] https://ec.europa.eu/programmes/horizon2020/

collected to be examined and considered and as a basis for reasoning, discussion, or calculation. In a research context, examples of data include statistics, results of experiments, measurements, observations resulting from fieldwork, survey results, interview recordings and images'.

Open data

Following Horizon 2020 open data is data that is free to access, mine, exploit, reproduce, disseminate and reuse[10]. The researcher can make the research data openly accessible in a repository or archive by using explicit licenses or waivers, such as the Creative Commons Licenses CC-BY, CC-BY-SA or CC0.

Information about licenses can be found at:

- https://wiki.creativecommons.org/wiki/Data_and_CC_licenses
- http://opendefinition.org/guide/data/
- http://opendefinition.org/licenses/

Metadata

In addition to depositing a dataset in a repository, the researcher should also give the appropriate information about the dataset, which is known as metadata. Metadata describes the dataset and makes it possible for others to find, understand, and reuse the data. Besides standard information such as the creator and contributors of the dataset, the title, year of publication, and access rights, it can be necessary to add documentation such as codebooks, lab journals, informed consent forms and used software. There are various metadata standards for different disciplines, describing a range of relevant additional information necessary for making specific types of datasets comprehendible to other users. For example, archaeological datasets require metadata about the spatial coverage area, while linguistics datasets require information about the language. Chapter 8 presents an in-depth analysis of metadata standards in the humanities.

Data archives and data repositories

According to the Science Europe Data Glossary[11] a data archive is 'a professional institution for the acquisition, preparation, preservation, and dissemination of research data'. A data repository is 'a place (data storage system, archive) that holds data sets, makes data sets available to use, and organizes them in a logical manner'. In practice, the terms *data archive* and *data repository* are often used interchangeably, including in this report.

[10] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot- guide_en.pdf

[11] http://sedataglossary.shoutwiki.com/wiki/Main_Page

2. Research data lifecycle and data management plans (DMP)

The research data lifecycle represents all of the stages of data throughout its life from its creation for a study to its distribution and reuse[12]. Research data exist longer than the research project where they were created. Researchers may improve the access to data at every phase of the life cycle and in new research projects the data may be reused by other researchers.

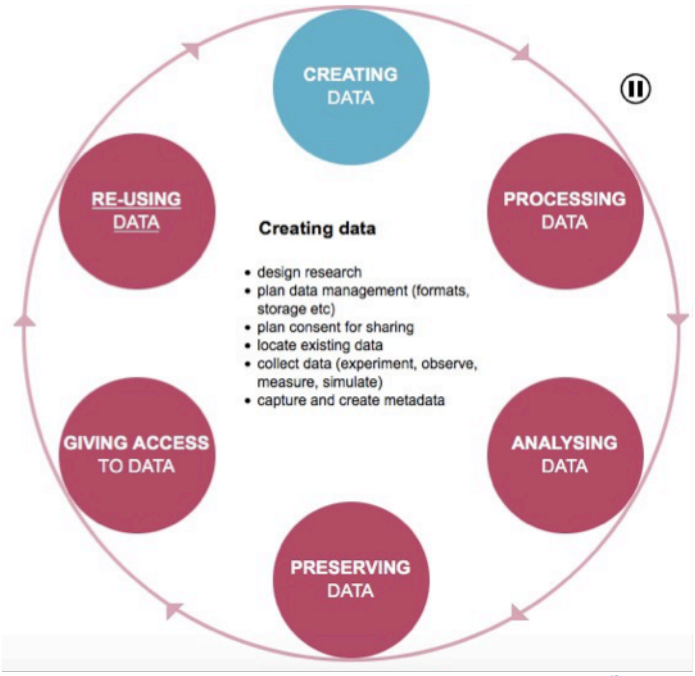

Figure 1: Research data lifecyce - (c) UK Data Archive[13]

'The data lifecycle begins with a researcher(s) developing a concept for a study; once a study concept is developed, data is then collected for that study. After data is collected, it is processed for distribution so that it can be archived and used by other researchers at a later date. Once data reaches the distribution stage of the lifecycle, it is stored in a location (i.e. repository, data archive) where it can then be discovered by other researchers. Data discovery leads to the repurposing of data, which creates a continual loop back to the data processing stage where the repurposed data is archived and distributed for discovery'.

The UK Data Archive gives an elaborate description of the research data cycle[13]. An important part of the research data cycle is the research data management plan (DMP). This is a formal document where the researcher, from the beginning of the research, describes what data and associated metadata and tools will be used, delivered and

[12] See the Science Europe Data Glossary: http://sedataglossary.shoutwiki.com/wiki/Research_data_lifecycle

[13] This research data lifecycle is retrieved from http://www.data-archive.ac.uk/create-manage/life-cycle

possibly shared both during and after the research. A DMP should be a 'living' document which can be adapted as necessary during the course of the research. Research funders increasingly require data management plans for research grants. Different templates for DMPs, also for the requirements of Horizon 2020, can be found in the DCC's DMPonline tool[14].

Also at the website of DANS (Data Archiving and Networked Services) in the Netherlands, an institute for data archiving with a strong focus on the humanities and social sciences, the researcher can find a DMP template[15]. This DMP consists of the following sections: administrative information (project title, principal researcher, funder(s) etc.); description of the data (existing data reused, new data generated, type(s) of data is concerned; file size etc.); standards and metadata (metadata standards, coding, software and hardware etc.); ethical and legal aspects (sensitive data, open data etc.); storage and archiving (storage and backup capacity during the project; storage after the project, expenses etc.).

[14] https://dmponline.dcc.ac.uk/

[15] http://www.dans.knaw.nl/en/about/organisation-and-policy/information-material

3. Stakeholders in the humanities open access research data landscape

The open access data landscape in the humanities includes the following main stakeholders:

- ***Individual researchers*** and ***research institutions*** are at the core of open access data publishing and use: they are the main data producers as well as consumers of digital research data. As data sharers, they need to trust that their data is preserved, accessible, and useable for the long term. As data users, the main concerns are the ability to find the data, and the authenticity and quality of the data.

- ***Archives, museums***, and other ***cultural heritage institutions*** (for example libraries) are important data providers in the humanities. Their main interests are to make their data available to the general public, and in the second place to researchers, as well as preserving the data for the future.

- ***Funding agencies*** benefit from promoting the optimal use and reuse of data in which funds were invested. They can do this by encouraging good data practices, investing in data infrastructure and raising data awareness.

- ***Digital repositories*** preserve and make data findable and useable for the long term, by e.g. using sustainable file formats, and providing persistent identifiers and informative descriptive data (metadata). Related to this are online data platforms that do not store data, but bring together metadata of research datasets, making them findable for data users.

- ***Academic publishers*** impose requirements on the availability of data connected to submitted and/or published papers, and provide identifiers to cite papers and link to related data. Non-academic publishers (for example societies) are also important in the humanities, however, for these the availability of data connected to publications is often less clear.

- ***The general public*** can access source data, research findings and educational tools through open access of data in the humanities. This also applies to educators and teachers interested in humanities, as well as NGOs and humanitarian organisations. The public is also increasingly involved in producing data through participation in citizen science.

4. The advantages and obstacles for researchers to share research data

4.1 Recent developments in data sharing

In the past few years the recognition of the value of (big) data, data sharing and proper research data management has sharply increased. The influential report 'Riding the Wave: How Europe Can Gain From The Rising Tide of Scientific Data'[16] (2010) of the EU High Level Expert Group on Scientific Data advocates a collaborative data infrastructure where researchers and other stakeholders could re-use the data.

Neelie Kroes, Vice-President of the European Commission responsible for the Digital Agenda, stated in her Opening Remarks at a Press Conference on Open Data Strategy in Brussels, 12th December 2011, that '***Data is the new gold***'[17]. One of the goals of the FP7 programme and particularly of Horizon 2020 is to achieve 'open data'. In the Open Research Data Pilot[18] of Horizon 2020 researchers working in nine dedicated research areas have to publish their research data as open data. It is possible to opt out, and it is also possible to opt into the pilot. From 2017 this applies to research data in all disciplines; opting out is still possible.

At the EU Open Science Conference in Amsterdam in April 2016, Carlos Moedas, the EU Commissioner for Research and Innovation, announced that the EU aims to build a European Open Science Cloud. This initiative is planned to be operational in 2020, when researchers will be able to store, share, and re-use data across disciplines and geographical boundaries[19].

These developments at the EU level are reflected at the national level of the of member states. Open access and data sharing is an important topic for national policy makers and funding organisations. For example, the government of the Netherlands aims to make 60% of all Dutch scientific publications open access by 2019, and a 100% by 2024[20]. National research funders increasingly require the inclusion of research data management plans in grant proposals or granted projects, and ask for research data resulting from funded projects to be made openly accessible.

[16] http://ec.europa.eu/information_society/newsroom/cf/itemlongdetail.cfm?item_id=6204

[17] https://ec.europa.eu/digital-single-market/en/news/data-new-gold

[18] https://www.openaire.eu/opendatapilot

[19] http://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

[20] http://www.openaccess.nl/en/in-the-netherlands/what-does-the-government-want

4.2 Advantages of data sharing

Reusing data

Researchers can use the data of previous research by others, for example for comparative research or meta-analyses. Open research data can also be used for validating the results of earlier research or for secondary analysis, i.e. answering new research questions with an existing dataset.

Looking at for example the use of EASY[21], the data archive at DANS in the Netherlands, we can see a clear increase of usage in the last years. As of May 2016 EASY contains nearly 32,000 datasets, mainly in the humanities (31.000) and social sciences (4.400)[22]. In 2015, about 30,989 datasets (180,000 files) were downloaded and could be reused. The data reviews section of the archive shows that users highly appreciate the reuse of data[23]

Using data in national and international cooperation

The number of national and international data research projects and infrastructures has grown substantially over the last five years. In these projects data of many scholars and projects are brought together and are connected, enabling new and wider collaborations and analyses. Recent examples include the European Holocaust Research Infrastructure (EHRI) [24] for holocaust studies, or the Advanced Research Infrastructure for Archaeological Dataset Networking (ARIADNE)[25], for the integration of archaeological datasets.

Economic benefit

Sharing data in national and international collaborative projects increases the return on investment because researchers do not have to collect all the data themselves. This point has been argued on many occasions and by many organizations, ranging from the OECD to the European Commission. The publication of Capgemini Consulting 'The Open Data Economy Unlocking Economic Value by Opening Government and Public Data'[26] describes the economic benefit for the government and the private sector. For the government the economic benefit is, among other things, increased tax revenues through increased economic activity, and increased service efficiency through linked data. For the private sector the benefits include reduced cost by not having to invest in conversion of raw government data, and better decision-making based on accurate information.

[21] https://easy.dans.knaw.nl/ui/browse

[22] A dataset can belong to more disciplines.

[23] http://datareviews.dans.knaw.nl/index.php

[24] http://www.ehri-project.eu/

[25] http://www.ariadne-infrastructure.eu/

[26] http://www.capgemini-consulting.com/resource-file-access/resource/pdf/opendata_pov_6feb.pdf

Academic integrity

With open access to data, published research results become verifiable, which in turn promotes scientific integrity. M. Bakker's PhD thesis 'Good science, bad science: Questioning research practices in psychological research (University of Amsterdam 2014)' shows that data sharing stimulates the transparency of science and therewith reduces the number of errors[27]. In recent years the focus on the issue of open data and scientific integrity is increasing. On this topic in the Netherlands the advisory report 'Responsible research data management and the prevention of scientific misconduct' of professor Kees Schuyt was published in 2012 (in 2013 the English version).[28]

4.3 The main obstacles for data sharing

Unwillingness of researchers to share their data

There are several obstacles for sharing research data. One of them is the unwillingness, for several reasons, of researchers to share their data. In the publication 'The Dutch data landscape in 32 interviews and a survey'[29] (2011), DANS presents a brief overview of what Dutch researchers think of sharing data in their field. One of the reasons for not sharing is the use of strictly confidential personal data, such as criminal records. Another group of researchers is simply convinced that others have no right to access their data. A third impediment is that a number of researchers are afraid that others will use their data to publish sooner. Another reason is that researchers are afraid their colleagues will misunderstand or misinterpret their data. A fifth reason is that the researchers see their research as 'frontier research', where one cannot re-use the data.

	Behavioural sciences	Humanities	Socio-cultural sciences	Social sciences	Total
Others benefit	25	16	35	33	27
Little in return	64	28	39	33	42
Privacy considerations	36	20	30	67	38
Possible misuse	43	48	26	46	41
Technical obstacles	11	4	-	-	4
Other	21	40	30	25	29

Figure 2: Percentage of those working in humanities and social sciences quoting objections to sharing their own data (n=100)

[27] http://dare.uva.nl/record/472604

[28] https://www.knaw.nl/en/news/publications/responsible-research-data-management-and-the-prevention- of-scientific-misconduct?set_language=en

[29] http://www.dans.knaw.nl/en/about/organisation-and-policy/publications

EU General Data Protection Regulation (GDPR)

As of 14 April 2016 the new EU General Data Protection Regulation (GDPR) replaces the former Data Protection Directive, which dated back to 1995. The GDPR may offer new impediments for sharing personal data. The main worry is that the GDPR will be so strict that it will restrain research on individual data, and that it will restrict data sharing. This is especially true for the biomedical and social sciences where the regulation of privacy-sensitive data from subjects or patients is much stricter. The GDPR will be directly applicable in all member states in two years. In the meantime, the EU countries should develop additional arrangements at the national level, such as the clarification of the *informed consent*, which is important for scientific research that involves personal data.

Lack of funding for long-term data preservation and data sharing

A third major obstacle for data sharing is that funders, in particular the National Research Councils, do not reserve money for long-term data preservation and data sharing when they subsidize research. If the funders demand depositing the data in a trusted data repository[30] and make data storage and sharing eligible for funding, this would give time, money, and the obligation to the researchers to do so.

No credits or other rewards for producing and sharing data

Last but not least, researchers do not get sufficient credit and other rewards for producing and sharing data. It is still the publication in a peer-reviewed journal that counts. Making data sets available should also be rewarded as an important scientific output. Journals adopting a data availability policy and data journals can be an important instrument to change this situation.

[30] http://www.datasealofapproval.org

5. Landscape of open access research data infrastructure

5.1 Historical development data archives / repositories

The first data archives with special-purpose repositories in Europe emerged around 50 years ago. These were data archives in the social sciences, for example the Dutch Steinmetz Archive (1964) and the British UK Data Archive (1967). In the seventies the text archives for linguistics and literary studies (e.g. Oxford Text Archive) arose, followed by historical archives in the eighties and nineties (e.g. Dutch Historic Data Archive (NHDA) and the British History Data Service (HDS)[31]). Archaeological data archives date back from the beginning of this century (e.g. the British Archaeology Data Services (ADS)[32] and the Dutch e-depot for Dutch Archaeology (EDNA)[33]. The advantage for the researcher of these special-purpose repositories is that the data can be preserved according to the recognized standards in the discipline of the researchers.

General-purpose data repositories have mostly emerged in the last decade. Universities started to develop data repositories, mostly combined with scholarly publications. Later researchers could submit their research data in general data sharing repositories, such as Zenodo[58], Figshare[52], Mendeley Data[34], DRYAD[54], Dataverse[55], and the EUDAT B2-tools[35]. These repositories contain all kinds of data types, from very different disciplines.

5.2 Number of repositories

Via the Registry of Research Data Repositories (Re3Data.org)[36], based on self-registry, one can find over 1,400 research data repositories. In the field of humanities there are 123 data repositories[37]. From these repositories in the humanities there are 39 based outside the EU, in countries like the US (26) and Australia (9). In Europe there are 94 data repositories, with the most in Germany (31), the United Kingdom (15), France (6) and the Netherlands (5). There are also collaborations like European repositories (e.g. Zenodo), or between two or more countries (e.g. Germany with France or with the Netherlands).

[31] http://hds.essex.ac.uk/

[32] http://archaeologydataservice.ac.uk/

[33] http://www.dans.knaw.nl/en/about/services/archiving-and-reusing-data/easy/edna?set_language=en

[34] https://data.mendeley.com/

[35] https://www.eudat.eu/

[36] Re3Data.org

[37] http://service.re3data.org/search?subjects[]=11 Humanities

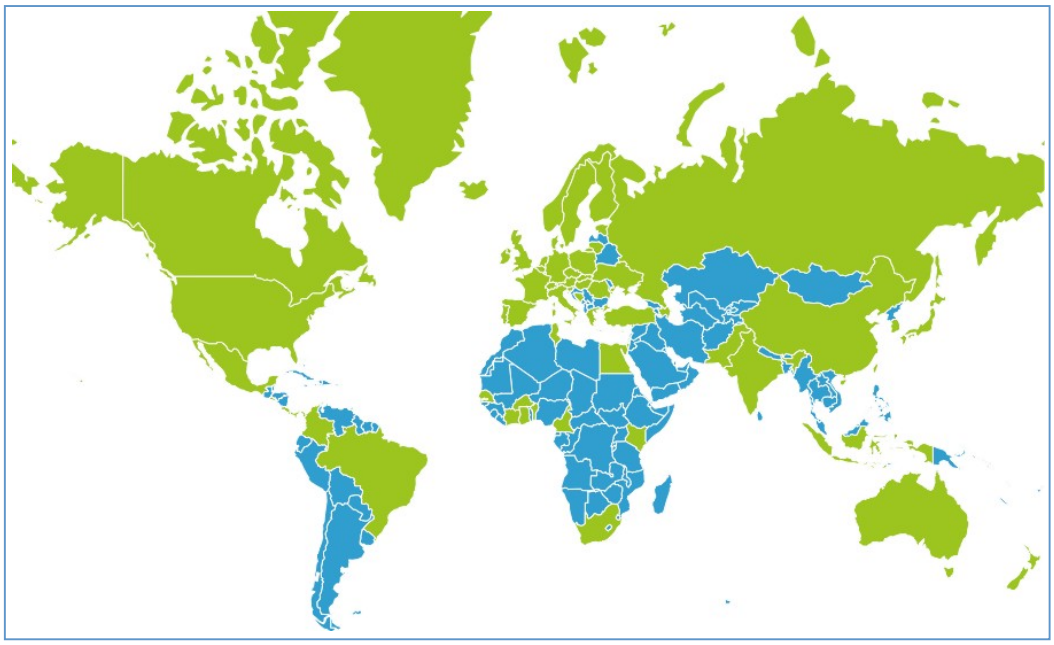

Figure 3: Countries with data repositories (in green) according to Registry of Research Data Repositories

5.3 Description of open international data repositories

When we take a look at the open, international data repositories we can distinguish: Dataverse, Dryad, EUDAT, FigShare, Mendeley Data, and Zenodo (descriptions of the repositories are retrieved from the Registry of Research Data Repositories, except for Mendeley Data):

Dataverse

Dataverse has been developed by Harvard University. 'The Harvard Dataverse is open to all scientific data from all disciplines worldwide. It includes the world's largest collection of social science research data. It is hosting data for projects, archives, researchers, journals, organizations, and institutions.' There are many communities that work together in platforms of Dataverse, for example the Dutch DataverseNL, a cooperation of nine institutions using the Dataverse platform.

Dryad

DataDryad.org is a curated general-purpose repository that makes the data underlying scientific publications discoverable, freely reusable, and citable. Dryad is an international repository of data underlying peer-reviewed scientific and medical literature, particularly data for which no specialized repository exists. The content is considered to be integral to the published research. All material in Dryad is associated with a scholarly publication.

EUDAT

The EUDAT project aims to contribute to the production of a Collaborative Data Infrastructure (CDI). The project´s target is to provide a pan-European solution to the challenge of data proliferation in Europe's scientific and research communities. The EUDAT vision is to support a Collaborative Data Infrastructure which will allow researchers to share data within and between communities and enable them to carry out their research effectively. EUDAT aims to provide a solution that will be affordable, trustworthy, robust, persistent and easy to use. EUDAT comprises 26 European partners, including data centres, technology providers, research communities and funding agencies from 13 countries. B2FIND is the EUDAT metadata service allowing users to discover what kind of data is stored through the B2SAFE and B2SHARE services which collect a large number of datasets from various disciplines. EUDAT will also harvest metadata from communities that have stable metadata providers to create a comprehensive joint catalogue to help researchers find interesting data objects and collections.'

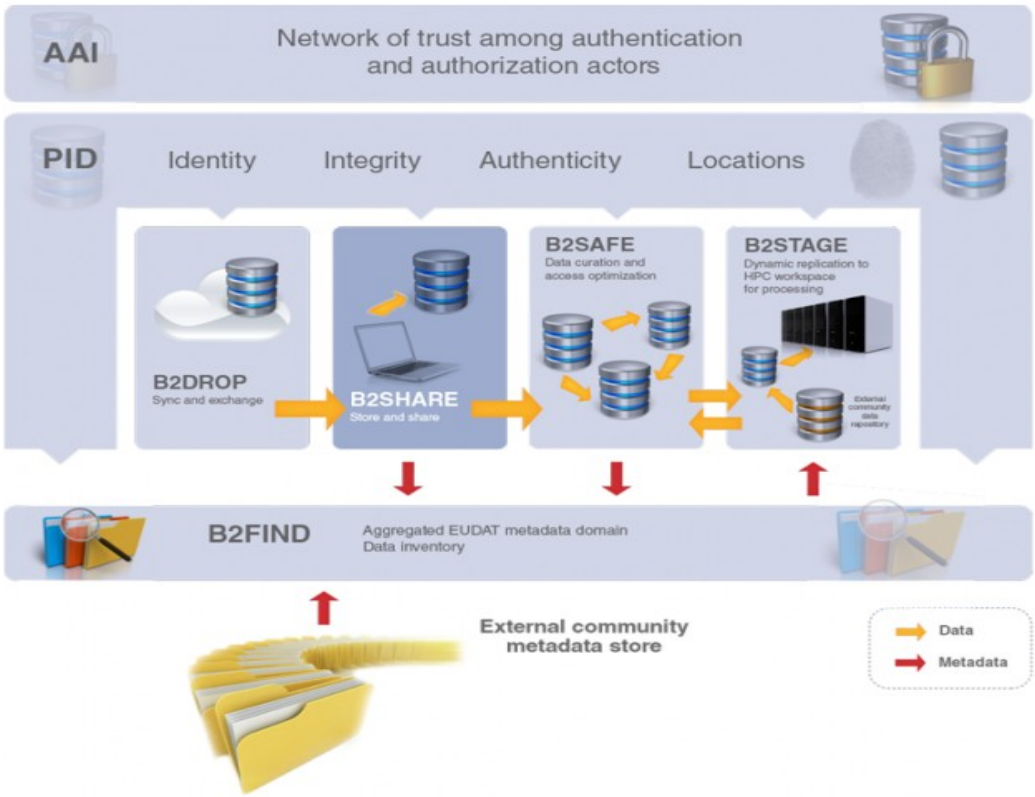

Figure 4: The core services of the B2 Service Suit

Figshare

'Figshare allows researchers to publish all of their research outputs in an easily citable, sharable and discoverable manner. All file formats can be published, including videos and datasets. Optional peer review process. Figshare uses creative commons licensing.' Figshare also contains research data in humanities.

Mendeley Data

'The platform allows researchers to upload the raw data from their research, and give it a unique identifier (a versioned DOI), making that research citable. For partnering journal websites, the article links to the research dataset on Mendeley Data, enabling readers to quickly drill down from a research article to the underlying data; while the dataset also links to the article. Researchers can also privately share their unpublished data with collaborators, and make available multiple versions of the data relating to a single research project, creating an evolving body of data.'[38]

Zenodo

'Zenodo builds and operates a simple and innovative service that enables researchers, scientists, EU projects and institutions to share and showcase multidisciplinary research results (data and publications) that are not part of the existing institutional or subject-based repositories of the research communities. Zenodo enables researchers, scientists, EU projects and institutions to: a) easily share the long tail of small research results in a wide variety of formats including text, spreadsheets, audio, video, and images across all fields of science. b) display their research results and get credited by making the research results citable and integrate them into existing reporting lines to funding agencies like the European Commission. c) easily access and reuse shared research results.'

Zenodo is a repository that is being harvested by OpenAIRE 2020[3]. OpenAIRE is a project of Horizon 2020 to promote the Open Access policy of the European Commission. OpenAIRE advises researchers to use Zenodo as repository for their research data (and publications), in case they do not have access to institutional or disciplinary repositories.

5.4 A comparison of infrastructures for publishing research data

In Annex 1[39], a comparison of a number of aspects of the following repositories is given: B2Share, Dataverse, Figshare, Dryad, Zenodo and Mendeley. These are all aspects that may be of particular interest to researchers. In this document no value judgments are given; the content of this document is for information purposes only.

[38] Text retrieved from: https://blog.mendeley.com/2015/11/09/put-your-research-data-online-with-mendeley- data/

[39] This 'Comparison of infrastructures for publishing research data' is an update and English translation of the Dutch document 'Vergelijking infrastructuren voor het publiceren van onderzoeksdata' written by Tessa Pronk and Kees van Eijden, 21-05-2014. Disclaimer: The authors accept no liability for any inaccuracies in the information provided. Using this comparative document is at the researcher's own risk.

These repositories, described above, are suitable for the entire research community and provide persistent identifiers to enable sustainable data reference.

5.5 Which repository to choose by the researchers?

Figure 5 shows the steps that are recommended by the H2020 OpenAIRE project for selecting a research data repository[40].

Figure 5: Steps to find a data repository

[40] See the OpenAIRE RDM briefing paper: https://www.openaire.eu/dissemination-material/briefpaper-rdm- infonoads

6. Certification standards of digital repositories

6.1 Certification of Trusted Digital Repositories

Trust is a key aspect of the relationship between repositories and their stakeholders. Data depositors, curators, consumers and funders each have expectations and requirements of repositories regarding the storage, preservation and dissemination of information. Data depositors need to be able to trust that their data is safe in a repository, and that the data will be accessible, useable, and readable on a long-term basis. For data consumers, relevant questions include whether the data are stored properly, whether the authenticity and integrity of the data are guaranteed, whether the data is of good quality, and whether the identifiers refer to the correct objects. The main issue for funding agencies is the optimal use and reuse of data in which funds were invested, and the long-term availability for reuse of these data.

In order to formally ensure the performance of digital repositories in all these aspects, several systems for audit and certification have been established. Certification on the basis of accepted assessment frameworks can be a significant factor in warranting the trustworthiness and sustainability of digital archives, which in turn promotes a culture of sharing data. For these reasons, there is increasing emphasis on certification of trusted digital repositories in the current digital data landscape.

6.2 Overview of standards and certification initiatives

Certification methods for digital preservation infrastructures have been in development for over a decade, with different organizations developing several procedures in parallel. The different initiatives were largely based on the principles, terminology and functional characteristics described in the 2002 Reference Model for an Open Archival Information System (OAIS), published by the Consultative Committee for Space Data Systems (CCSDS). One of the first to define the characteristics of a Trusted Digital Repository (TDR) was the 2002 Research Libraries Group (RLG) and the Online Computer Library Centre (OCLC) Working Group of Digital Archive Attributes[41]. This resulted in the publication entitled Trusted Digital Repositories: Attributes and responsibilities (2002)[42], which later resulted in TRAC: Trustworthy Repositories Audit & Certification: Criteria and Checklist (2007). Several other lists of criteria and certification standards were developed over the years, including:

- DRAMBORA: Digital Repository Audit Method Based on Risk Assessment (DCC & DPE, 2007)[43]

[41] http://www.crl.edu/archivingpreservation/digital-archives/metrics-assessing-and-certifying-0

[42] http://www.oclc.org/content/dam/research/activities/trustedrep/repositories.pdf

[43] http://www.repositoryaudit.eu

- Data Seal of Approval. Quality Guidelines for Digital Research Data (2009, 2013)[44]
- Audit and Certification of Trustworthy Digital Repositories. Recommended Practice (CCSDS, 2011)[45]
- ICSU World Data System certification (2012)
- ISO 16363: Audit and certification of trustworthy digital repositories (2012)[46]
- NESTOR Seal (2016)[47] [48]based on the DIN 31644: Criteria for trustworthy digital archives (2012)[49]

Alignment of DSA and WDS certification

In recent years, the Data Seal of Approval (DSA) and the ICSU World Data System (ICSU-WDS) are collaborating to harmonise their certification requirements and procedures, and to set the stage for a global shared framework including other standards. The DSA-WDS Partnership Working Group of Repository Audit and Certification, a working group of the Research Data Alliance, was set up to develop Catalogues of Common Requirements and Procedures, based on the criteria that were already in put in place by the DSA and ICSU-WDS. In the course of 2016, both DSA and ICSU-WDS will replace their current certification criteria with the new shared Catalogue of Common Requirements.

6.3 European Framework for Certification

In 2010, the European Framework for Audit and Certification of Digital Repositories[50] was set up by three groups working on standards for TDRs: Data Seal of Approval, the Repository Audit and Certification Working Group of the CCSDS, and the DIN Working Group 'Trustworthy Archives – Certification'. The framework consists of three levels of certification that offer increasing trustworthiness:

1. Basic Certification is granted to repositories that obtain Data Seal of Approval (DSA) certification. DSA, initially developed by DANS, ensures that research data can be processed in a high-quality, reliable manner, provided the 16 guidelines for self- assessment are followed.
2. Extended Certification is granted to Basic Certification repositories that perform an additional structured, externally reviewed and publicly available self-audit based on ISO 16363 or DIN 31644. The DIN 31644 standard, an initiative by NESTOR, is a catalogue of 34 criteria that trusted digital repositories should meet. The ISO 16363 standard presents over 100 metrics for different aspects of a digital repository.

[44] http://datasealofapproval.org

[45] http://public.ccsds.org/publications/archive/652x0m1.pdf

[46] http://www.iso.org/iso/catalogue_detail.htm?csnumber=56510

[47] http://files.dnb.de/nestor/materialien/nestor_mat_17_eng.pdf

[48] http://files.dnb.de/nestor/materialien/nestor_mat_08_eng.pdf

[49] http://www.nabd.din.de/cmd?level=tplart-

[50] www.trustedrepository.eu

3. Formal Certification is granted to repositories which, in addition to Basic Certification, pass a full external audit and certification based on DIN 31644 or ISO 16363.

Granting of these certificates will allow repositories to show one of three symbols (to be agreed) on their web pages and other documentation, in addition to any other DSA, DIN or ISO certification marks (Figure 6).

Figure 6: The certification marks of the Data Seal of Approval, the nestor Seal, and ISO

7. Open access data publication and data citation

This chapter focuses on the publication of research data in the open access publication sector, including the issue of data citation and its relationship with primary sources, in the humanities and social sciences.

7.1 Introduction

Publishing convention inherited in Social Sciences and Humanities pushes scientists to pursue an objective of perfection and completeness before releasing monographs and books, but the increased volume of data produced and new governance model are calling for some evolution (Maxwell 2015). Usually, research data were assimilated to some adjunct material, found in annex, appendix, separate volume, even partial or absent; Prost and al. (2015) use the expression 'dark data' to describe the part of 'small' data produced and partly published. For example, in ethnography, fieldwork generates a large amount of data that helps to build a deep contextual knowledge but that is not directly embedded in publication. Primary data is not published and so the provision of this material escapes accessibility. But scholars are now re-investigating raw data, finding new perspectives, connecting interpretations and shifting paradigms (Roorda 2014). In this context, relevance is no more set by publication in prestigious journals, it is built also as the dynamic of the social and scientific 'ongoing discourse' (Maxwell 2015).

As publication is not anymore limited to the print sphere, value can shift from compactness to completeness (Mooney and Newton 2012). Following this assertion, shared data does not diminish in value because it can be used by other and can contribute to further methods or analysis.

The disjunction of research data and research claims is a profound modification from the monographic perspective of publication. Prost and al. (2015) explain that when data is submitted as a kind of appendix, the dissertation becomes a 'data vehicle', where data are published together with the dissertation or as part of it. When data is available on a server, without the dissertation, it transforms the dissertation into a 'gateway to data' and afford an 'agile publishing' based on earlier and more frequent releases, opened to review, commentary, new editions and supplementary material (cf. Armstrong, 2010; Maxwell & Fraser, 2011; Raccah, 2012; Omas & Hunt, 2010; Maxwell, 2015).

7.2 Barriers to open data citation

Numerous issues have been mentioned in the literature on the topic. We decided to highlight some of them:

Data-related barriers:

- Diversity of formats: compared to conventional scientific publications (e.g. journal articles or monographs), research data comes with a considerable wider range of

data formats, metadata schemas and types of content. This makes it difficult for citation and referencing standards as they usually rest upon standardised infrastructure services. But the currently very broad range of research data formats demands for diverse infrastructure services.

- Fragmentation: incomplete, inadequate, or even missing description of data sets or individual data.
- Unconsidered use cases: Missing organisation, making research data not suitable for further reuse because of a lack of structure or organisation.

Technical barriers: When data is not clearly separated from the dissertation, or when it is glue together in a pdf file instead of being properly published in adequate file format (spreadsheet, image, text, database, etc. (Prost and al. 2015).

Legal barriers:

- Privacy issues, including research data from surveys, experiments, interviews or biographies, that include personal information allowing identification of the respondent.
- Third party copyright: some academic publications, for instance PhD dissertations, might include copyrighted elements (maps, photographs, text samples, etc.) that cannot be reproduced and disseminated without authorization, even by fair use or copyright exception.

Other barriers: We can also express here some concerns about the issues of funding and sustainability, provenance, identity and attribution, versioning, granularity (Altam and Crosas 2013), trust and resistance (Cliggett 2013), curation and status of data (Mayo and al. 2015), or the variety of licences, the widespread adoption of 'non-commercial' licences (Moore 2014); and more generally the great variation of practices and standards between and within disciplines (Austin and al. 2015; Cliggett 2013; Sperberg-McQueen in Uhlir 2012).

7.3 Publishing and circulation of credit

During the past few years, there has a been a notable rise of awareness regarding the data citation issue. The Joint Declaration of Data Citation Principle[51] from the Data Citation Synthesis Group 2014 (JDDCP) presents eight framing principles to make data transparently available for verification and reproducibility. As stated in the declaration, 'Data citation, like the citation of other evidence and sources, is good research practice and is part of the scholarly ecosystem supporting data reuse.':

[51] https://www.force11.org/group/joint-declaration-data-citation-principles-final

1) ***Importance***: 'Data should be considered legitimate, citable products of research. Data citations should be accorded the same importance in the scholarly record as citations of other research objects, such as publications'.

2) *Credit and Attribution*: 'Data citations should facilitate giving scholarly credit and normative and legal attribution to all contributors to the data, recognizing that a single style or mechanism of attribution may not be applicable to all data.'

3) 'In scholarly literature, ***whenever and wherever a claim relies upon data, the corresponding data should be cited'***.

4) ***Unique Identification***: 'A data citation should include a persistent method for identification that is machine actionable, globally unique, and widely used by a community.'

5) ***Access***: 'Data citations should facilitate access to the data themselves and to such associated metadata, documentation, code, and other materials, as are necessary for both humans and machines to make informed use of the referenced data.'

6) ***Persistence***: 'Unique identifiers, and metadata describing the data, and its disposition, should persist -- even beyond the lifespan of the data they describe.'

7) ***Specificity and Verifiability***: 'Data citations should facilitate identification of, access to, and verification of the specific data that support a claim. Citations or citation metadata should include information about provenance and fixity sufficient to facilitate verifying that the specific time slice, version and/or granular portion of data retrieved subsequently is the same as was originally cited.'

8) ***Interoperability and Flexibility***: 'Data citation methods should be sufficiently flexible to accommodate the variant practices among communities, but should not differ so much that they compromise interoperability of data citation practices across communities.'

7.4 Inspiring initiatives

We can distinguish between different kinds of innovative actors in the field:

Repositories and other services offering citable identifiers

- ***Figshare***[52] provides a Digital Object Identifier (DOI) for a submitted dataset so it can be cited like a usual publication. In this regard, Thomson Reuters' Data

[52] http://figshare.com/

Citation Index[53] has to be mentioned as it could make a real contribution toward a cultural shift for establishing data publishing practice in scientific communities.

- ***Dryad***[54] *Digital Repository* is a curated resource that makes the data underlying scientific publications discoverable, freely reusable, and citable. It provides a general-purpose home for a wide diversity of data types.
- ***Dataverse***[55] is an open source web application to share, preserve, cite, explore and analyse research data. It facilitates making data available to others and allows to replicate others' work. Researchers, data authors, publishers, data distributors, and affiliated institutions all receive appropriate credit. A Dataverse repository hosts multiple dataverses. Each dataverse contains dataset or other dataverses, and each dataset contains descriptive metadata and data files (including documentation and code that accompany the data). It is worth to mention that Dataverse is now integrated with Open Journal System[56].
- ***Datacite***[57] is another repository with a special focus on developing and supporting methods to locate, identify and cite data and other research objects using the standards behind persistent identifiers for data.
- ***Zenodo***[58] is a Github integrated repository that enable researchers to share and preserve any research outputs in any size, any format and from any science.
- ***CrossCite***[59] is a new project trying build bridges between CrossRef and DataCite as a cross-platform citation service and DOI resolver.

7.5 Policies (journals)

An increasing number of publishers and journals, mainly from the natural sciences, address the issue of data publication and citation in their style guides: American Meteorological Society, American Sociological Association (Machine Readable Data Files; References for data sets must include a persistent identifier for future access, as assigned by digital archives), University of Chicago Press (Scientific databases), the Council of Science Editors (Databases on the Internet), National Library of Medicine (Part of a Database on the Internet), Nature Scientific Data, GigaScience (Biomed Central), F1000Research, Geoscience Data Journal (Wiley), etc.
Although enforcement of data publication is an effective strategy, (Fuchs et al 2012, Moore 2014), recent examples demonstrate that other incentives can be fruitful too. For instance the journal Psychological Science[60] uses badges to 'acknowledge Open

[53] http://wokinfo.com/products_tools/multidisciplinary/dci/
[54] http://www.datadryad.org/
[55] http://dataverse.org/
[56] http://projects.iq.harvard.edu/ojs-dvn/home
[57] https://www.datacite.org/node
[58] http://zenodo.org/
[59] http://www.crosscite.org/
[60] http://pss.sagepub.com/

Practices' provided by the Center for Open Science. Another asset of providing DOIs to datasets, beyond enabling accurate citation, is to give the possibility to track reuse. It is reasonable to assume that researcher would be more likely to deposit data in a repository if they can gain academic credit through a data journal (Callaghan et al. 2009; Harley, 2010; Moore 2014).
In this perspective, data journals comprise an important element of the publishing ecosystem as they can provide peer-review of data sets, shorten the delay of publication, and give attribution and credit to data managers and contributors who might not be involved in the analysis of a study, and therefore would not be eligible for author credit on an analysis paper. This would also assure the discoverability and understanding of data (quality and provenance) (Moore 2014). We can notice the interesting Science.ai[61] project that aims at simplifying the publishing pipeline and reducing production costs by more than 75% with an end-to-end solution native to the web. It is also worth mentioning the interest in research data and used software, in the journals from Ubiquity Press[62]: Journal of Open Archeology Data, Journal of Open Psychology Data, Journal of Open Research Software, Open Health Data or Open Journal of Bioresources.

7.6 Data Editorialization (OECD)

OECD[63] publication department is genuinely tackling the issue of hosting and making citable data upon which analysis and publication are based. It is a good example of data editorialization and of the integration of data with publications, specifically in open access.

Third-Party Providers, like ***data aggregators***, are also a key component of the ecosystem because they often work beyond the original data provider's subject or institutional focus, and some data providers enrich their metadata (e.g. with data-publication links, keywords or more granular subject matter) to enable better cross-disciplinary retrieval (Austin and al. 2015). In this regard, Huma-Num data publication service Nakala[64], coupled with its search engine Isidore[65], attributing handle identifiers to datasets, is a good example of innovative third party services providing the missing link between open research data and publications in humanities.

7.7 Recommendations

Based upon current literature, we can propose the following recommendations:

[61] https://science.ai/
[62] http://www.ubiquitypress.com/site/publish/#metajournals
[63] https://data.oecd.org/
[64] https://www.nakala.fr/
[65] http://www.rechercheisidore.fr/

- ***Data curation***: Some data cannot be reused only because of legal reason (Prost and al. 2015). ***Training and assistance*** must to be provided to researchers in adequate means to achieve a clear separation of text and data, submitted differently in open and ideally non proprietary format, in separate and organised data set, and including metadata of good quality.
- ***Flexible approach***: Commitment to open access, disciplinary sensitive approach and collaboration between the continuum of stakeholders is crucial to build shared perspectives (Prost and al. (2015).
- ***Reward structures*** must be in place to encourage data publication and citation: a proper tool for scholarly acknowledgment.
- ***Promotion***: Citation for data must be publicized as an essential component of science, accelerating and widening scientific research. Normative practice must emphasize on identification, retrieval, attribution of research data, and the possibility of restrictive application procedures.
- ***Reuse tracking via citation metrics***: data providers can ensure citation by including quality metadata, suggesting formatted citation and making data citation a compulsory condition for data reuse.
- ***Attribution of Persistent Identifiers*** to datasets and sub-links to each piece of a set to achieve a good level of granularity, and taking into account the *version* of referred objects, for example with CrossMark[66] or a Github inspired system.
- ***Peer review of data*** by researcher and by editorial review: Data journals, Data articles and Data reviews (metadata, integrity, discoverability, interoperability and indexation).
- ***Commitment to persistence:*** a resolving authority or domain owner has to be trustworthy in regards of its 'reasonable chance to be present and functional in the future (Starr and al. 2015).
- ***Tools:*** A great way of sensitization is the proliferation of tools, like Mendeley DOI attribution service[67] (Force11 compliant) or Zotero's integration with institutional repositories can play a major role in spreading new practices and technical knowledge from a user experience point of view.

As a final remark, we can use Silvello's (2015) *requirements that a data citation methodology must fulfill*: (a) uniquely identify the cited objects; (b) provide descriptive metadata; (c) enable variable granularity citations (dataset as a whole, a single unit, or a subset); and (d) produce both human- and machine-readable references.

7.8 Conclusion

While significant concerns are expressed about reproducibility and false positives being reported as fact (Colquhoun, 2014; Rekdal, 2014; Begley & Ellis, 2012; Prinz, Schlange &

[66] http://www.crossref.org/crossmark/

[67] https://data.mendeley.com/

Asadullah, 2011; Greenberg, 2009; Ioannidis, 2005), research data should be treated as first class material, and should be archived, indexed and cited just like textual publications (CODATA-ICSTI Task Group, 2013; Altman & King, 2006; Uhlir, 2012; Ball & Duke, 2012; Starr and al., 2015).
Conventions need to be established between all stakeholders but with a special attention to the risk of fragmentation between professional specialization. OECD's exemplar integration of its data and publication services is shedding light on a key issue: the skills of a data publisher that would be closely involved in the editorialisation of data sets is a real asset. In regard of the digital promises, new editorial forms have to be invented and implemented even if, at first sight, it may appear to be in contradiction with the conservatism of practices based on database, monographs and 'once and for all' writings. As we have seen, in order to overcome reluctance, better information, researcher training and accompaniment, qualified staff, innovative incentives and efficient reward systems need to be widely spread to help the production of 'publishable' and reusable data.

7.9 References

Altman, Micah, and Mercè Crosas. 'The Evolution of Data Citation: From Principles to Implementation.' IASSIST Quarterly, 37, 2013.

Austin, Claire C, Bloom, Theodora, Dallmeier-Tiessen, Sunje, Khodiyar, Varsha, Murphy, Fiona, Nurnberger, Amy, Raymond, Lisa, et al. 'Key Components of Data Publishing: Using Current Best Practices to Develop a Reference Model for Data Publishing,' 2015. doi:10.5281/zenodo.34542.

Cliggett, Lisa. 'The Qualitative Report 2013 Volume 18, How To Article 1, 1 - 11 http://www.nova.edu/ssss/QR/QR18/ cliggett1.pdf.' Accessed April 18, 2016.

Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014 [https://www.force11.org/group/joint-declaration-data-citation-principles-final]

Maxwell, John W. 'Beyond Open Access to Open Publication and Open Scholarship.' Scholarly and Research Communication 6, no. 3 (October 22, 2015). http://src-online.ca/index.php/src/article/view/202.

Mooney, Hailey, and Mark Newton. 'The Anatomy of a Data Citation: Discovery, Reuse, and Credit.' Journal of Librarianship and Scholarly Communication 1, no. 1 (2012). doi:10.7710/2162-3309.1035.

Moore, Samuel, ed. Issues in Open Research Data. Ubiquity Press, 2014. http://www.ubiquitypress.com/site/books/detail/12/issues-in-open-research-data/.

Prost, Hélène, Cécile Malleret, and Joachim Schöpfel. 'Hidden Treasures: Opening Data in PhD Dissertations in Social Sciences and Humanities.' Journal of Librarianship and Scholarly Communication 3, no. 2 (2015). doi:10.7710/2162-3309.1230.

Roorda, Dirk, and Charles van den Heuvel. 'Annotation as a New Paradigm in Research Archiving.' ResearchGate, October 7, 2014. https://www.researchgate.net/publication/269711156_Annotation_as_a_New_Paradigm_in_Research_Archiving.

Silvello, Gianmaria. 'A Methodology for Citing Linked Open Data Subsets.' D-Lib Magazine 21, no. 1/2 (January 2015). doi:10.1045/january2015-silvello.

Starr, Joan, Eleni Castro, Mercè Crosas, Michel Dumontier, Robert R. Downs, Ruth Duerr, Laurel L. Haak, et al. 'Achieving Human and Machine Accessibility of Cited Data in Scholarly Publications.' PeerJ Computer Science 1 (May 27, 2015): e1. doi:10.7717/peerj-cs.1.

Uhlir, P. F., and National Research Council (U.S.), eds. For Attribution--: Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop. Washington, D.C: The National Academies Press, 2012.

8. Metadata

8.1 Introduction

There are a large number of specialized vocabularies; in March 2016 (Vatant et al. 2013) identified 542 vocabularies used in the Semantic Web (see Figure 7). It may be noted that the boundaries between conceptual models (or ontologies) and vocabularies are not clearly defined, for example FRBR[68] is referenced as a vocabulary by (Vatant et al. 2013) while from our point of view it is rather a conceptual model. However some ontologies such as the CIDOC-CRM[69] or HADOC[70], promoted by the French Ministry of Culture, have the ambition to embrace the 'Cultural Heritage' (CH) as a whole as to move beyond the institutional boundaries (museums , libraries, archives).

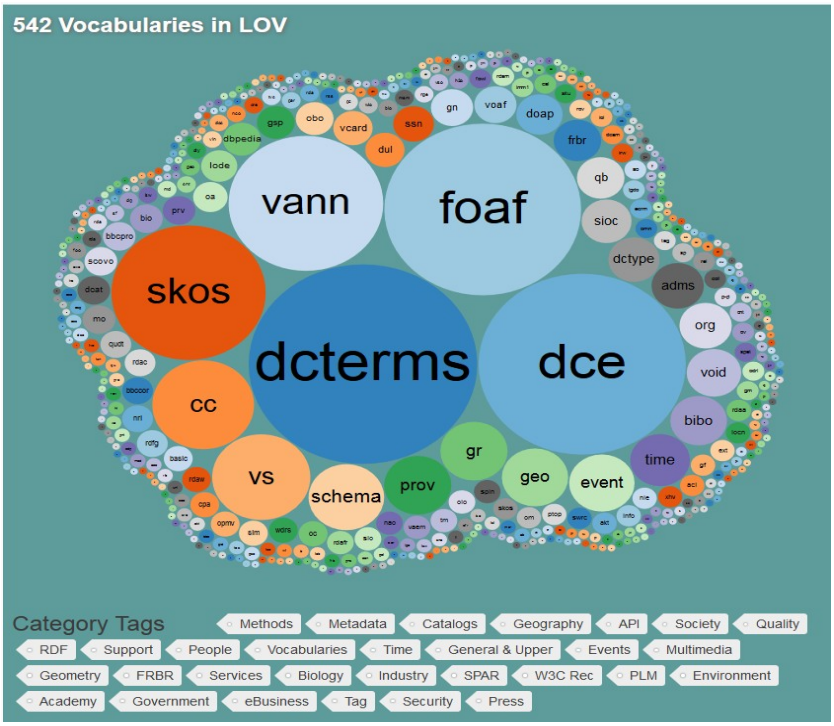

Figure 7: 542 vocabularies used in the semantic web

Concerning the target of the HAS and DARIAH projects, we propose to distinguish three areas that cover scholars' activities involved in Arts and Humanities, because those areas have indeed specific needs. These are the archival, electronic scientific text encoding and bibliographic fields.

[68] All abbreviations are explained at the end of this report

[69] http://www.cidoc-crm.org/

[70] http://www.culturecommunication.gouv.fr/

8.2 The Archival Field

In the archival field, the classification task is one of the missions that best embodies the characteristics of this sphere. To accomplish this mission, three major international standards have emerged in the years 1990-2000: ISAD (G), ISAAR (CPF), and ISDF are the references. ISAD (G), the general and international standard approved by the International Council on Archives and adopted as a standard by various countries, is the cornerstone for archival description.
The main objective is to facilitate research and exchange of information between archives departments. Based on the strong principle of non-redundant information and the notion of heritage of properties, ISAD (G) defines a set of twenty-six elements (six are mandatory) divided in seven areas. In the early 90s, this format was a syntactic encoding format expressed first in SGML and in XML. In 1998, it took the form of a DTD and offered other advantages (hyperlinks, etc.). The notion of archival collections, which involves the concepts of level and hierarchy, is at the heart of this model. First, the goal is to reflect the structure of archives and links between components of the document, and secondly, to provide a simple formalism to preserve the principle of inheritance information between levels.

The SNAC collaborative research project[71] between archive, library and museum members is based on ISAD (G). So far, it has produced a prototype built on 3.7 million descriptions. **'The prototype has achieved sufficient scale to be both a useful reference source, and a means to locate millions of historical resources located in more than 4,000 repositories around the world'**[72]. Among several forthcoming features, the SNAC consortium announces serializations of EAC-CPF graph data as RDF and GraphML.

8.3 The Electronic Scientific Text Encoding Field

In the field of description of the text contents, in its most general sense, which covers both manuscripts, printed, written language resources or oral resources transcripts, the TEI[73] (Text Encoding Initiative) established in late 1980s is the reference. **'The Text Encoding Initiative (TEI) is a consortium which collectively develops and maintains a standard for the representation of texts in digital form. Its chief deliverable is a set of Guidelines (…) In addition to the Guidelines themselves, the Consortium provides a variety of resources[74] (…) and software[75] (…) developed for or adapted to the TEI'**. The TEI website displays a list of 172 projects[76], from all around the world, which uses the TEI.

[71] socialarchive.iath.virginia.edu/snac/search

[72] socialarchive.iath.virginia.edu/home_cooperative.html

[73] http://www.tei-c.org/index.xml

[74] http://www.tei-c.org/Support/Learn/

[75] http://www.tei-c.org/Tools/

[76] http://www.tei-c.org/Activities/Projects/

The TEI is totally immersed in the XML syntax, one can conceive of it as a DTD (or XML Schema) specific to the text annotation. To date, there is no mention of a vocabulary proposal expressed RDF / RDFS / OWL.

8.4 The Bibliographic Field

In the bibliographic domain, two models (or vocabularies), DC[77] and FRBR[78] coexist and are used predominantly.
The DC was conceived in 1995 in Dublin (Ohio) by OCLC and NCSA. The current version, called 1.1, is the RFC 5013 recommendation of the IETF (Internet Engineering Task Force) validated in 2007, ANSI / NISO Standard Z39.85-2007 and ISO 15836. The 15 proposed elements, all optional and repeatable, focus on the generic description of bibliographic items and aim at a minimal interoperability between different descriptive systems. This generic model has been enriched by the Qualified Dublin Core elements which offers additional qualifiers. The main criticism of the DC is the fuzzy semantics of its elements, such as, for example, 'creator', 'coverage' or 'type'. This is a major drawback.
For the 'dc:type' element, its semantic blur has important implications for harvesting OAI-PMH stores. For example, to be harvested by some institutions, it is mandatory to use a predefined values set for the field 'dc: type' (this is the case of the BNF in France, or OpenAIRE in Europe). If these sets do not correspond, the data producer is obliged to build multiple streams OAI-PMH, which is engineering time consuming. To resolve this issue and manage the proliferation of 'dc: type', the solution adopted by ISIDORE is to use an open taxonomy of 'dc:type' associated with a <skos: hiddenLabel>.

The FRBR model was designed in the late 1990s. Actually, to be more precise, we should refer to a set of models, as the International Federation of Library Associations and Institutions offers three models: FRBR, FRAD and FRSAD developed and respectively approved in 1997, 2009 and 2010. They cover the bibliographic data and copies, authority data and subject. In a nutshell, FRBR offers a model for bibliographic records and parts of copies, FRAD (Functional Requirements for Authority Data) models the content of authority records, and FRSAD (Functional Requirements for Subject authority data) the relationships between bibliographic and subject authority file. The concept of work is at the heart of these models, which is relevant for the bibliographic point of view, but not for the cultural heritage one.

[77] dublincore.org

[78] www.ifla.org/frbr-rg

8.5 The Heritage Field

In the heritage field, the CIDOC CRM model[79] (CIDOC Conceptual Reference Model) designed in the late 1990s is the reference. In 2006, the CIDOC-CRM has been published as an international standard by the ISO (ISO 21127: 2006). The concept of event, which is a cornerstone of the cultural heritage field, is at the heart of this model. The CIDOC CRM model is often called semantic model by its designers: 'The CIDOC-***CRM*** Represents an 'ontology' for cultural heritage information, i.e. it describes in a formal language the explicit and implicit concepts and relationships relevant to the documentation of cultural heritage. The primary role of the CIDOC CRM is to serve as a basis for mediation of cultural heritage information and thereby provide the semantic 'glue' ***needed*** to transform today's disparate, localised information sources into a coherent and valuable global resource.'[80]. Specifically, the model defines 93 classes (or entities) and 161 properties.

8.6 Other Models

The study of different existing platforms (see Chapter 9 'Best practices') show that other vocabularies are used in the field of Arts and Humanities. It is essentially FOAF, SKOS and ORE.
FOAF[81] is a project devoted to linking people and information using the Web. FOAF collects a variety of terms; some describe people, some groups, some documents. Main FOAF terms are grouped in two categories. The **Core** category of which classes and properties form the core of FOAF. They describe characteristics of people and social groups that are independent of time and technology; as such they can be used to describe basic information about people in present day, historical, cultural heritage and digital library contexts. The **Social Web** category gathers, in addition to the FOAF core terms, number of terms for use when describing Internet accounts, address books and other Web-based activities. To date, FOAF proposes 13 classes and 62 properties.

SKOS[82] is an area of work developing specifications and standards to support the use of knowledge organization systems (KOS) such as thesauri, classification schemes, subject heading systems and taxonomies within the framework of the Semantic Web. It provides a standard way to represent knowledge organization systems using the RDF. It must be pointed out that ISO 25964-1 thesaurus standard, published in 2011, has been aligned with SKOS by members of the Working Group responsible for ISO 2596. They have developed a set of linkages between the elements of the ISO 25964 data model and the ones from SKOS, SKOS-XL, and MADS/RDF.

79 http://www.cidoc-crm.org/
80 www.cidoc-crm.org
81 http://xmlns.com/foaf/spec/
82 www.w3.org/2004/02/skos/

OAI-ORE[83] defines standards for the description and exchange of aggregations of Web resources. These aggregations, sometimes called compound digital objects, may combine distributed resources with multiple media types including text, images, data, and video. The goal of these standards is to expose the rich content in these aggregations to applications that support authoring, deposit, exchange, visualization, reuse, and preservation. To date, ORE proposes 4 classes and 8 properties.

8.7 General Synthesis

It must be pointed out that in his later publications the ontology FRBR is closely based on the conceptual model CIDOC-CRM. From our perspective, CIDOC-CRM can be seen as a model or as a vocabulary. A model in the sense that the semantics of each entity and each property is defined both in natural language, but also in the formal language of OWL; vocabulary in the sense that it is possible to use for a specific use, some of the properties offered by the model. One of CIDOC-CRM strengths is to be semantically defined and implemented in RDF / OWL.

We would like to stress some points. First, the extent of the description of different models, from 15 properties for the DC to 160 for CIDOC-CRM. Secondly, except for CIDOC-CRM, the rigidity of models associated with a blur in the semantics of the proposed description. Third, the TEI holds a special place in the sense that its proposals concern both the description of the text object and the content of this item. Fourth, there are two categories of models. One category, oriented work description (or document) as the DC, and other one oriented description of the events that affect an object such as the CIDOC-CRM. We also note the emergence of models which focus on the processes that affect the life cycle of an object (Kettula & Hyvönen 2012). Finally, it must be pointed out the use of SKOS vocabulary is mandatory to be able to process data enrichment by using multilingual thesauri.

After this brief synthesis, we want to emphasize the difficulty of predicting what will be the standard that will prevail in the coming five years. For example, in France, the BNF seems to tend to the choice of FRBR; the Ministry of Culture and Communication promotes the use of harmonized model for the production of cultural data, developed under the HADOC program (very closed to CIDO-CRM); the National Archives, participating in a working group of the ICA / LRE, are engaged in the development of an ontology described in OWL; at European level, Europeana[84] promotes its own EDM vocabulary, etc. This instability should be taken into account, which means that our methodological proposals (WP 7.2 & 7.3) will aim to ensure an adaptability of the target system.

[83] www.openarchives.org/ore/

[84] pro.europeana.eu/

8.8 References

KETTULA S., HYVÖNEN E., (2012), « Process-centric cataloguing of intangible cultural heritage », in *Proceedings of CIDOC 2012 - Enrich Cultural Heritage*, Helsinki, Finlande, 2012.

VATANT Bernard, VANDENBUSSCHE Pierre-Yves, « Linked Open Vocabularies » 2013, <http://lov.okfn.org/data>.

8.9 Abbreviations

BNF: French National Library
CIDOC-CRM : CIDOC Conceptual Reference Model
CH : Cultural Heritage
DC : Dublin Core
DTD : Document Type Definition
EDM: Europeana Data Model FOAF: Friend of a Friend
FRBR : Functional Requirements for Bibliographic Records
HADOC : Harmonisation des données culturelles
IFLA : International Federation of Library Associations and Institutions
ISAD : International Standard Archival Description
ISAAR : International Standard Archival Authority Record for Corporate Bodies, Persons and Families
ISDF: International Standard for Describing Functions
NCSA : National Center for Supercomputing Applications
OAI-ORE:Open Archives Initiative Object Reuse and Exchange
OAI-PMH : Open Archives Initiative Protocol for Metadata Harvesting
OCLC : Online Computer Library Center
RDF: Resource Description Framework
SGML: Standard Generalized Markup Language
SKOS: Simple Knowledge Organization System
TEI: Text Encoding Initiative
XML: Extensible Markup Language
OWL: Web Ontology Language

9. Case study: the French open access research data ecosystem

In this chapter we describe several infrastructures from the French open access research data ecosystem as case study, which may function as examples of best practice for HaS. The French case has been picked for several reasons: through CNRS and INRIA as DARIAH affiliated partners a comprehensive first-hand access to the information is given, the French infrastructure – partly for reasons of the national political system – is quite centralised and therefore very suitable for a full picture. This situation is not given for the other partners in this work package – Germany and the Netherlands – meaning that case studies would involve more resources (in terms of working hours) that allocated for this work package. However a comparable case study for Germany has been compiled in 2013 and is open access available through nestor – the German competence network for digital preservation[6].

9.1 State-of-the-art open access research data for the Humanities in France: The French Open Access Research Data Ecosystem[85]

The CNRS, in connection with the University of Aix-Marseille, has implemented an ecosystem which aims to cover the entire lifecycle of the production of scientific data and publications. This ecosystem is based on the following organizations:

Open Editions[86] offers comprehensive services in journal publications, books, scientific blogs and scientific events in open access.

CCSD (Centre pour la Communication Scientifique Directe), which depends on CNRS, INRIA and Université de Lyon, offers a set of services for the management of open archives (HAL SHS[87])

The Very Large Facility TGIR Huma-Num[88] offers a range of services dedicated to the production and reuse of data.

In this abstract, we focus on the tools offered by Huma-Num that fall within the scope of WP7.1 and WP7.2 of HaS DARIAH.
In this abstract, we focus on the tools offered by Huma-Num that fall within the scope of WP7.1 and WP7.2 of HaS DARIAH.

The workflow implemented by Huma-Num has been built on interoperability. The aim is to foster the exchange and dissemination of metadata but also of the data itself via standardized tools and lasting, open formats. These tools, developed by Huma-Num, rely on Semantic Web technologies, mainly for their auto-descriptive features and the opportunities for enrichment that they provide.

85 Abstract. See Annex 2 for the extended description.

86 https://www.openedition.org

87 http://halshs.archives-ouvertes.fr

88 www.huma-num.fr

The first objective is to promote the sharing of data so that other researchers and communities can reuse the data in an interdisciplinary perspective, and if need be with other methods. For example, a map is a scientific object that can be analyzed and described from the point of view of a geographer or that of a historian. More generally, these services which rely on the principles and methods of the Semantic Web (RDF, SPARQL, SKOS, OWL) make it possible to document or re-document data for various uses without locking them in inaccessible silos. Other interoperability technologies complete them, such as the OAI-PMH. Another important point is to make the storage of data

The second objective is to prevent the loss of data by preparing their preservation over the long term. The documentation associated with appropriate formats, which are the basis of data interoperability, greatly facilitates the archiving process.
All the services developed by the Huma-Num accompany the life cycle of research data and are designed to meet the needs of scholars in Humanities and Social Sciences:

- SHARE and DISPLAY. The NAKALA tool offers services to store, display and share documented and standardized metadata and data based on interoperable technologies;
- DISSEMINATE. The NAKALONA tool in connection with the CMS Omeka provides means to editorialize data stored in NAKALA and offers the features of this CMS such as its search engine;
- TAG and PUSH. The ISIDORE tool enriches the data stored relying on several disciplinary thesauri and provides the functionality of a faceted search engine in order to ensure better visibility.

9.1.1 NAKALA: SHARE and DISPLAY

Noting that many teams and research projects do not have the necessary digital infrastructure that will provide a persistent and interoperable access to their digital data, Huma-Num has implemented a service called NAKALA exposure. NAKALA offers two types of services: one to give access to the data and another one to expose metadata. By relieving scholars of technical management, it enables them to concentrate on the scientific value of their data. Data hosted by Nakala may be editorialized with the NAKALONA pack[89] (combining Omeka and Nakala) developed and managed by Huma-Num; they can be shared via a SparqlEndPoint or via the OAI-PMH protocol and searched via the multilingual and multifaceted access platform ISIDORE (see below).
The key points of NAKALA are the following:

- W3C standard languages used;
- Secure servers located in Europe;
- Sustainability. NAKALA is managed by the Very Large Facility Huma-Num which is a unit of the CNRS, guaranteeing continuity of service;
- PID management based on Handle;

[89] http://www.huma-num.fr/services-et-outils/diffuser

- Open source development based on reliable and proven components (Handle server, PROAI OAI PMH server and Virtuoso Triple Store server);
- The whole NAKALA vocabulary is described in Annex 2.

9.1.2 COCOON: Specialized for ORAL Corpora

Cocoon (COllections de COrpus Oraux Numériques), managed by two CNRS research units (Laboratoire de Langues et civilisations à tradition orale and Laboratoire Ligérien de Linguistique), is a platform that provides services for managing digital oral corpora. By oral corpus we mean recordings speech (audio, video or other measures of physiological activity), annotated (transcriptions, translations, etc.). Services cover the entire lifecycle of the data, from archiving in one system to final archiving. The identification system used PURL (Handle and ARK identifiers are also added when a resource is harvested by ISIDORE or archived at CINES). Metadata are accessible through the OAI-PMH protocol as well as an -Endpoint (with RDF encoding following the 'Europeana Data Model (EDM)', Cocoon warehouse is harvested especially by OLAC organizations CLARIN and Isidore or through portals such as 'Corpus of speech' of the French Ministry of Culture.
Sustainability in long-term data is ensured for an interim period by the French archiving platform CINES [90] and the futures will be under the French National Archive's responsibility. Some figures: 10 741 records; 3 300 transcripts, 5 000 hours of recordings, 7 To of data.

9.1.3 ISIDORE: HARVEST and SEARCH

ISIDORE is a platform allowing access to digital data in the Humanities and Social Sciences. Its architecture relies on the languages of the semantic web (RDF/RDFS/OWL) and provides open access to data. ISIDORE is a managed by the Very Large Facility TGIR-Huma-Num (CNRS, Aix-Marseille University, Campus Condorcet) and implemented by the Center for Direct Scientific Communication (CCSD / CNRS). The key points of the ISIDORE platform are the following:

- **Targeted harvesting** of metadata and scientific data structured according to international standards available in open access;
- **Indexing of unstructured data** (full text of a scientific article, for example) and of structured data (documentary metadata, for example);
- **Standardization** of metadata and enrichment of data relying on vocabularies recognized in the community (DC, DCterms, FOAF, ORE, RDFS, SKOS);
- **Multilingual** (English, French, Spanish) search GUI exploiting the richness of structured data and vocabularies to make the user an actor of his search;
- **SparqlEndPoint** on sources and indexed data: In 2013, the TGIR Huma-Num and DANS developed a prototype (Proof of Concept) in order to show the connection between two repositories NARCIS (DANS) and ISIDORE. The connections relied on

[90] https://www.cines.fr/en/

the two SparqlEndpoints and the alignment of the disciplinary vocabularies used by these systems (cf. Annex 2). This proof of concept demonstrates the compatibility between ISIDORE and OPENAIRE. In other words, a query in OPENAIRE can dynamically search in ISIDORE and/or in other SparqlEndPoints managed by other stakeholders. This kind of decentralized architecture based on several SparqlEndpoints is more resistant to failure and ensures a great scalability and sustainability

- Smartphone applications;
- Supplying metadata enriched by several multilingual thesauri;
- Possible integration of the search engine Isidore in another environment by providing widgets (IMOCO).

9.1.4 General Synthesis

The French open access research data ecosystem can be represented as a multi-layer system which aims to deal with the various stages in the scholarly content life-cycle as well as to further a culture of data sharing in the Humanities and Social Sciences.
The first layer offers, through OpenEdition, open journals, open books, blogs in which scholars can submit their scientific papers, conference CFPs or write blog posts. OpenEdition is run by the Centre for open electronic publishing (Cléo), a unit that brings together the Centre National de la Recherche Scientifique (CNRS), the Université d'Aix-Marseille, the École des Hautes Études en Sciences Sociales (EHESS) and the Université d'Avignon et des Pays de Vaucluse.
Open Archives (HAL, HAL-SHS, TEL, Theses) constitute the second layer in which scholars deposit their scientific papers published by journals managed by OpenEdition or other publishers. CCSD and ABES (Bibliographic Agency of Higher Education) run these services.
The third layer is dedicated to managing and securing scientific metadata and data with NAKALA. These metadata and data are the product of descriptions or experiments which were presented by scholars in scientific publications.
The fourth layer is constituted by the ISIDORE and NAKALONA platforms which tag, enrich and push data and metadata produced by the other three layers. Both platforms share these data by providing access to two SparqlEndpoints, making them interoperable with other platforms such as NARCIS, OPENAIRE or any platforms based on RDF triplestore. It must be pointed out that ISIDORE harvests far beyond French data providers since American, Belgian, Spanish, and Italian data providers (library, research centers, etc.) have already requested to be harvested.

9.2 A newcomer in the French ecosystem: Ortolang

The panorama described in Annex 2 regarding the French ecosystem corresponds to a general layer of services operating at the scale of DARIAH, that is « Arts and Humanities ». Another content hosting offering in France has been recently set up which deals specifically with linguistic resources. This platform, named Ortolang for (Open Resources

and TOols for LANGuage) has a role to play in the general data repository landscape in several respects:

- It is under integration in the French ecosystem (in particular into the Isidore research space) making Ortolang one element in the global French contribution to DARIAH.
- Even if centred on linguistic data, Ortolang hosts in practice any kind of language-based resource, whether written documents (and a lot of research data in Art and
- Humanities is text based), spoken interaction of any combination of text and other communication form.
- Most of all, we claim that some of the features of Ortolang may be of interest in the HasDariah context.

Therefore, we do not describe here the features that are usual requirements for repository platforms such as persistent identifiers, single sign on, long-term archiving (through the solution provided by Humanum), etc. We focus here on the way Ortolang tackles the data all along its life cycle and highlight the specificities of the platform.

9.2.1 Tackling the data life cycle from the onset.

It is a requirement on most serious hosting platforms that the data should be:

- properly described by appropriate metadata;
- conditioned in normalized or at least well described formats;
- secure in terms of reuse by means of proper licencing terms (preferably open licences), legal terms (related to legal issues regarding persons), etc.

When a content-hosting platform requires such high standards, it necessarily happens very late in the life cycle of the data: in practice, no research data is ever born with such characteristics. The way Ortolang deals with this issue relies on the separation between two central functions of a repository, namely: « hosting contents » and « publishing contents ».

9.2.2 Workspaces

From a producer's point of view, just after having created an account on the platform, the first thing to do on Ortolang is to create a workspace. Such workspaces may be viewed as private areas in which producers deposit data, work upon them, fill in metadata, etc. The only requirement Ortolang makes upon workspaces is that producers eventually aim at making data accessible to a larger community than the producers themselves (they are strongly encouraged towards offering access at least to the academic community (research and teaching)). Even though in practice users usually work upon their own machines and regularly update the data in their workspaces, depositing in a workspace offers the following advantages:

- The data is secured (Ortolang is based upon a reliable computing service)
- The data is shared among those people to whom the workspace creator gave rights

- Visualization tools (for instance modified versions of TEI boilerplate) help in curating the data.
- Tools external to the platform may access data through an API. In particular tools aiming at data curation are of much relevance here.

However, workspaces are not continuously versioned, it is the user's responsibility to tackle this issue.

9.2.3 Asking for publications

The step that immediately follows the elaboration of the data into the life cycle is the action of enabling access to other users to the data. In Ortolang, this corresponds to publishing a workspace. In order to do so, the producers are required to fill in mandatory metadata, stating the access rights on any object of the workspace make a publication request.
This operation forbids any further modification of the workspace and leads to a workflow involving moderators of the Ortolang platform. This workflow may result into one of two possibilities:

- The publication is rejected (and the rejection is motivated for instance because legal issues are raised). In such a case, a message is sent to the workflow owner and the workspace is made modifiable again (so that the producer may tackle the issue(s).
- The publication is accepted. In this occurrence, a new object is accessible on the platform and the workspace is made modifiable again. The published version will not be modified, but after some time, it may happen that a new publication (resulting in a new version of the data) is asked for.

The overall idea is that:

- Published versions have to be fixed content: some users want a reference (a PID) to a clearly identified and un-mutable contents
- Data has to live, corrections have to be made, sometimes more recordings in a series have to be added, etc. This occurs not only by means of data reuse (which would correspond to brand new workspace) but also because a particular corpus evolves.

9.2.4 Conclusion

Ortolang manages a workflow dedicated to publishing workspaces. However, from the outside world, the difference between a published workspace as opposed to a not yet published one is a matter of whom may access to its contents. In each case, providing a 'cloud based' repository that is securely available for web services is a must either for data curation or because it is a privileged means of data reuse. We are convinced that in a near future, web oriented version of language tools will be available, and we are also convinced that they might be the glue that fits DARIAH pieces together.

9.3 Managing terminologies: OpenTheso

9.3.1 What is OpenTheso?

Opentheso is an open source web-based collaborative solution for managing thesauri, taxonomies and other controlled vocabularies. Such a tool is central to connecting research data on the semantic level: it allows research teams to easily build and manage their own scientific terminology, share it on the web and map it to other vocabularies. It can be used in different contexts: archives and records management system, library catalogues, databases, research blogs, etc.

Opentheso has been developed at the French National Center for Scientific Research (CNRS) since 2007 for the French archaeological libraries network Frantiq91 (39 French National Centre for Scientific Research, Ministry of Culture and regional authorities' libraries), by scientific research teams (such as ZooMathia), by the French facility Huma-Num (in relation with its search engine ISIDORE) and by the Lyons Hospitals network, originally as a collaborative thesaurus management tool. It is currently used by researchers, librarians and engineers dealing with research data.

9.3.2 Dealing with a huge variety of specialized terminologies

A huge diversity of vocabularies is created to describe, analyse and retrieve scientific data and publications, throughout the entire cycle of scientific activity. Those terminologies are created by different professionals:

- Researchers
- Librarians
- Publishers
- Archivists
- Curators

Each professional conforms to its own community's standards, do not share the same goals, and do not always address the same audiences. They sometimes address a very small and specialized community (ex: researchers) but sometimes much broader audiences (ex: university students and lecturers, national or international audiences). Even though data and publications are stored in databases, the tools used are different: mainstream databases, reference software for managing bibliographies, library information systems, archival description software, Geographical Information Systems.

Furthermore, terminologies are themselves like living entities: new concepts are created, and, depending on the school of thought and on the period, different terms are privileged. Terminologies partake of the scientific activity and debate and reflects a point of view.

9.3.3 OpenTheso: enabling the semantic interoperability of research metadata

Within the ecosystem developed by the French National Center for Scientific Research (CNRS), OpenTheso is an open source software that complies with the W3C standards. It aims at managing thesauri and terminologies, enabling the scientific communities to:

- Build and manage their own scientific terminology, share it on the web and map it to other vocabularies;
- Ensure the interoperability of the terminologies used (persistent identifiers, conforms to ISO and semantic web standards, SKOS-compliant);
- Discuss and define terminologies collectively and at a distance;
- Connect to the other open source software in order to use those standardized terminologies;
- Manage multilingualism.

It can thus be connected with a content management system such as Omeka that can be used to display scholarly collections and exhibitions or a library catalogue, such as Koha.

Opentheso is a collaborative tool which has been developed to enable distance co-working, with a user-friendly interface (smooth learning curve). Users only need a web-browser to visualize or manage the concepts. A workflow enables administrators to suggest and discuss concepts (as « candidates »).

Opentheso can manage several structured vocabularies or thesauri at once, as well as polyhierarchical vocabularies. An image can be uploaded for each concept, which can be useful to disambiguate or illustrate a concept; this feature can also be used to reach pedagogical goals since it enables the user to visualize and differentiate concepts. Scope notes can be used to define each concept.

It has been developed to enable Knowledge Organization Systems interoperability both at a semantic and at a technical level, with a focus on the open linked data and the semantic web.

- Semantic interoperability:
  - Mapping interface (enable matching concepts with other vocabularies)
  - Multilingualism
- Technical interoperability:
  - SKOS format (Simple Knowledge Organization System, W3C recommendation)
  - SKOS and csv exports and imports
  - Allocation of a persistent identifier (ARK) for each newly created concept
  - REST and SOAP web services

Opentheso complies with ISO 25964:2011 and ISO 25964-2:2012 standards (Information and Documentation. Thesauri and Interoperability with other vocabularies), which aims at

facilitating relevant information retrieval, through the standardization and possible interconnection of controlled vocabularies in the context of the semantic web.

Each concept is labelled with a preferred term and, according the ISO standards, relationships include:

- Equivalence (between synonyms and near-synonyms, also used for translations in different languages)
- Hierarchical (between broader and narrower concepts)
- Associative (between concepts that are closely related in some non-hierarchical way)
- Thematically: Facets can also be used to group concepts.

It can be downloaded on GitHub. It has been funded by the CNRS since 2007 and it is Open Source: Licence CeCILL_C, GNU GPL.

9.3.4 An example of what can be done

The multilingual thesaurus for Archaeology PACTOLS:

- is used by 38 archaeology libraries in connection with the open source Information Library System Koha;
- has thus been one of the first thesaurus used to index research data in the ISIDORE platform since it was SKOS-compliant;
- has been mapped to the specialized vocabulary of ArSol, an archaeological database used to process stratigraphic data in order to ensure its semantic interoperability (it also uses the CIDOC-CRM ontology) since it uses persistent identifiers and standardized formats;
- has been used to map the vocabulary of an archaeologist database describing 'fanums' (gallo-roman sanctuaries. The data was about to be lost since it used non standardized technologies. The database has been exported from a FileMaker Pro system and the data model has been standardized to fit the requirement of the semantic web;
- is used to index the INRAP's (French National Institute for Preventive Archaeology) website;
- has been translated into 7 languages;
- is used as a basis for several research teams;
- is about to be mapped with other terminologies (ex : French Ministry of Culture vocabulary for Archaeology).