DELIVERABLE

Project Acronym: eCloud - Europeana Cloud

Grant Agreement number: 325091

Project Title: Europeana Cloud: Unlocking Europe's Research via The Cloud

Europeana Cloud Deliverable 1.4 Content priorities for Humanities and Social Sciences research communities

Revision: Final

Authors:

Ingeborg Versprille (CERL)
Eliza Papaki (Athena RC)
Marian Lefferts (CERL)

Contributors:

Vicky Garnett (TCD)
Roberta Cacciaglia (The European Library)
Marie-Claire Dangerfield (Europeana)
Cecile Devarenne (Europeana)
Stefan Ekman (UGOT/SND)
Agiati Benardou (Athena RC)
Costis Dallas (Athena RC)
Thomas Baldwin (CERL)

Editor:

Friedel Grant (LIBER)

Revision History

	When?	Who?	Organisation	What?
1.	06/11/14	I. Versprille	CERL	Initial draft of Chapters 3-4
		E. Papaki	Athena RC	Initial draft of Chapter 2
2.	20/11/14	I. Versprille	CERL	Processed edits by F. Grant
				Processed comments by A. Benardou
				Added recommendations
		M. Lefferts	CERL	Added Executive Summary
3.	1/12/14	I. Versprille	CERL	Processed comments by WP1 Project Partners and Europeana staff
4.	17/12/14	I. Versprille	CERL	Processed comments
				Final edit
				Final draft report to reviewers
5.	20/1/15	I. Versprille	CERL	External review by members of the Research Communities Advisory Board
				Final edit
				Final report

D1.4 Content priorities for Humanities and Social Sciences research communities

Executive summary

This Deliverable 1.4. is the result of work carried out in *Task 1.2.2 Matching Europeana content to the Research Communities*. The aim of this task was to match existing content in Europeana and The European Library to the humanities and social sciences research communities (the 'target audience'). The report therefore builds on the outcomes of tasks *1.1.3 Research Communities identification and definition* (D1.1) and *1.2.1 Desk research into the state of the art on scholarly content use* (D1.2.).

For the content analysis, the report made use of previous content analysis work undertaken by Europeana and The European Library, but in addition much effort went into providing a current content analysis of both the Europeana and The European Library datasets. The indications of subject matter that were available in the Europeana and The European Library datasets were used to match their content to identified academic disciplines.

Based on this analysis and what we have learnt about scholarly content use, recent research into user requirements and the expectations of the target audience, the report presents its conclusions in the form of further recommendations for the content strategy and the future development of Europeana Research, which are fully outlined in chapter 5. The recommendations around several issues, which are summarised here:

- ✓ Target audience: it is of eminent importance to exactly define the intended user group in order to determine the service level to be provided by Europeana Research.

- ✓ Content strategy: clearly state the goals of Europeana Research and the plans for its future development, provide information about the selection policy[1] and an exact definition of what (type of) content and metadata the service will contain (contextualisation).

- ✓ Content: source high quality content with comprehensive metadata.[2] Improved handling of multilingual resources and metadata would make Europeana Research an even more

---

[1] Stating selection criteria motivating the choice for certain content and content and resource types are available and how they can be used.

[2] Europeana Cloud Deliverable 1.5 (3) 'Expert Forum Tools and Content for Social Science Research', p. 17; Link to Expert Forums: http://pro.europeana.eu/web/europeana-cloud/results/-/document_library_display/p6BV/view/1861918 (9 January 2015)

powerful resource. Based on the study of digital scholarly content use by the target audience[3] as well as the content analysis of the Europeana and the European Library portals[4], preferences for certain types of content re a) Resource type, b) Geography, c) Chronology and d) Subject were distilled:

a) Resource type: text: in addition to the traditional text and image aggregation, users of Europeana Research especially value full-text, sound and video.[5]

b) Geography: the material currently available is rich and diverse. Its potential could be enhanced by including more material from Eastern European countries.

c) Chronology: actively aggregate content from the 19th and 20th centuries to ensure Europeana Research offers unique research material, not readily available in other digital resources.

d) Subject: it is likely that research communities interacting with Europeana Research will require tailored research corpora with critical mass focused on specific subjects. Readily available content for the humanities includes: history, musicology and philosophy. For the social scientists: political science, world politics, ideology and propaganda; social history, sociology, (popular) culture, religion, possibly genealogy.[6]

✓ Tools: the data in Europeana Research will be accessible via an API, which makes it likely that Europeana Research will mostly be used by teams of academics who can harnes some technical resources. The report identifies a substantial risk in this, as the common perception of API development would be a programmer's, rather than that of a researcher.[7] Any tools researchers may want to use on top of the data will be determined and/or developed by the group, although Europeana Research may wish to support sharing and reuse of tools.

The report concludes that the target audience for Europeana Research is fluid with ever changing research questions being asked by interdisciplinary groups of scholars in an ever changing mix of academic disciplines (historians and economists, or art historians and scientists, or social sciences and sociology students, etc.), and with an ever evolving set of skills, including those related to data mining. Those developing Europeana Research will need to be in constant communication with their scholarly users, in order to work with them to prepare the best tailored service based on an extremely rich but variable resource.

---

[3] See Chapter 2

[4] Described in Chapters 3-4

[5] As this content may not be readily available in this form and quantity in other digital resources.

[6] For social scientists the relevance of the described datasets is not always easy to determine.

[7] Citing Jennifer Edmond, Trinity College Dublin during a meeting: Alastair Dunning, 'Drafting priorities for Europeana Research', 30 July 2014) http://pro.europeana.eu/web/europeana-cloud/blog/-/blogs/trackback/drafting-priorities-for-europeana-research (20 November 2014)

Table of Contents

List of Figures and Tables

Figures

Tables

Glossary of Terms

API: in computer programming, an **Application Programming Interface** (**API**) is a set of routines, protocols, and tools for building software applications.[8]

Aggregator: an aggregator in the context of Europeana is an organisation that collects metadata from a group of data providers and transmits it to Europeana. Aggregators also support the data providers with administration, operations and training.[9]

Collection: (in the context of Europeana/TEL): the term Collection will have different meanings depending of the context in which it is used. Within The European Library context a collection is defined as a set of criteria for selecting resources from the broader information space according to the interests of users". Within the Europeana context, a collection is defined as: a set of Metadata that has been intentionally-defined at an institutional or thematic level.[10]

Content (as defined by Europeana): as described in the Data Exchange Agreement (DEA), the term 'content' refers to a physical or digital 'object' that is part of Europe's cultural and/or scientific heritage, typically held by the data provider or by an aggregator of the data provider. By 'metadata' we mean the textual information (including hyperlinks) that may serve to identify, discover, interpret and/or manage digital objects.[11]

Data: individual pieces of information. Data as an abstract concept can be viewed as the lowest level of abstraction, from which information and then knowledge are derived. Data is least abstract, information next least, and knowledge most. Data becomes information by interpretation. [12]

Data Provider: an organisation that contributes Metadata describing Content that it is offering online. Note: This should be used instead of Content Provider or Provider[13].

Dataset: 1) A dataset is a multidimensional array of data elements, together with supporting metadata.[14] 2) An administrative unit representing a suitable amount of Content or Metadata for ingestion or download.[15]

---

[8] http://en.wikipedia.org/wiki/Application_programming_interface (15 December 2014)

[9] http://pro.europeana.eu/about/partners (16 April 2014)

[10] http://pro.europeana.eu/glossary (16 April 2014)

[11] <http://pro.europeana.eu/documents/866067/983522/D3.8+%26+MS12+Content+and+Metadata+Strategy +and+Plan+2013> (29 April 2014)

[12] http://en.wikipedia.org/wiki/Data (25-02-2014); See for more information: White paper Europeana: http://pro.europeana.eu/c/document_library/get_file?uuid=cb417911-1ee0-473b-8840-bd7c6e9c93ae&groupId=10602

[13] http://pro.europeana.eu/glossary (16 April 2014)

[14] http://www.hdfgroup.org/HDF5/Tutor/crtdat.html (15 April 2014)

[15] http://pro.europeana.eu/glossary (16 April 2014)

Desk research (also known as secondary research): involves the summary, collation and/or synthesis of existing research rather than primary research, where data is collected from, for example, research subjects or experiments.[16]

Digital object: a digital representation of an object that is part of Europe's cultural and/or scientific heritage. The Digital Object can also be the original object when born digital[17].

Digital tool: One definition of a digital tool has been proposed by Hughes[18]
"I would suggest that a digital tool is any piece of software that can be used to gather, analyse and/or process data. A useful distinction can be drawn between tools that enable existing (i.e. analogue) research processes to be conducted better and/or faster, and tools that enable researchers to ask, and answer, completely new research questions. As digital objects are rarely only processed with just one tool (and should be re-useable and re-purposable in the future), interoperability and in/output interfaces are crucial for evaluating tools, whichever category they fall into."

Linked Open Data: Linked Open Data is a way of publishing structured data that allows Metadata to be connected and enriched, so that different representations of the same content can be found, and links made between related resources.[19]

Metadata: the textual information and hyperlinks that serve to identify, discover, interpret and/or manage Content.[20]

Persistent Identifier: (PID) a long-lasting reference to a digital object—a single file or set of files.[21]

Primary Data: while taking into consideration the scholarly discussion on the terminology of sources, this Deliverable report adopts the term 'primary data' for computerised data sets, rare books and manuscripts, maps and charts, photographs, artefacts, moving image or sound recordings, while 'secondary' material to describe journals, newspapers, books and electronic publications accordingly. (…)[22]

---

[16] http://en.wikipedia.org/wiki/Secondary_research (3 December 2014)

[17] http://pro.europeana.eu/glossary (16 April 2014)

[18] L.M. Hughes (ed.), *Evaluating and measuring the value, use and impact of digital collections,* (London, UK: Facet Publishing, 2012), p. 43

[19] http://pro.europeana.eu/glossary (16 April 2014)

[20] http://pro.europeana.eu/glossary (16 April 2014)

[21] http://en.wikipedia.org/wiki/Persistent_identifier (24 April 2014)

[22] Europeana Cloud Deliverable 1.2 'State of the art report on digital research practices, tools and scholarly content use', 2013, pp. 13-14 http://pro.europeana.eu/web/europeana-cloud/results/-/document_library_display/p6BV/view/2240207 (8 January 2015)

1. Introduction to the task

This Deliverable explores the extent to which the content in the portals of The European Library and Europeana matches the needs of the humanities and social sciences research communities. It is has been compiled by Work Package 1 of the Europeana Cloud project – a Work Package whose main mission is to assess researcher needs and behaviours.

The objectives of the Europeana Cloud project are to provide new content, new metadata, a new linked storage system and services for researchers. The project will also develop a new service labelled Europeana Research to discover and re-use content from the project partners The Europeana Foundation, The European Library and Poznan Supercomputing and Networking Center.[23] European national, research and university libraries will provide digital content via The European Library to Europeana Research, and Europeana brings in a large, continually updated, data set aggregated from the other constituents of the GLAM sector. The project is also used to experiment with the aggregation of digitised content (i.e. actual digital objects) and to engage with a new and different user audience by supporting researchers to exploit the content in the Cloud infrastructure[24] and collaborate in the development of tools and services based on this resource. This 'new and different user audience' consists of humanist and social sciences research communities, defined in Chapter 2 as the 'target audience' to be serviced by Europeana Research.

1.1 Main Aims of this Deliverable

The main aim of this report is to inform and prioritize the ingestion of content for humanities and social sciences communities from existing providers into Europeana Research, via Europeana and The European Library.[25]

To do this, a content analysis was undertaken to identify the gaps in the existing content in the portals of Europeana and The European Library. Here, we present the results of this analysis, take a look at the target audience of Europeana Research and their specific needs, explain the methodology used and analyse the results.

Much of the work in this report incorporates or builds upon work outlined in earlier Deliverables.

In Chapter 2, for example, we focus on the target audiences for Europeana Research and their use of digital scholarly content. In order to do this, the communities for which Europeana Research will be created first had to be explored. This was undertaken by The Institute for War, Holocaust and Genocide Studies (NIOD), Data Archiving and Networked Services (DANS), University of Gothenburg (UGOT) and Athena Research and Innovation Center in Information Communication & Knowledge Technologies (Athena RC) and resulted in Deliverable 1.1 Research Communities' Identification & Definition report.

---

[23] Competitiveness and innovation framework programme 2007-2013: Annex 1: Description of Work, p. 7

[24] Deliverable 1.1 'Research Communities Identification and Definition Report', p. 2

[25] Competitiveness and innovation framework programme 2007-2013: Annex 1: 'Description of Work' : 'Europeana Cloud: Unlocking Europeana's Research via The Cloud', pp. 5-6

To be able to match the content of Europeana Cloud to its target audience, relevant literature about user behaviours was studied by Athena RC, CERL[26] and UGOT. The desk research[27] done in this part of the Work Package delivered some preliminary results about the way researchers in the humanities and social sciences interact with digital content. Another part of the desk research concentrated on the use of digital tools by researchers. This resulted in Deliverable 1.2, from which information was also incorporated into Chapter 2.

Throughout this deliverable, we also drew on earlier desk research into the digital scholarly content use of the target audience and the usefulness of the Europeana portal for research. This was assessed through a web survey and three Expert Forums.[28] The web survey was executed by Athena RC and NIOD, while the Expert Forums were held to investigate which tools and content Europeana Research needs to have or develop to engage humanities and social sciences research communities. A short presentation of the key findings of the Expert Forums held so far (one more is scheduled for July 2015) is available on the Europeana Cloud blog[29].

The actual analysis of the existing content, the methodology used and the representation of the results as well as the identification of possible gaps in the portals of The European Library portal and Europeana are described in Chapters 3 and 4 respectively. A structured description of the content of the portals was prepared by staff at The European Library and staff at Trinity College Dublin and based upon preparatory work done in the project by staff at the Europeana Foundation and CERL.

Finally, in Chapter 5, we present some recommendations for the content strategy and future development of Europeana Research, for further discussion in the course of the Europeana Cloud project.

1.2 How the results will be used

The results of this research will form the basis for a more in-depth analysis of user requirements[30] and for the development of a content strategy for Europeana Research, resulting in Deliverable 1.6, 'Content Strategy Report'. These tasks also fall under the remit of Work Package 1.[31] Furthermore, the results will feed into the activities of Europeana Cloud Work Packages 3 and 4:

- ✓ Work Package 3, titled 'Exploiting Europeana Cloud with services and tools for researchers', will examine the user behaviour studies and studies of digital research practices to identify

---

[26] Consortium of European Research Libraries – www.cerl.org (1 April 2014)

[27] Secondary research (also known as desk research) involves the summary, collation and/or synthesis of existing research rather than primary research, where data is collected from, for example, research subjects or experiments. http://en.wikipedia.org/wiki/Secondary_research (3 December 2014)

[28] Europeana Cloud Research Community Web Survey T1.3.4, 2013; Expert Forums: http://pro.europeana.eu/web/europeana-cloud/results/-/document_library_display/p6BV/view/1861918

[29] http://pro.europeana.eu/web/europeana-cloud/blog/-/blogs/researchers-tell-u… (1 April 2014)

[30] Europeana Cloud Deliverable 1.3 'User requirements analysis and case studies report', 2013 http://pro.europeana.eu/web/europeana-cloud/results/-/document_library_display/p6BV/view/2240207 (8 January 2015)

[31] T1.2.3 Content Strategy Development. Competitiveness and innovation framework programme, p. 6 http://pro.europeana.eu/web/europeana-cloud/results/-/document_library_display/p6BV/view/1861926

future users and develop personas[32] for Europeana Research. These findings will serve as input for the development of dedicated tools and services for Europeana Cloud.

- ✓ Work Package 4 deals with the ingestion of content and metadata development. The main objectives of WP4 are the ingestion of new metadata and content, the enhancement of the Europeana Data Model (EDM) to stimulate the (re-)use of data by researchers and to improve the discoverability of data. CERL and the Europeana Foundation will use the content strategy developed in Work Package 1 as a means to select relevant content for Europeana Research.

---

[32] Descriptions of the types of researchers who will form the target user group for Europeana Research.

2. Mapping the target audience for Europeana Research with the use of digital scholarly content

Shaping content priorities for the research target audience of Europeana Research, namely humanists and social scientists, requires studying and understanding their digital user needs regarding scholarly content use. Building Europeana Research based on the user requirements of humanities and social sciences researchers is vital for the successful development of a research centered service that will meet existing as well as future digital needs of researchers. Aiming to provide feedback into the design and implementation of Europeana Research, this chapter will try to synthesize information on defining these two communities as the target audience of the project. This will be done by shedding light on the differences and similarities regarding their scholarly content use as these were identified through desk and empirical research during the course of the Europeana Cloud Project. Relying therefore on work already conducted in this project, this chapter aims to contribute to the wider discussion of shaping and informing content needs for humanities and social sciences research communities. Their content needs will be studied based upon their scholarly content use documented in the literature as well as in the Europeana Cloud Expert Forums, Web Survey and Case Studies report, leading to suggestions for the user needs to be addressed in the service.

Capturing scholarly content use, focusing particularly on Europeana, can be a demanding task due to the lack of information documenting how researchers use and interact with this content. The Europeana User Survey, undertaken outside the auspices of this project, did uncover instances of academic usage but did not go into further detail. [33] This gap was attempted to be filled/reduced by following a mixed methodology of both background and empirical evidence. This methodological conjunction would allow for both tracing answers in the literature as well as identifying scholarly trends through empirical research thus leading to more concrete conclusions regarding scholarly content use by humanists and social scientists. Without intending to repeat discussions already presented in the context of the desk research and empirical background, conclusions here will be gathered, compared and analysed aiming to take this overview a step forward into informing content priorities for Europeana Research.

For the purposes of this Deliverable, the target audience examined here confines to the above-mentioned research communities. Besides this target audience, there are other stakeholders such as: research infrastructures and content aggregators, developers, IPR specialists and the extended Europeana community of services, policy makers and funders. [34] These will not be dealt with in this report.

---

[33] Europeana, 'Results of the Europeana User Survey 2014', http://pro.europeana.eu/pro-blog/-/blogs/2222979 (14 October 2014)

[34] Europeana Cloud Deliverable 6.1 "Stakeholder Engagement & Infrastructure Plan", p. 8 http://pro.europeana.eu/web/europeana-cloud/results/-/document_library_display/p6BV/view/2240207 (8 January 2015)

2.1 Identifying and defining the research communities

This section draws upon the conclusions & recommendations of Deliverable 1.1 Research Communities Identification and Definition Report. It describes how these communities can be defined and what method has been used to define them in the project so far. Furthermore, it will explain why researchers within six subject domains in both humanities and social sciences have been chosen as 'prime candidates to be supported by Europeana Cloud'.

Among the aims of the Europeana Cloud Project one is to approach and engage research communities. For this reason, the definition of the audience was prioritised since the beginning of the project by focusing on specific communities. Having a set target audience would allow the study of their needs in greater depth as well as the formulation of ultimately more effective suggestions for the development of a researcher centered service. Setting as starting point the aims of Europeana Research for facilitating "access to Europe's cultural and intellectual heritage" for "researchers undertaking digitally-enabled research" in the disciplines of humanities and social sciences, Europeana Cloud was oriented towards these communities as they "are amongst those most familiar with using items from cultural collections" but also "because these fields are actively developing adjoining European initiatives such as DARIAH and CESSDA".[35] Guided by the content of Europeana and the lately-defined aims of this new service, the target audience is defined as "a group of researchers working in a common subject domain that is either part of the humanities or social sciences, who use similar computational methods to create, analyse and disseminate a certain resource type of a digital resource".[36] These disciplines, however, function as "umbrella terms" under which several sub-disciplines lie. For this reason, this wide scope of disciplines was attempted to be specified through a focus on particular subject domains which would be considered as "prime candidates" to be supported by Europeana Cloud and Europeana Research.[37] In the case of the humanities domain, the characteristics of this selection were for sub-disciplines that a) have a wide variety of types of (re)sources used in their research and b) are distinct from each other. In the case of the social sciences domain, the selection was made for projects that have been awarded funding by national grants committees. Having defined the framework of researchers that will occupy the project, Europeana Research thus inherits this set of target audiences which is presented in Table 1.

Table 1: Target audience of Europeana Cloud and Europeana Research

---

[35] Europeana Cloud Deliverable 1.1 'Research Communities Identification and Definition Report', 2013, pp. 2-3
[36] Ibid., p. 4
[37] Ibid, p. 12

Humanities	Social Sciences
Archaeology	**(Social) Anthropology**
History	**Social/human/economic/ political/cultural geography**
Law	**Gender Studies**
Linguistics	**Economic and Social History**
Musicology	**Political Science**
Philosophy	**Sociology**

Source: Deliverable 1.1 'Research Communities Identification and Definition Report'

Chosen out of a list of sub-disciplines constituting humanities and social sciences, these "prime candidates" present differences and similarities regarding scholarly content use, their user needs and therefore their expectations of the new service. On the other hand, setting content at the core of the analysis, there is an obvious need for various resource types for researchers in both disciplines alike. Without ignoring the fact that some researchers or scholars might require particular, targeted material for their research, the diversity and variety of resource types in humanities and social science research is vital and a common ground for this audience. The combination of the resource types text, image, video, sound and 3D, all identified within Europeana, is desirable for their research. This allows for the ranking of content according to usefulness, significance and extent of frequency in projects and research outcomes. Researchers will be approached here by focusing on their individual disciplines, stressing in this way their distinct requirements in such a digital space.

2.2 Defining user information behaviour of the target audiences: demonstrating the use of digital scholarly content through background and empirical evidence

This section looks at the use of digital scholarly content and its influence on the research practice through desk research undertaken for the D1.2 State of the art report on digital research practices, tools and scholarly content and empirical research presented in D1.5 Case Studies Expert Forum Report (defining the typical needs of researchers who will use Europeana Research) and Research Communities Web Survey M1.3. It will mention the similarities and differences between humanities and social sciences with regard to content use focusing on scholarly content in general as well as Europeana content. Furthermore, it will evaluate, based upon case studies, the opinion of the researchers using Europeana and describe what their user needs are regarding Europeana.

A first step towards the understanding and documentation of scholarly content use was desk research conducted during the first months of the project. Since the literature hardly refers to the use of Europeana content by researchers, it was attempted to explore the subject through references to scholarly content use in general, in the context of information behaviour literature, by focusing on the types of content in Europeana, namely text, image, video, sound and 3D. These types alongside their subtypes guided the background research towards an overview on the use of digital content by researchers today and, in consequence, on the use of the types of content available in Europeana. The following table presents the subtypes of each resource type as identified in for the survey used for the content analysis of the Europeana datasets (see Table 2). Moreover, the literature which nurtured this desk research was mainly journal papers and books in the area of information behaviour published since 1990, as well as various European research projects and infrastructures alongside their deliverables such as DARIAH, CESSDA, NeDiMAH, CENDARI, EHRI etc.[38]

Table 2: Europeana content and resource subtypes

TEXT	IMAGE/3D	VIDEO	SOUND
Blue Print	Archaeological sites	Advertising film	Music
Books	Artefacts	Amateur film	Sound Effects
Calligraphy	Book bindings	Animated film	Speech
Chapbooks	Book plates	Compilation	
Diaries	Buildings	Documentary	
Documents	Cartoons	Educational film	
Encyclopaedias and Dictionaries	Ceramics	Experimental film	
Essays	Collotype	Feature film	
Field notes	Costumes	Loop	
Finding aids	Cultural property	Music video	
Folios	Daguerreotype	Newsreel	
Guidebooks	Digital images	Outtake	
Incunables	Drawings	Series	
Letters and Correspondence	Engravings	Short film	
Lyrics	Equipment	Trailer	
Magazines, Journals, Yearbooks	Furnishings	TV film	
Manuals (user guides)	Heraldry	Video installation	
Manuscripts	Illuminations		
Monographs	Illustrations		
Newspapers	Jewellery		
Pamphlets	Lithography		
Patents	Photographs		
Periodicals	Maps		
Programs	Metalwork		
Rare books	Monuments		
Registers	Mosaics		
Scripts	Musical instruments		
Text books	Paintings		

---

[38] For a full list of literature and research infrastructures initiatives on scholarly content use see Europeana Cloud Deliverable 1.2 'State of the art report on digital research practices, tools and scholarly content use', pp. 6-12 http://pro.europeana.eu/web/europeana-cloud/results/-/document_library_display/p6BV/view/2240207 (8 January 2015)

	Photographic plate		
	Photographs		
	Postcards		
	Posters		
	Prints		
	Sculpture		
	Sheet music		
	Sketches		
	Specimen		
	Stained glass		
	Studies		
	Technical drawings		
	Textiles		
	Toys		
	Watermarks		
	Woodcuts		
	Woodwork		

Source: Europeana – Dataset Information Form (Appendix A)

Relevant research mainly focused on rating the essence of each source, digital or analogue, for personal research while more recent articles engaged in a new discussion on the potential offered in the exploitation of this digital content, a question that mostly connects digital content to digital research practices. Focusing on surveys which examine the research practices of both humanists and social scientists, it is suggested that according to the resource type, or even the subtype, the information behaviour varies among these disciplines.

As far as image is concerned, as the largest resource type within Europeana[39], desk research showed that image concentrates overall more diverse behaviours. While a problematic fuzzy distinction exists whether there is reference to its digital or analogue counterpart, its rated value seems to be higher for humanists rather than for social scientists.[40] For example, a survey among academic researchers and postgraduate students conducted in 2001 in the United Kingdom recorded that photographs and still images are considered essential sources for 10% of the social scientists and 36% of the humanists.[41] Unlike image, text does not present such significant variations. It is overall rated as the most valuable source by both disciplines while e-journals are ranked higher in value and use. An interesting survey conducted in 2009 by the Research Information Network Report specifically for e-journals has characterised this type of content as "a huge success".[42] However, there are also variations documented on the essence of its subtypes according to the research needs.[43] The rest of the resource types, video, sound and 3D, are considered more specialised sources that mostly enrich the main bulk of information composed by texts and images. On the other hand, while they are not valued as significant sources, they are

---

[39] See paragraph 4.3.1 for a distribution of resource (sub)types over the Europeana portal

[40] Europeana Cloud Deliverable 1.2 'State of the art report on digital research practices, tools and scholarly content use', 2013, p. 17

[41] Education for Change Ltd, SIRU, 'Researchers' use of libraries and other information sources: current patterns and future trends' (Brighton: The Research Partnership; University of Brighton, 2002), p. 20. http://www.rslg.ac.uk/research/libuse/LUrep1.pdf (23 October 2014)

[42] David Nicholas, et al. *E-journals: their use, value and impact,* (London: Research Information Network, 2009) , p. 14 http://www.informationr.net/ir/reviews/revs337.html (12 January 2015)

[43] See conclusions in Deliverable 1.2 'Digital research practices, tools and scholarly content use', p. 17

increasingly perceived as sources for reaching innovative results. As a survey conducted in 2005 by the British Library recorded, researchers expressed a vivid interest towards sound and video material among the new possibilities offered nowadays by digital content and practices. [44]
Here lies an opportunity for the Europeana portal: if the video content in the portal grows, considering that it is "highly valued among historians", it will certainly add to the attractiveness of the portal for these researchers.[45]

In an effort to illustrate these points raised in the literature during the desk research, a new survey will be presented here aiming to add to the discussion and verify or negate points presented in Deliverable 1.2 of the project. This survey dates from 2009 through 2011, was conducted by the British Library and entitled 'Researchers of Tomorrow: the research behaviour of Generation Y doctoral students'.[46] Despite its somewhat restricted audience, it is extremely interesting that by combining qualitative and quantitative methods this survey "is to date the longest and most intensive study of research student activity in relation to information-seeking and research behaviour".[47] It should be acknowledged here that this survey examines content use through an information seeking perspective, referring "only to the external research information and sources that the doctoral students were looking for at a particular moment in time" without taking into account the process of creating and managing their own original data.[48] For this reason, it should be perceived here as indicative of the practices of younger scholars mainly regarding their content-wise "external" research.

Its sample focuses on young scholars born between 1982 and 1994 from a wide range of disciplines such as social sciences, arts and humanities, engineering and computer sciences, biological sciences, physical sciences, medicine dentistry and health, biomedical and veterinary sciences.[49] More specifically, the study contains 6,161 'Generation Y' students and 7,432 older students, a fact which enables comparison with the attitudes and behaviour among both age groups of researchers. This interdisciplinary demographic cohort allows for comparisons and observations regarding content needs of researchers of a wider disciplinary background allowing thus conclusions to be drawn specifically for humanists and social scientists.

---

[44] Jones K. Spark and R. Bennett, '*E-resources for research in the humanities and social sciences: a British academic policy review*', (London : British Academy, 2005), pp. 45-46 http://www.britac.ac.uk/policy/eresources/e-resources.cfm (12 January 2015)
[45] Deliverable 1.2 'Digital research practices, tools and scholarly content use', p. 16
[46] Julie Carpenter, 'Researchers of Tomorrow: the research behaviour of Generation Y doctoral students', *Information Services and Use* 32(1) (2012), pp. 3-17
http://www.jisc.ac.uk/publications/reports/2012/researchers-of-tomorrow (31 October 2014)
[47] Ibid, p. 9.
[48] Ibid, p. 19
[49] Ibid, pp. 13-15

Fig. 1 Type of research resource found and subject discipline
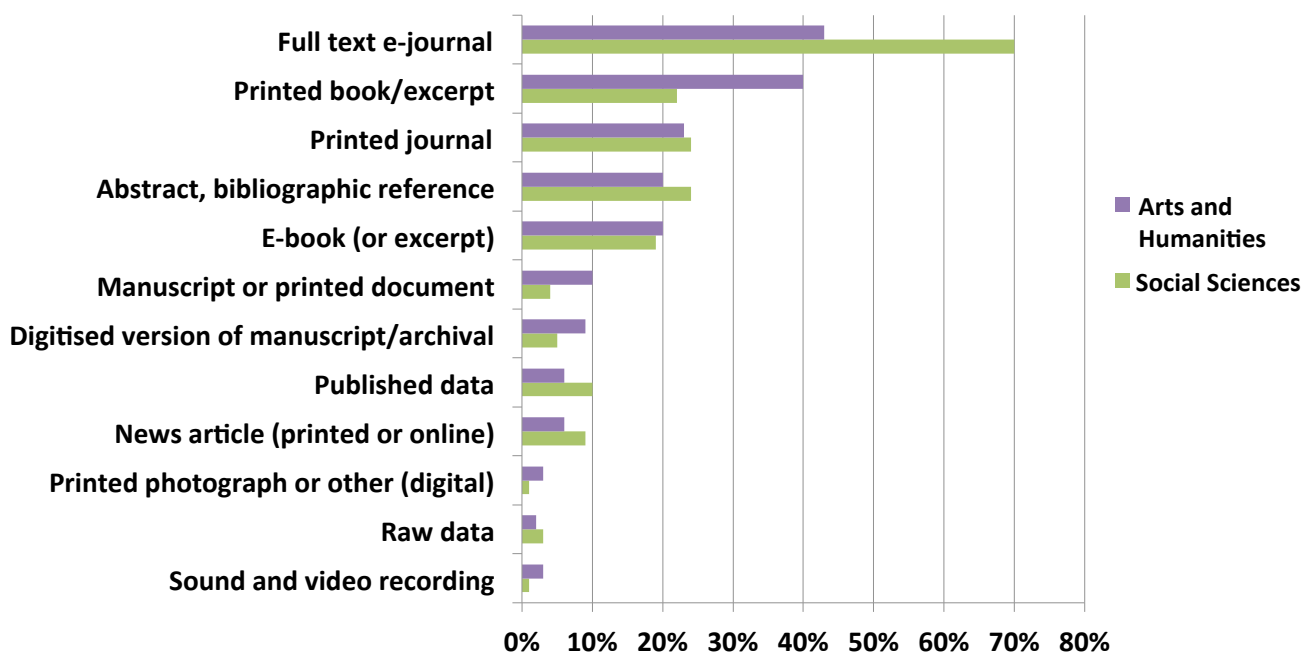
Source: Julie Carpenter et al., 'Researchers of Tomorrow' (2012), 21.

In the question on what students find of value, the data gathered from the responses of the survey sample suggested a - relative uniformity in the kinds of resources and materials sought and used - by doctoral students.[50] According to the survey "the overwhelming majority ended their information-seeking incident with a book (e-book or print), a journal article (e-journal or print), a reference or abstract of an article, and not primary or original source material such as data, photographs or newspaper articles, or archival material".[51]

Figure 1 illustrates this conclusion by demonstrating that text is the most valued type of content among both disciplines while the rated essence of its subtypes varies according to the research needs and research question.[52] This conclusion is not a surprise and it rather confirms the discussion presented before. The preference for content-type text expressed by researchers seems to conflict with the main focus of content within Europeana which is image. Audio-visual material on the other hand is indeed showcased as low rated in essence among researchers in the survey. Taking as starting point that Europeana is quite rich in all resource types, these observations do not necessarily interpret how researchers perceive and use its content, particularly if taking into account that these findings are mostly linked to non-primary data.[53] It could be said therefore that the most valuable source among researchers focuses on text; however, users access and use

---

[50] Carpenter et al., 'Researchers of Tomorrow', p. 20

[51] Ibid, p. 20

[52] This observation is also met in Deliverable 1.2 'Digital research practices, tools and scholarly content use', p. 17

[53] Definition of primary data for the purposes of this Deliverable report can be found in Deliverable 1.2 'Digital research practices, tools and scholarly content use', p. 13.

Europeana with different aims oriented by the type and wealth of content available there. What do researchers actually think of Europeana? How do they think that it could be developed in order to best meet their research needs? And, if possible, how do they use its content? These questions will be examined using the empirical research conducted during the course of Europeana Cloud.

Aiming to define user information behaviour of researchers regarding the use of digital content, a web survey was designed and implemented and Expert Forums were held which brought together researchers from both disciplines as well as people from the project for discussing tools and content in humanities and social sciences research. Both tasks constitute rich, and to an extent unique, sources of information while they will be enriched here by the Case Studies report on the actual use of innovative digital tools and services also produced in the context of the project.

During the summer of 2013 a Web Survey was conducted aiming to gather evidence-based data from the humanities and social sciences research community. This data would help in analysing their digital research practices, tools and content they adopt in their research, examining in particular the potential use of content from Europeana and The European Library within Europeana Cloud. Out of a total number of 339 people who answered the survey, 75% primarily consider themselves as humanists and 25% as social scientists. This sample does not encourage safe conclusions to be drawn on similarities and differences among these disciplines; it will however shape an overview of their digital research practices nowadays by analysing their replies separately.

In the question "how often have you used the Europeana portal in the last 12 months", the largest percentages concentrate on "never" and "a few times" for both disciplines. Even with minor significance, the percentage of humanists that use Europeana "a few times" exceeds the "never" percentage by 51% to 39% replies while for social scientists these percentages are 40% to 46% respectively. It is interesting to note here that The European Library has been accordingly used "a few times" during the last 12 months by 30% of humanists in contrast to 70% who have "never" used it while the percentages for social scientists are more evenly distributed among 56% who have "never", 25% "a few times" and, surprisingly, 19% "at least once a month" used this service. Considering that the largest resource type in the Europeana portal is image while in The European Library is text, these percentages could perhaps mirror the greater engagement of humanists with images, as documented earlier in the literature, in contrast to social scientists who value text higher. On the other hand, the fact that social scientists use The European Library more often than humanists contradicts with the content of this service which, if searching by discipline, is wider for the humanities.[54]

Focusing on questions that deal with the content, Figure 2 shows how researchers reacted to the question "assess the importance of the format of a resource as you seek relevant information for your research".[55] It seems that both disciplines significantly value the format of a resource (e.g. text, image, video, jpg, .pdf, etc.) from somewhat to extremely important with percentages reaching overall 86% and 94% for humanists and social scientists accordingly.

---

[54] See the webpage of The European Library, section 'Discover by Discipline' in http://www.theeuropeanlibrary.org/tel4/ (accessed 15 October 2014)

[55] The term 'format of a resource' corresponds to the term 'resource type' used before to analyse Europeana content. The term 'format' was adopted in the Web Survey in terms of familiarity to the audience.

Social sciences, however, seem to have a clearer overview of their content needs considering the balanced calibration of their answers' rates.

Fig. 2 Assess the importance of the format of a resource (e.g. text, image, video etc.) as you seek relevant information for your research
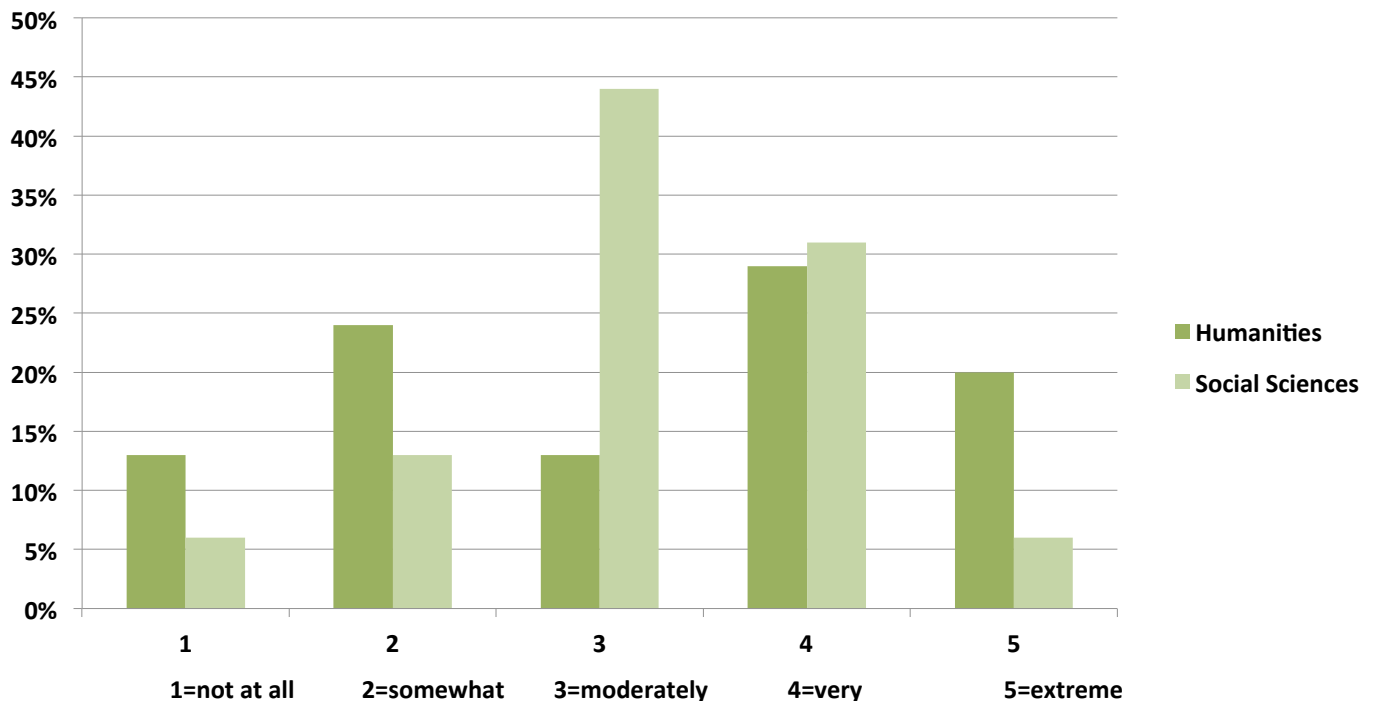
To the question "assessing the importance of the object type related to a resource" (e.g. oil painting, pottery, book etc.) humanists present a quite balanced fluctuation among all five answers primarily ranking these properties of a resource as moderately important while for social scientists "not at all" does not exist in the map. For them, the object type related to a resource seems to be mainly "somewhat" but overall important. Finally, in the question "assessing the importance of the subtype of a resource" (e.g. correspondence file, short story, landscape painting etc.), humanists have ranked these properties as mainly "somewhat" and "moderately" important while social scientists have been more 'generous' in evaluating them as "somewhat", "moderately" and "very" important. Interpreting these results in terms of tracing similarities and differences in the content needs of this target audience, it could be suggested that as all researchers engage with digital content they can all evaluate its significance according to their research needs and had thus an opinion when answering this Web Survey. While there is no information on what type of content they mostly use, as this is a question often addressed and answered in the literature, having available metadata of the resource, such as its subtype, its object type related to it etc., are considered essential information for using that resource by both disciplines alike while the extent of its significance varies. This need for metadata was clearly emphasized in the three Expert Forums conducted in 2013 which aimed at informing the Content Policy for Europeana Research. Through discussions held during these Forums between project members and external experts on the specific case of Europeana, this task of empirical research explored among other questions the usefulness of its content today and the research it can support in its current form.

Moving on to a brief presentation of the discussions held at the Expert Forums and the main suggestions formulated there, the first Expert Forum aimed to investigate the uses for Europeana

by researchers in both the humanities and the social sciences by developing case studies that would reflect their typical needs.[56] After completing a warm-up exercise, named Europeana Treasure Hunt, which allowed experts as well as project members to explore the Europeana portal having specific tasks to complete, the discussion that followed resulted in interesting feedback on the portal, its services as well as its content. These observations were summarised under the following broad subject headings:

- Insufficient Metadata
- Date Ranges not Recognized
- User Ranking
- Multi-lingual Resources
- Spatial and Temporal Mapping of Results
- More Transparent Citation Methods
- Tool Development
- Europeana as a Teaching Resource

These headings could be further distinguished into problems traced, even during the Europeana Treasure Hunt, and suggestions for future development of the portal. The first two headings refer to the first type of observations with "insufficient metadata" being characterized as one of the major factors that prevent Europeana to be useful as anything more than a discovery tool.[57] On the other hand, two of the headings referring to suggestions for the portal, namely "user ranking" and "multi-lingual resources" have also to do with the metadata issue. More precisely, participants of this Expert Forum suggested a "star-ranking system" so that users could rank the metadata provided taking into consideration the completeness and quality of the data. Another suggestion calls for a thesaurus or translation tool to be also built into Europeana which will automatically link multilingual metadata so as to enable researchers to find all items associated with a search term.[58]

The second Expert Forum conducted dealt with tools and content for humanities Research. Focusing on content, four principal questions guided the discussion, namely:

1. Is there content in Europeana that is currently useful? What content?
2. Does this content need to be improved on or added to Europeana?
3. What new content (subtypes, formats) would you like to see added to Europeana?
4. What are the biggest gaps in content in Europeana from a humanities perspective?[59]

These questions inspired interesting discussions among experts who reached several conclusions regarding the current status quo of Europeana among humanists. Some experts characterised Europeana as "already lending itself well for use in introductory teaching activities". This observation adds to one of the discussions held during the first Expert Forum for Europeana to strengthen its resources for teaching. Humanists, however, recognise this as a current service

---

[56] See Europeana Cloud Deliverable 1.5 (1) 'Expert Forum 1 Case Studies', p. 3. http://pro.europeana.eu/web/europeana-cloud/results/-/document_library_display/p6BV/view/1861918 (9 January 2015)

[57] Ibid, 8.

[58] Ibid. The translation tool has meanwhile been built in into the Europeana portal.

[59] Europeana Cloud Deliverable 1.5 (2) 'Expert Forum Tools and Content for Humanities Research Report', p. 10.

provided by Europeana stressing though the importance of the provision of better-quality metadata. References to poor metadata were unsurprisingly often made in this Expert Forum as well by experts pointing out that "at present it is not feasible to establish the portal's coverage of (Meta) data on any given subject".[60]

"Participants would like to see many more textual collections added" as "humanities scholars are mainly interested in digital texts".[61] This observation, recorded in the Expert Forum, brings in mind the desk research alongside its conclusions presented in Deliverable 1.2 on the types of content more highly valued among researchers. It is no surprise that humanists suggested the enrichment of textual sources as it is already suggested in the literature and the new survey presented before that both humanists and social scientists value text as the most significant source for their research. For this reason, among the questions that also came up and has been documented in this Expert Forum report is if Europeana aims to provide access to scholarly/scientific journal articles.[62]

"Europeana currently is not thought of as a research tool".[63] How can this be changed? Humanists alongside project members reached conclusions complying with earlier findings in the project and, more particularly, in Deliverable 1.1. These findings re-introduce the need for Europeana to develop enhanced calibration of the metadata of individual items as well as entire research collections with "relevant resource descriptors and identifications of possible deployment in humanities and social science research".[64]

The third Expert Forum on Tools and Content for Social Science Research was more or less developed having the same aims and structure. A main difference in the content-related suggestions provided by researchers here is for Europeana to implement a new strategy for prioritising a limited number of areas in order to have the highest possible quality, and metadata comprehensiveness, of the existing content.[65] This suggestion derives from the same need detected in previous Expert Forums as well for better-quality metadata. Listing this as one of the problems Europeana content faces today, experts agreed on ideas submitted in earlier project meetings for a user ranking system for metadata quality, for example.

A new addition in the discussion was searchability. Social scientists recommended several modifications to search refinements, so as to enable researchers to narrow their results, implementation of new tactics for smoother navigation among the different resource types and finally they stressed the importance of reliable search terms. Interestingly, contextualisation was also one of the issues that came upon among the participants emphasising the need for the content to be more contextualised rather than providing isolated pieces of content with sufficient contextual information.[66] This could be practically illustrated by offering researchers the function possibility to view relevant content based on previously viewed objects. Similar to the Expert Forum for humanities research, this project meeting also stressed the need to achieve a sufficiently high standard of metadata for its content adding another significant dimension.

---

[60] Deliverable 1.5 (2) 'Tools and Content for Humanities Research', p. 20
[61] Ibid, p. 18
[62] Ibid, p. 21
[63] Ibid, p. 17
[64] Deliverable 1.1 'Research Communities Identification and Definition Report', p. 15
[65] Europeana Cloud Deliverable 1.5 (3) 'Expert Forum Tools and Content for Social Science Research', p. 17.
[66] Deliverable 1.5 (3) 'Tools and Content for Social Science Research', p. 20

Taking into consideration that "researchers must always be able to evaluate their sources of data", "it would be important for Europeana to increase its transparency by providing clear information on who is behind the project, the provenance of the objects, and the selection policy involved".[67]

In an effort here to understand and document how researchers in the humanities and social sciences use and interact with Europeana content, and thus form suggestions for their needs in terms of content, in the preceding sections we followed the journey of Work Package 1 through desk research, case studies, Web Survey and Expert Forums to assess their needs. To sum up, desk research revealed that overall according to the resource type, or even the genre, the information behaviour varies among these disciplines. Especially the resource type 'image' concentrates more diverse behaviours while its rated value seems to be higher for humanists rather than for social scientists, a conclusion which is also drawn in the Web Survey. In addition to this, text is more equally valued with e-journals ranked higher while the rest of the resource types, video, sound and 3D, are attributed less significance as they could be characterised as specialised sources. A significant input of the Web Survey in this discussion is that while there is no information on what type of content researchers mostly use, a question often addressed and answered in the literature, having available metadata of the resource, such as its genre, its object type related to it etc., are considered essential information for using that resource by both disciplines alike while the extent of its significance varies. This conclusion sheds light to a final finding, the issue of metadata, and more particularly the lack thereof. Improving the metadata quality of Europeana content is one of the major recommendations underlined during all three Expert Forums urging attention in future Content Policies to be enforced by Europeana.

2.3 User profile of the digital humanities researcher and social scientist researcher

Background and empirical research has led to an initial formulation of user requirements of humanists and social scientists in using scholarly content, and in particular Europeana content. The needs and problems researchers face today will be synthesised here in an effort to shape user profiles of the digital humanities researcher and social scientist researcher. User profiles aim to illustrate the conclusions presented earlier in this report and contribute in the discussions held in the following chapters for the development of the Content Policy for Europeana Research.

Imagine now real users of this disciplinary background accessing the Europeana portal with a set of research goals and personalised needs. User A is an Art Historian conducting her PhD research on the art collection of an 18th century Dutch family.[68] She applies digital humanities techniques in her research and she is engaged in discussions on new upcoming digital tools that could help her in the analysis of content and, ultimately, in writing her dissertation. She mainly uses images for her research and therefore she visits several digital libraries and services, such as Europeana, to gather her evidence. She mainly accesses the Europeana portal to search for content on art and Dutch paintings. However, she is confronted with a wide number of results, most of them irrelevant to her topic. When she types in, for example, "Dutch paintings of 17th century", the search results include works of earlier or later periods as well. Apart from that, willing to cite the relevant retrieved

---

[67] Deliverable 1.5 (3) 'Tools and Content for Social Science Research', p. 22
[68] This user profile is inspired by one of the case studies explored in Europeana Cloud Milestone for Task 1.3.5 'Evaluating Use of Innovative Tools in Research', p. 22

content in her thesis, user A explores its metadata aiming to find more information and evaluate the source. She notes that the quality of the available metadata there is rather poor, a fact that makes her hesitant to cite Europeana as the source for the retrieved content. She is rather led in visiting the original institution that provided the content and its metadata to the portal asking for more information so as to acquire scientifically valid references. However, she considers this process as rather time consuming for her research. Textual sources are also significant in her work accompanying visual sources. For this reason, user A visits mainly repositories of e-journals aiming to find secondary literature that will serve as background evidence to her research.

User A adopts several digital tools to annotate and analyses the art collection and its textual sources. More particularly, one of the main digital tools that she uses is "HyperImage".[69] She accesses this tool from its webpage and it mainly helps her to "drag and drop" images for annotation.[70] She would also like to have the possibility to share this outcome with her colleagues for input and comments. This is a function provided by another digital tool that she uses on textual sources. This digital tool is called "Annotateit" and she uses it for "sharing annotation on online documents", "inviting her colleagues to help her scan through a series of relevant texts for the topics of interest for her essay and share their annotations, comments and tags with her".[71] She would also like to have access to a variety of digital tools via the same service allowing her thus to discover new tools as well as enable her to learn more on digital humanities research. This user could be characterised as a keen but non tech-savvy Humanist willing to produce new results by using digital content and adopting digital tools without being perfectly comfortable with all technological advancements.

User B is a Sociology PhD student researching social networks and their impact on diverse themes such as socioeconomic status, local energy governance and travel behaviour for example.[72] He is an enthusiastic user of new digital technologies applied in social sciences research and he thus explores all possible ways to discover, analyse and present the results of his research. Text is his core resource type while his research is constituted of mainly primary data. Regarding secondary literature, he often visits The European Library portal aiming to search for relevant publications that will support his research and enhance his knowledge on the outreach of social networks in literature as well. However, the problem that he often faces is that he has no full-text access to those sources, which hinders him when trying to gather his content.

Text provides him with the opportunity to use several digital tools to edit, annotate, publish, share and visually observe bibliographic metadata and publication flows, for example. Some of the tools that he adopts in his research are Pundit and TiNYARM.[73] The first tool allows him not only to annotate the texts, but also to look up various concepts and persons occurring in the texts, while TiNYARM helps him to share publications that he already read and publications that his supervisor suggested him to read as well as to know what other research groups are reading.[74] Apart from that, he also uses tools that enable him to get an insight in the topic of social networks and their

---

[69] T1.3.5 'Innovative Tools in Research', p. 19; http://hyperimage.ws/en/

[70] Ibid.

[71] Europeana Cloud Deliverable 3.1 'Personas, scenarios and use cases', pp. 9-10; http://annotateit.org/

[72] This user profile is inspired by one of the case studies explored in T1.3.5 'Innovative Tools in Research', p. 32

[73] For the tool Pundit see Deliverable 3.1 'Personas, scenarios and use cases', p. 40; https://thepund.it/. For the tool TiNYARM see Ibid., p. 37

[74] Deliverable 3.1 'Personas, scenarios and use cases', p. 10

adoption. Such a tool is More! which is a mobile web application showing different links to social networks that a speaker at a conference is actively participating in for checking the speaker's profile, his publications and recent tweets.[75] Finally, another specialised tool that user B adopts in his research is NodeXL which is a network analysis and visualisation tool that adds features and functions to the Excel spreadsheet.[76] His engagement with digital tools is not his only technological dealing. User B is a quite tech-savvy user, comfortable and familiar with a wide variety of digital practices who actively deals with API's as well exploring the potentials provided there for improving and conducting his research.

---

[75] Deliverable 3.1 'Personas, scenarios and use cases', p. 12
[76] T1.3.5 'Innovative Tools in Research', p. 30

3. Content analysis I – Introduction and The European Library

The availability of large-scale shared datasets and databases has changed modern-day research practices, enabling researchers to do collaborative research on important resources.[77] At the same time, researchers such as those in the CENDARI[78] project, which aims to integrate dispersed digital archives and resources for research on medieval and modern European history, acknowledge the difficulty of getting an overview of all the relevant information for their research.[79]

The European Library and Europeana portals both hold extremely large collections and datasets and therefore potentially enable the kind of collaborative research that is currently so attractive to researchers. To evaluate the usefulness of these datasets for research, a content analysis was undertaken with two main goals:

1. To match the existing content in The European Library and Europeana to the humanities and social sciences research communities and to reveal the potential value of these datasets for both communities;

2. To identify the gaps in the existing content.

Several issues hindered a thorough and uniform analysis of the datasets. In the first place, the content in the datasets was not aggregated using a rigorous content strategy,[80] which reduces the predictability of what one might expect in the databases, secondly, the quality of the metadata was not adequate to execute a complete and comprehensive content analysis, and thirdly the access to the metadata was not direct, but filtered through either the portal interfaces or an instrument like a survey.

No Content Strategy
For many institutions, a content strategy is vital. It determines what will be aggregated and how one set of content will be connected with other material in for example, the database, portal or service. This in turn is vital for researchers, who need to understand the context in which a particular item or collection of materials exist. The datasets in the portal of The European Library and Europeana, however, were not developed using a content strategy. They grew organically and came from a wide variety of providers and projects.[81] This inclusive approach is a fast and straightforward way to collect information, but the disadvantages are apparent in both the inconsistency of metadata and the type of content, which tends to lack the critical mass needed to enable research in certain subject areas. The datasets in The European Library and Europeana

---

[77] http://www.bl.uk/reshelp/findhelpsubject/busmanlaw/contentstrategy.pdf (6 August 2014)

[78] http://www.cendari.eu/about-cendari/ (6 August 2014)

[79] (…) the CENDARI project intends to produce a powerful platform providing scholars with online access to physically dispersed historical information. http://www.cendari.eu/about-cendari/ (13 November 2014)

[80] A content strategy describes why an organisation selects what it selects and it sets out the principles which will inform selection in future. Definition based upon: British Library, From Stored Knowledge to Smart Knowledge: The British Library's Content Strategy 2013-2015
http://www.bl.uk/aboutus/stratpolprog/contstrat/british_library_content_strategy_2013.pdf

[81] Europeana gathers content via national, regional and project-based aggregators, thus providing access to all cultural content offered. The European Library aggregates datasets offered by 48 national libraries of Europe and leading research libraries.

portal offer a great variety of possibly attractive content but it is not always available in great quantities and it is difficult for users to have an overview of what is available and to select items of interest. Steps have been taking to address this issue by grouping information together via thematic and generic projects like Europeana Fashion[82] and Europeana Travel[83]. On the whole, however, it remains difficult to execute a content analysis on a detailed level.

Metadata consistency and quality

Another issue hindering the content analysis is the quality and inconsistency of the metadata. Metadata inconsistency is a common issue with digital libraries, as was for instance reported by Zavalina: 'digital libraries aggregating metadata from different sources inevitably face problems with metadata consistency (…)'[84] One of solutions for this problem was provided by the participants of the first Expert Forum who recommended that Europeana should encourage its data providers to 'provide complete metadata in future'.[85]

The quality of the metadata certainly is an issue. This was for example illustrated by the participants in all three Expert Forums, who strongly recommended improvement of the metadata in the Europeana portal.[86] The content analysis was for example hindered when we aimed using the subject terms for the analysis.

Searching for items using subject terms in both portals is almost impossible because most of the items in both The European Library and Europeana are not indexed on subject[87]. The reasons for this are varied; sometimes they were not indexed on an item-level by the data providers.[88] In other cases, the cataloguing methods of the data providers do not comply with the aggregation procedures of The European Library and Europeana, for instance when a text value is provided rather than a URI to a vocabulary or when abbreviations are used by the cataloguers. These will probably be in the natural language of the cataloguer, making it difficult to translate.[89]

This makes it difficult to assess which subjects are covered by which datasets. It also makes it impossible to determine *how many* items relevant to a certain subject are present in any given

---

[82] http://www.europeanafashion.eu/portal/home.html

[83] http://www.europeanatravel.eu/

[84] Oksana L. Zavalina, 'Free-text collection-level subject metadata in large-scale digital libraries: a comparative content analysis', International Conference on Dublin Core and Metadata Applications, DC-2011, The Hague, 21-23 September 2011, p. 148. http://dcevents.dublincore.org/IntConf/dc-2011/paper/download/50/19 (3 June 2014)

[85] Deliverable 1.5 (1) 'Expert Forum Case Studies Report', 2013, p. 13

[86] All three Expert Forums can be found at: http://pro.europeana.eu/web/europeana-cloud/results/-/document_library_display/p6BV/view/1861918 (9 January 2015); See also paragraph 2.2

[87] 'Indexed on subject' is here used in the sense of making a connection between an item and a subject by attributing controlled-vocabulary metadata to the item, drawing value from a formally-maintained (controlled) list of terms. This in turn allows for searching using controlled subject terms.

[88] Oksana L. Zavalina, 'Free-text collection-level subject metadata in large-scale digital libraries, p. 147

[89] See paragraph 4.2

dataset. As researchers often search using subject terms or key words[90], this diminishes the worth of the digital portals for research purposes.[91]

Indirect access to the data

The analyses of both datasets was executed by a variety of means, such mapping content against terms in a survey, analysis of textual descriptions of collections, or making use of the indexes available in the contexts of both portals. It was not possible to use algorithms and computational powers because of the inconsistency in the metadata. This indirect form of access to the data means that the analysis was both subject to human interpretation (when using the surveys and collection descriptions) and shaped/limited by the interface and indexes available in the portals.

We will now look at these issues in more detail, by examining the datasets of each portal separately.

3.1 The European Library

At the time of writing, The European Library portal enables users to cross-search over 24 million pages of full-text content, 18 million digital objects and 119 million bibliographic records, spread over 608 collection descriptions, each containing one or more datasets.[92] At the end of 2013, the Central Index of the portal contained 125 million records, up 17% from 2012.[93]

Performing a content analysis of this database is daunting, not only because of its size but also because the collection descriptions in the portal are very heterogeneous. Groups of datasets can be described as a whole or separately. Sometimes a collection level description refers to a group of datasets while only one dataset is online available. The collection level descriptions are also uneven and range from a textual description of a catalogue to bibliographic records, digital items or a combination of data types.

The textual descriptions of catalogues in The European Library are not indexed in its portal, making it difficult to retrieve them. Researchers can only browse the collection level descriptions by accessing them through a special menu option 'discover' followed by 'browse collections'.[94] Furthermore, only some data has been indexed by subject. This means that although The European Library unites diverse resources from 48 national libraries and some research libraries, researchers are not provided with a transparent overview of all the information held in the portal. The research value of The European Library is therefore diminished because it is precisely the gathering of information from diverse resources which enables research. The European Library portal allows for a 'discovery by discipline', which might be a helpful tool in this case.

---

[90] Europeana Cloud Research Community Web Survey T1.3.4, 2013:, Q4_ Rate the level of importance of using the following entry points, or attributes, as you seek relevant information for your research; Q6_ Assess the importance of the following properties of a resource as you seek relevant information for your research; Deliverable 1.2 'Desk research into the state of the art on scholarly content use', p. 31

[91] The preferred search strategy of post graduate students and professors is measured as by subject (47%) and free text (46.8%). In: L.M. Hughes (ed.), *Evaluating and measuring the value, use and impact of digital collections,* (London, UK: Facet Publishing, 2012), pp. 90-91

[92] http://www.theeuropeanlibrary.org/tel4/aboutus (22 October 2014)

[93] The European Library – Business Plan, p. 14

[94] http://www.theeuropeanlibrary.org/tel4/discover (3 December 2014)

In the next section, the methodology used for the content analysis is described, followed by the analysis itself and text and figures.

3.2 Methodology

The following content analysis of the datasets aggregated by The European Library is based on a 2013/2014 document prepared by staff at The European Library. It contains detailed collection descriptions for about 60% of the 608 collections in the portal (368 as of May 2014), plus important additional information such as subject, period, language and country coverage. The terminology used to determine the resource types and subject terms in the document corresponds with the terminology used in the Europeana 'Dataset Information Form'.[95] This form was also used for the Europeana survey on which the analysis of the Europeana dataset is based (see Chapter 4).

The document was completed in May 2014. At the time of writing of this report, 24 datasets had been removed or suspended from the portal on request of the providers because the content became outdated. These datasets have been left out of the analysis, leaving 344 datasets to review.

---

[95] See Appendix A

As noted above, there is a great variety of material in The European Library portal:

	Type	Description	Number of collection level descriptions
A	Catalogue(s)	Textual description of the content of catalogue(s)	203
B	Online catalogue	Description of the content of the catalogue and search box to search for bibliographical records	7
C	Catalogue and hybrid dataset	Combination of datasets partly consisting of a catalogue containing bibliographical records and partly of digital items	19
D	Dataset containing metadata with digital items	Dataset exclusively consisting of digital items	115
Total:			344

Table 3 Variety of material in The European Library

A. Catalogues

Of the 344 collection level descriptions available for analysis, 203 are nothing more than textual descriptions of catalogues. So, in The European Library portal, a majority of collection level descriptions do not link directly to bibliographical records. For Europeana Research, it is recommended that better use be made of these textual descriptions to gain access to this wealth of information.

Since through the collection level descriptions via The European Library portal, the bibliographical records could not be identified, these have of necessity been excluded of this analysis. Therefore, of a reported 1.25m records in 2013 (see 3.1), an estimated one third are ultimately covered by this analysis.

C.-D. Datasets holding online available items

The 7 datasets in B can be searched online, but do not contain online available items; they have therefore been left out of this analysis. The other datasets in The European Library portal holding online available items, C-D (115 + 19 = 134) contain either digital images or full-text items. These will probably satisfy the information needs of social scientists and humanities scholars and, as

convenience and speed are important reasons for academics to consult digital resources.[96] The convenience of having the research material directly available, without having to consult a library, is highly appreciated by researchers. In other words, direct access to material has become as important as the discovery experience.[97]

This coincides with a recommendation for the Europeana portal to move 'from exploration and discovery to in-depth descriptions and interconnectedness'.[98] Levels of satisfaction for the humanists and social scientists who stumble upon the datasets, containing a combination of digital items and bibliographic records or only bibliographic records will probably be lower, despite the fact that there is some value in being made aware of items that one previously did not know existed.

Given the vast levels of digitisation still to be done, it is likely that it will not be possible to offer access to digital objects for all library collections in the immediate future. A 2014 survey on digitisation by ENUMERATE concluded that just 12% of collections in Europe's libraries had been digitised. An earlier 2012 survey done by ENUMERATE showed that only 4% of national library collections were digitally available, compared to 12-14% for higher education, special and other libraries.[99] This means that both collection-level descriptions and bibliographic records will continue to be important finding aids for researchers. 'The transition to digital is by no means finished yet' as is illustrated by Elsbeth Kwant by picturing the history of Western textual communication so far on a twenty-four hour clock, showing only the last twelve hours (the previous twelve being manuscript age as well). This figure shows we have spent not more than half an hour on the digitisation of metadata and about ten minutes on the digitisation of content (images and text):

---

[96] Research Information Network, Reinventing research? : information practices in the humanities (London: Research Information Network, April 2011), p.71; http://www.rin.ac.uk/our-work/using-and-accessing-information-resources/information-use-case-studies-humanities

[97] Lynn Silipigni Connaway and Timothy J. Dickey, The Digital Information Seeker: Report of the Findings from Selected OCLC, RIN, and JISC User Behaviour Projects, (Bristol, UK: JISC, 2010) http://webarchive.nationalarchives.gov.uk/20140702233839/http://www.jisc.ac.uk/publications/reports/2010/digitalinformationseekers.aspx (9 January 2015), p. 18

[98] Deliverable 1.5 (2) 'Expert Forum Tools and Content for Humanities Research Report' (December 2013), p. 3 http://pro.europeana.eu/web/europeana-cloud/results/-/document_library_display/p6BV/view/2240207 (28 August 2014)

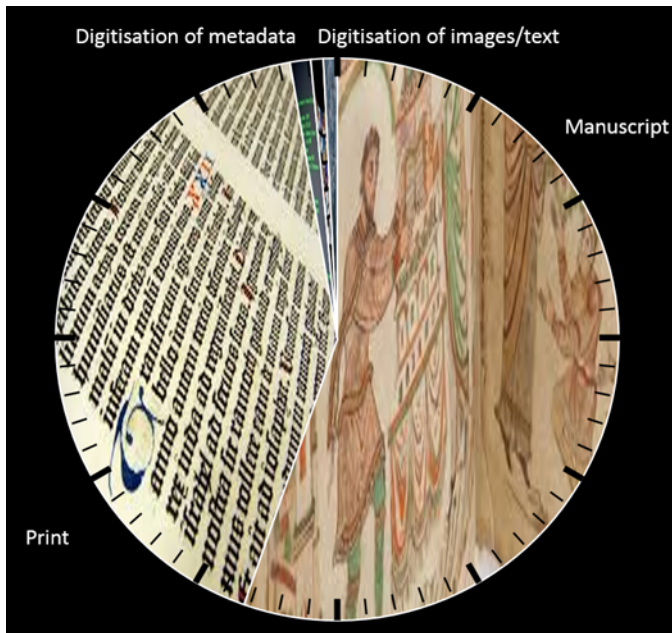[99] http://www.enumerate.eu/en/statistics/ (6 August 2014)

Fig. 3 Parchment to portal, textual communication in the West
(Source: Elsbeth Kwant, 'Changing roles between authors and readers')[100]

3.3 Visual representation of the content analysis

In this section, the content in The European Library portal (e.g. items available online and datasets) will be matched to the needs of various research communities. The analysis concentrates on items which are available online because these can be retrieved directly by searching in the portal. This therefore creates a more realistic picture of the content of the datasets available to researchers. Consequently, the term 'datasets' in the following sections refers to datasets holding items available online. Gaps in the content will also be identified.

Size	Number of datasets	Number of items
Very Large	13	10 mil+
Large	9	100,000-10,000,000
Medium	55	1000-100,000
Small	64	1-1000
Total number of datasets	128	

Table 4 The 128 datasets holding online available items

---

[100] Elsbeth Kwant, *Changing roles between authors and readers: a framework for thinking about change in the creative communication chain*, 30 November 2014, n.p.
http://www.researchgate.net/publication/269571005_Changing_roles_between_authors_and_readers_-_a_working_paper_%28for_discussion%29 (15 December 2014)

The content analysis focuses on the relative distribution of the categories a) resource type, b) country, c) period and d) subject over the range of items in the datasets which are available online. The results of the content analysis are presented in text and figures. On the basis of the collection-level descriptions prepared in The European Library's own report, only a relative distribution of content in these categories can be produced. Datasets may contain content from more than one category and it cannot be determined from the collection level descriptions (nor through a review of the datasets in the portal), exactly how many items in each dataset may be attributed to a specific resource type, country, period or subject.

If, in a collection-level description, an indication of resource type, country, period or subject occurs, then the *total number* of items in the dataset is counted. If more than one indication is encountered, the items in the dataset are added to the total for *both*, so the items may be counted more than once in these figures. This is the reason that the total number of occurrences of indications of resource type, country, period and subject are not given, but are instead presented relative to one another. This approach is described as 'relative distribution'.

3.3.1 Relative distribution of resource type among items available online

Resource types – broadly speaking texts, images, sounds, videos and 3D drawings – are not indexed and cannot be searched through the portal of The European Library. Instead, they are analysed with the aid of more refined categories such as pamphlets, illustrations and drawings. All the resource types are listed in the Europeana survey[101].
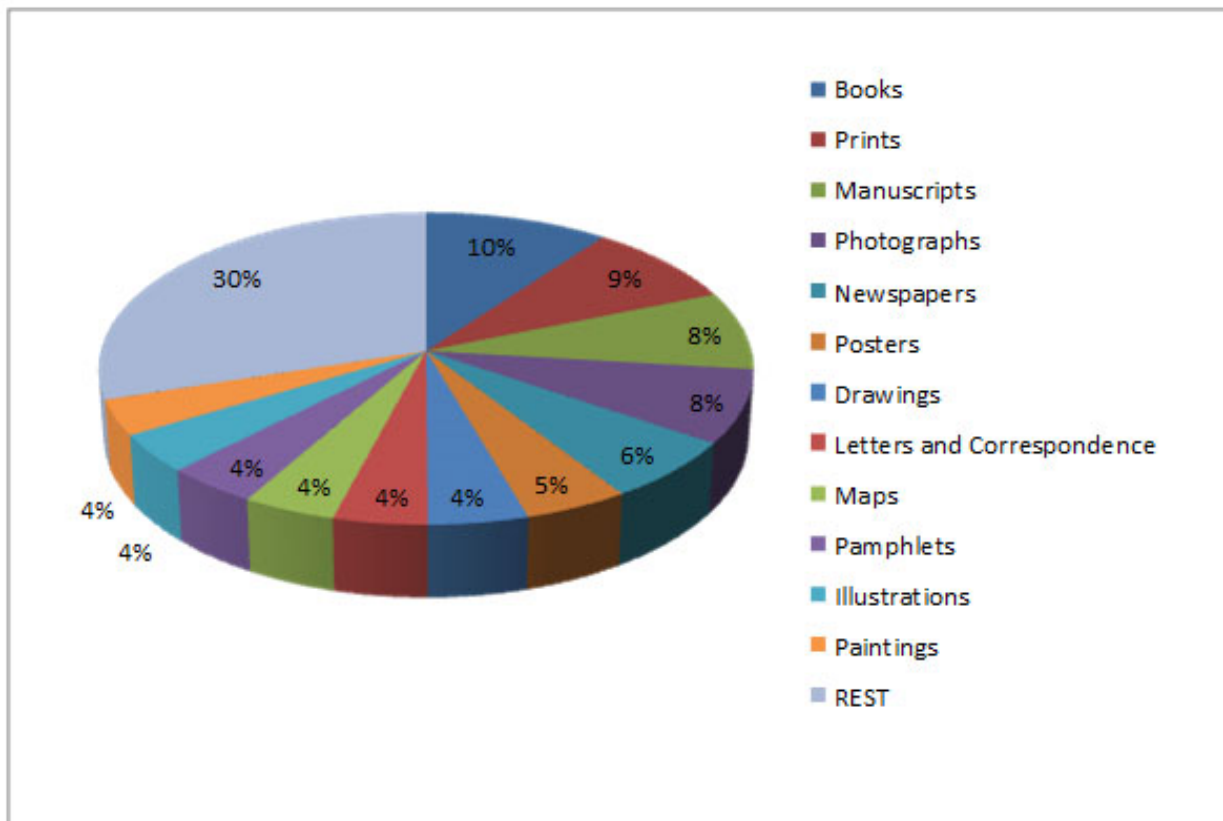
---

[101] See Appendix A

Fig. 4 **Resource types**: 12 most commonly occurring resources types relative to all occurring resource types

This graphic illustrates the 12 most commonly occurring resource types.

As might be expected from a library aggregator, books and manuscripts make up at least a fifth of the online available items and a further 40% relate to visual materials, such as prints[102], posters, drawings and photographs (which in themselves make up almost 10% of the collection).

Newspapers are relatively well represented in the collection (9%). This is understandable, because many libraries have already started to digitise their newspaper collections and The European Library has ingested the data from the Europeana Newspapers project in its portal.

Posters (5%) are the sixth category of resource type in the portal. A more detailed description of some of these collections of Posters can be found in section 4.1.3.

Music is notably missing from the 12 most common resource types. This might seem odd since music is the most common subject term (see Section 3.3.4) but this discrepancy is easily explained by the fact that the subject term music covers a large number of items (e.g. a book about music, a manuscript containing music notations, documents of a symposium). Music as a resource type,

---

[102] The resource type 'print' refers to: 'images formed by transfer from one surface or source to another. Usually created with ink(s) and produced in multiple impressions';
http://www.loc.gov/pictures/collection/tgm/item/tgm008237/ (21-08-2014)

however, indicates three different broad categories of resource type: sheet music (part of the image category), music videos (category: video) and music (category: sound).

Looking at the large amount of subject items covered by music, together making up one large collection, musicologists interested in music (as opposed to material about music as a resource type) will find a reasonable amount of items in their field of interest. Depending upon the point of view of the musicologist, he or she could experience the relative absence of the resource type music as a gap in the collection of datasets.

Earlier research in the Europeana Cloud project concluded that researchers in the subject domains history, law, linguistics, musicology and philosophy engage to a certain extent with all resource types listed in Europeana.[103] As we have seen, traditional text-based information is well-represented in The European Library portal, as are images (e.g. prints, photographs and illuminations). Digital humanities researchers wanting to combine research using the resource types text and image will therefore find good possibilities in the portal. Also noted from earlier research was a preference of both digital humanities and social sciences research communities for full-text content.[104]

For social sciences research communities, The European Library portal offers the same relevant resource types as for the digital humanists, namely text and image. Since for these researchers the content of the data and the tools to assist in the analysis of data are of particular importance[105], the availability of detailed metadata is crucial.

Looking at specific resource types, video, sound and moving image datasets are present.

- There are many videos in particular but they are spread among few datasets, which might indicate a large research corpus with little variety.
- Sound and moving image are only available in very small quantities and in few datasets. This can be considered as a gap in The European Library portal.
- 3D objects are not represented at all, which means that the needs of archaeologists (who more and more often engage with 3D objects) will also not be met by searching in the portal.
- Five datasets adhere to the type diaries, together making up one medium-sized dataset. For social scientists interested in this resource type, the research material is available in a reasonable variety but in a small quantity.
- Posters and maps are available in few datasets, but these datasets could prove very useful for research as they may contain unique material, which could be difficult to find elsewhere.

---

[103] Deliverable 1.1 'Research Communities Identification and Definition Report', pp. 12-14 and associated Excel-files, 'D.1.1_Communities_Table_SocialSciences', 'D.1.1_Communities_Table_Humanities'
[104] Full-text: The complete work in either print, electronic, or microfilm format. http://www.bc.edu/libraries/help/howdoi/howto/glossary.html (4 September 2014)
[105] Deliverable 1.1 'Research Communities Identification and Definition Report', pp. 11-12

3.3.2 Geography: relative percentage of total dataset referring to specific countries

In this section, we look at the geographical distribution of digitally available items on The European Library. By this we mean countries to which datasets refer, rather than the countries from which the datasets have been delivered. We include historical empires in the list, such as Austria-Hungary, as well as modern-day countries. Unlike the data's country of origin, which is indexed in The European Library portal, the information to which the collections *refer* to is not indexed, and not ever analysed before.

It must be noted that the evidence for this analysis is limited, as not every dataset can be linked to a specific country.[106] However, where references to geographical coverage were available in the detailed collection descriptions, these were recorded and analysed below. Ultimately, 55% of the collection metadata in the portal contained information about geographical coverage.[107]

---

[106] See paragraph 3.2 Methodology

[107] Oksana L. Zavalina, 'Free-text collection-level subject metadata in large-scale digital libraries, p. 151
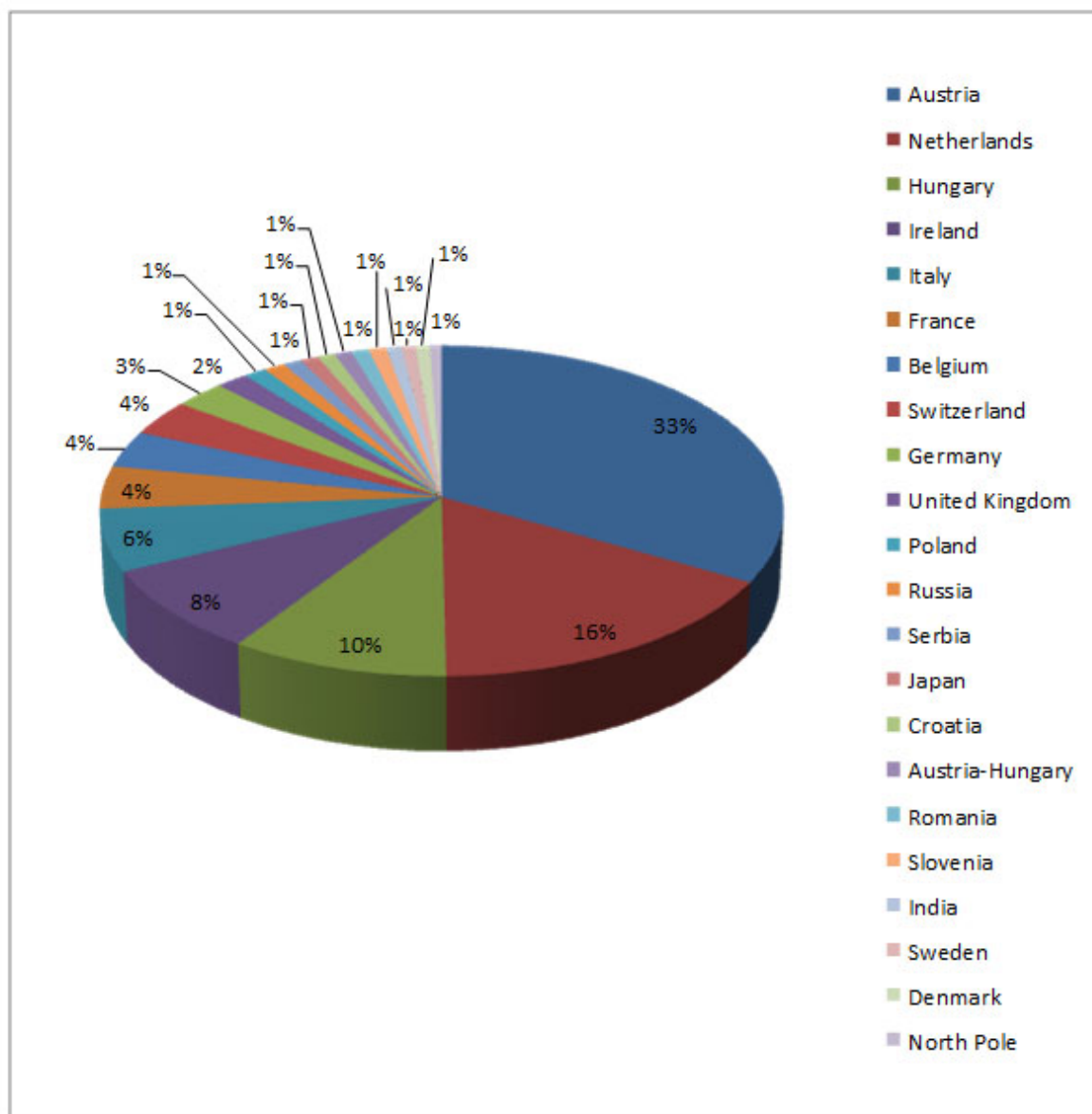
Fig. 5 **Geography:** relative percentage of total datasets referring to specific geographic regions

The European Library portal references over 32 geographic regions, mostly modern European countries but also countries such as Japan and India and empires of the past such as the Carolingian and the Ottoman Empires.

The vast majority of items cover Austria and the Netherlands, followed by Hungary, Ireland and Italy. As Austria, Hungary, Ireland and Italy are relatively large European countries; this is not surprising; the second place for the Netherlands is, despite its much smaller size, striking. The Europeana portal also contains a considerable amount of items from the Netherlands (see Fig. 9). Countries and geographic regions covered by less than 10,000 items are not represented because they represent less than 1% of the total number of online searchable items. From the Eastern European countries only Hungary is well represented; not much material has yet been digitised from Eastern Europe, potentially this is a valuable resource for digital information. The European Library could possibly stimulate the digitisation efforts in these countries by intensifying the cooperation, thus filling a gap in the portal.

Portugal and Spain, one of Europe's largest countries, are not represented. Unfortunately, some collections from the National Library of Spain have been removed on their request in the past few months, leaving a gap in the portal. Intensifying the cooperation with the national libraries of Portugal and Spain may also lead to filling this gap.

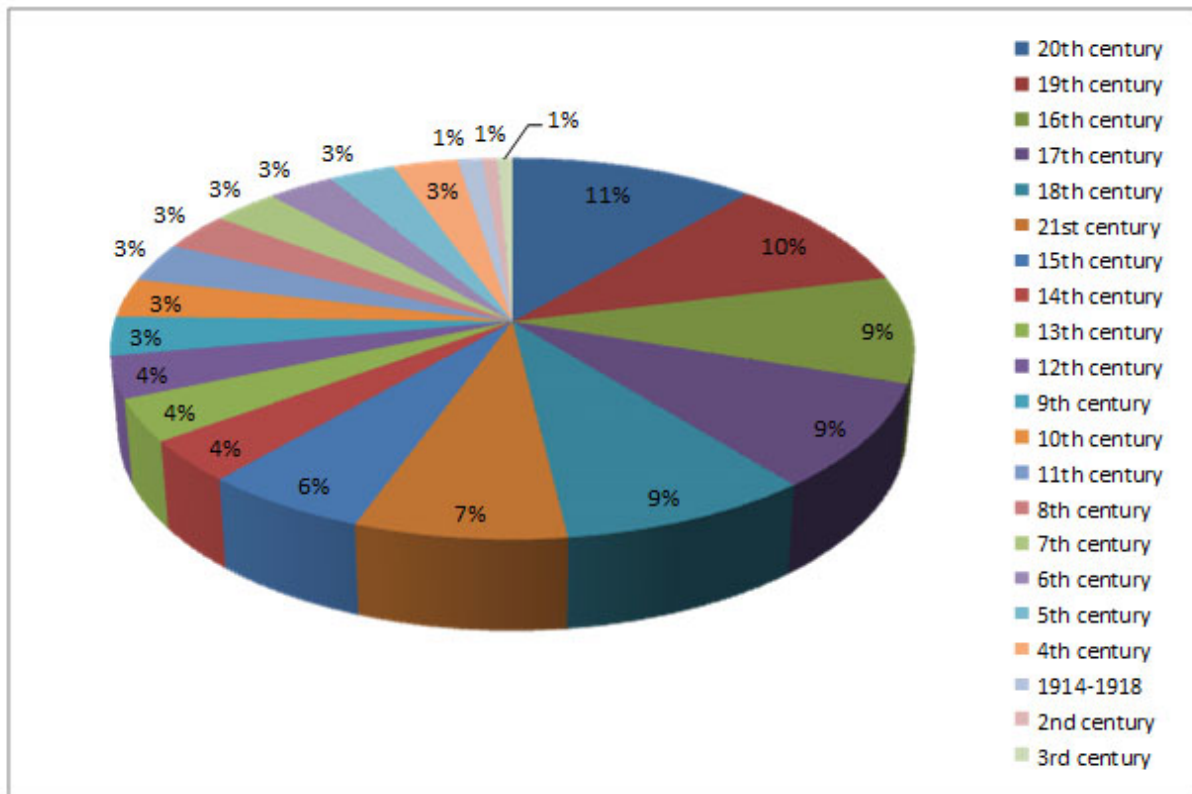### 3.3.3 Chronology: temporal distribution of online available items

Fig. 6 Chronology**:** temporal distribution of online available items in The European Library portal

When measuring the chronology in The European Library portal, we have to take into account that not all online available items are dated; the temporal coverage in the portal was measured at 63% in 2011.[108] This graphic shows that most online-available items cover the twentieth century, closely followed by the nineteenth and the sixteenth to eighteenth centuries. Part of the twentieth century items relate to World War One, due to the centenary of the War in 2014 and the participation of The European Library in the Europeana 1914-1918 project.

Largely due to the copyright limitations put on modern materials, the projects involved in digitising twentieth century materials are relatively scarce. Sourcing 20th century material is challenging because of what is sometimes called the '20th century black hole'; primarily due to copyright protection.[109] Part of this 'black hole' can be filled in by aggregating user-generated content such as Europeana did with the *Europeana 1914-1918* project.

Surprisingly enough this seems to be coupled with a lack of interest in studying twentieth century material on the part of the digital humanities community.[110] Still, considering the enormous

---

[108] This figure is derived from: Zavalina, Free-text collection-level subject metadata in large-scale digital libraries, p. 151. A more recent figure is not available.
[109] Chapter 3, pp. 13-14; Chapter 4, p. 55
[110] Alastair Dunning*, 'The Great Twentieth-Century Hole Or, what the Digital Humanities Miss',* DH Benelux, June 2014 https://availableonline.wordpress.com/author/xcia0069/ (26 August 2014)

increase in print production in the nineteenth and twentieth centuries and considering that at least part of that material (e.g. publications commissioned by public companies and governmental bodies) is not subject to copyright, these centuries seem to be underrepresented in the dataset of The European Library. European libraries typically hold large physical collections in the form of manuscripts and printed books for the fourteenth to eighteenth centuries and these have long since been the subject of digitisation projects. This is evident in the graphs above, where each century is more or less equally represented. The pre-medieval period is considerably less represented, as can be expected because there is far less material left from this period to be digitised.

3.3.4 Subject coverage: distribution of subjects over the online available items in the datasets
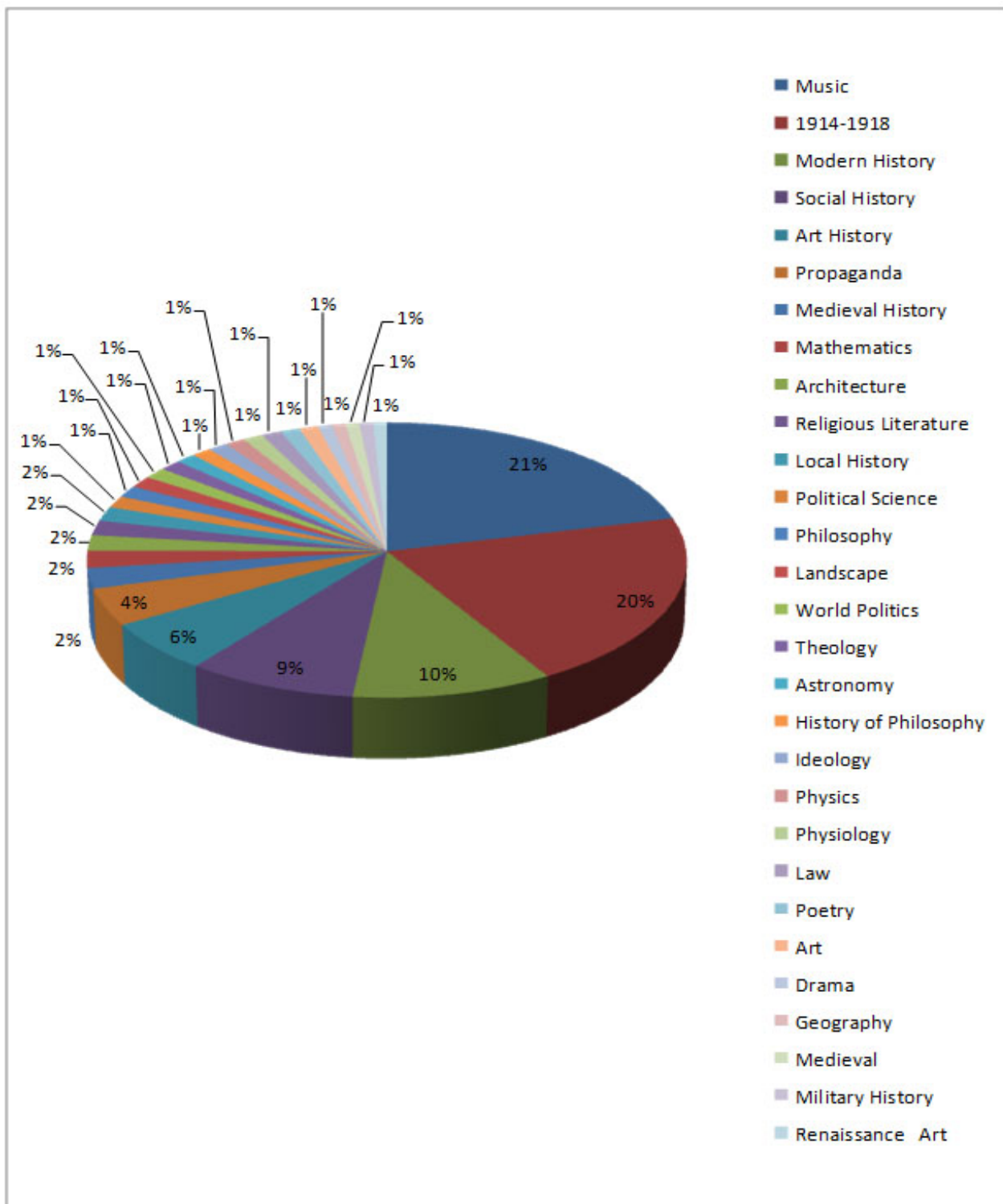

Fig. 7 **Subjects** covered by at least 1% of the datasets

This figure depicts the subjects covered by datasets that represent at least 1% of the total of digitally available items in the portal (i.e. large and medium datasets holding at least 24,000 items).

Music and 1914-1918 make up a large part of the collection of datasets. The relatively large part of the dataset covering propaganda (4%) probably correlates with the extensive collection of material related to World War One (1914-1918) in The European Library portal.

Earlier in the project, six subject domains in both the humanities and social sciences were identified as the target audiences for Europeana Cloud and Europeana Research:[111]

Table 5 Target audiences for Europeana Cloud and Europeana Research

Humanities	Social Sciences
Archaeology	(Social) Anthropology
History	Social/human/economic/ political/cultural geography
Law	Gender Studies
Linguistics	Economic and Social History
Musicology	Political Science
Philosophy	Sociology

Source: D1.1 'Research Communities and Definition report'

Using table 5, the domains were matched to the subject categories. The following conclusions were drawn for matching the datasets to the **humanists:**

- The subject term Archaeology does not exist and it would seem **Archaeologists** will not easily find relevant material in the datasets (although relevant material may be lurking in the Local History, Medieval History, and Landscape categories).
- For **Linguists** only three datasets are marked with that particular subject term, making up a very small portion of the total content (smaller than 1%).
- **Historians** are more likely to find interesting research material, depending on their chosen subject. Considering the history-related subject categories 1914-1918 (20%), modern history (10%), art history (6%), medieval history, local and military history (all under 2%), are all divided over a considerable number of datasets, this leads to a large quantity of research material with much variety. This is not the case for researchers interested in art history, as this subject is covered by one large collection, bringing together a large quantity of research material with less variety.
- **Law** makes up 1% of the total datasets, about 35,000 items, divided over eight datasets, a medium-sized collection of datasets with large variety.
- For humanities scholars interested in **philosophy** and the history of philosophy, 2% of the datasets are available, together making up a medium-sized dataset of about 80,000 online available items. A considerable amount, with much variety, spread over 19 datasets in total.

To summarize, historians and musicologists are most likely to find relevant research material among the described datasets and, to a somewhat lesser degree, philosophers. Working with this material in practice using computational tools and methods has proven to be challenging, as was

---

[111] For more information on the subject domains see Chapter 2, pp. 1-2

concluded by studying one group of philosophers.[112] Some research potential exists for humanities scholars interested in law, but there is room here to strengthen the offering by increasing the variety of available datasets and the number of digitally available items. If The European Library wants to serve all of its pre-defined subject domains, it must acquire a considerable amount of data in a variety of complementary datasets covering subjects such as Archaeology and Linguistics.

For **social scientists** the relevance of the described datasets is not always easy to determine. It depends on the subject domain in which the scientists work.

- For the subject domain of **Geography** the datasets hold a small quantity of research data, about 25,000 items, divided over six different datasets, in total making up one medium dataset.
- Few datasets cover the domains **Gender Studies** and **Economic History**. One large dataset covers **Gender Studies** but this subject category as well as the category Economic History needs to be built up further.
- The subject domain **Social History** (9%) is well covered by datasets, providing a large amount of online available items divided over eight large datasets, resulting in a large quantity of research material with a large variety.
- The subject domain of **Political Science** (1%, 4 datasets), together with **World Politics**, **Ideology** (1%, 3 datasets; 1%, 1 dataset) and **Propaganda** (4%, 6 datasets) in total make up 7% of the datasets divided over 14 datasets, which is a considerable amount with research potential.
- For the two remaining subject domains of **Social Anthropology** and **Sociology**, matching the datasets to the interests of the researchers is complex. A search for the subject term social anthropology offers no separate datasets about this subject but searching the portal for the same term generates around 6,500 items. This is little compared to the material available on other subjects. It has proven difficult to exactly determine what search terms social anthropologists and sociologists would use. As sociologists and social anthropologists deal 'with all types of human endeavours and cultural expressions' they might find interesting material among the datasets[113] but they will have to be creative in their search strategies to locate materials relevant to their research.
- Some subdomains in Sociology might offer relevant research material, for instance Sociologists interested in Religion will find relevant materials, as around 4% of the datasets cover subject areas like Religious Literature (2%), Theology (1%), Buddhism, Christianity, Hinduism, Islam, Islamic law and Judaism (1%). In total these subjects are covered by 19 datasets holding religious literature, so a large variety with small quantity of research material, 15 datasets covering theology and 14 datasets each covering the diverse religions. Adding these up creates a large dataset with a large variety of research material.

---

[112] Hein van den Berg, Gonzalo Parra, Anja Jentzsch, Andreas Drakos, and Erik Duval, 'Studying the history of philosophical ideas: supporting research discovery, navigation, and awareness'. In: *Proceedings of the 14th International Conference on Knowledge Technologies and Data-driven Business* (i-KNOW '14). ACM, New York, NY, USA, Article 12 , 8 pages. DOI=10.1145/2637748.2638412 http://doi.acm.org/10.1145/2637748.2638412 (20-11-2014)

[113] Stefan Ekman of the University of Gothenburg has consulted several colleagues active in the subject domains social anthropology and sociology to determine what search or subject terms they would use to search for material in The European Library portal. The outcome was 'there is no limit to the search terms they would use, as they are interested in all types of human endeavours and cultural expressions'.

- Sociologists interested in (popular) culture may also find relevant materials, as the portal contains a considerable number of datasets and online available items covering the subjects music (21%, 5 large datasets holding a large quantity of online available items), poetry, art (both covered in a large number of datasets 11, 13, in a small quantity of online available items) and drama (covered in 3 datasets in a relatively large quantity).

As not many datasets cover the subject terms probably used by social scientists, the relevance of the datasets for these research communities is hard to determine.

In the next chapter, the datasets in the Europeana portal are matched to the humanities and social sciences research communities by way of a content analysis to identify the gaps and strengths in the content.

4. Content analysis II - Europeana

The Europeana Foundation[114] is the operator of the Europeana service. The Foundation provides access to the digitised datasets of museums, libraries and archives across Europe via the digital portal and search engine Europeana[115].

Full details about projects and institutions which have supplied data to the Europeana portal, is provided on another website, called Europeana Professional. This website provides background information to cultural heritage professionals and technologists. It is also the official source for technical information, metadata standards and case studies, and brings together information about current and finished projects undertaken by Europeana and its partners.[116]

This chapter focuses on the datasets held in the Europeana portal. To clearly distinguish between the many parts of Europeana, a reference to Europeana is made when the whole organisation and its services is meant, while the Europeana Foundation refers to the operating organisation.

4.1 Background information on the Europeana portal

Europeana aggregates metadata and thumbnails[117] for its portal. It does not store content itself but rather links to content held by data providers.[118] The aggregation process for the portal is intricate: 150 different aggregators supply metadata collected from around 2,300 institutions. This leads to a large variety of data, divided over thousands of datasets. Datasets are created during the ingestion process into the portal. A dataset is defined by Europeana as: 'an administrative unit representing a suitable amount of content or metadata for ingestion or download'.[119] In some cases, a dataset consists of a set of metadata intentionally-defined at an institutional or thematic level but in other cases, the metadata may cover several diverse topics. Datasets may also be comprised of metadata from more than one provider. This is a complicating factor when it comes to the analysis of the Europeana dataset as a whole.

A difficulty the Europeana portal shares with The European Library portal is the lack of subject indexing. The 33 million metadata records (as of the third quarter of 2014) have few or no subject references.

One of the reasons for this is that the subject field in itself is not mandatory for ingestion, but is part of a set of mandatory fields from which the contributing organisation may select which it contributes to Europeana. Europeana enriches the supplied metadata records which feature subject terms by linking them to online vocabularies, using Gemet (General Multilingual Environmental) Thesaurus and DBpedia.[120]. The use of Gemet has the advantage that it

---

[114] http://pro.europeana.eu/foundation (28 April 2014)
[115] http://www.europeana.eu (28 April 2014)
[116] http://pro.europeana.eu/home (28 April 2014)
[117] Thumbnail: a reduced and/or low resolution version of the Digital Object, normally limited by pixel dimensions, commonly used as the basis of the Preview. http://pro.europeana.eu/glossary (10 September 2014)
[118] Content in this sense means: 'a physical or digital object that is part of Europe's cultural and/or scientific heritage, typically held by a data provider'; http://pro.europeana.eu/glossary (9 September 2014)
[119] http://pro.europeana.eu/glossary (16 April 2014)
[120] http://www.eionet.europa.eu/gemet/en/about/ (13 November 2014)

enables users to find items using subject terms in more than one language, as it provides a link between the subject term written in the initial language and the term in another language. Very recently, after this content analysis was undertaken, some large datasets were enriched by using the Art and Architecture (AAT) thesaurus, leading to more refined metadata with subject representation.[121] However, providing the granularity of subject terms that researchers requires is still an ongoing issue for Europeana.

And again, the recommendation from the Expert Forum to encourage Europeana's data providers to 'provide complete metadata in future' applies.[122]

4.2 Methodology

For the content analysis of Europeana's metadata and thumbnails, several different approaches were discussed and tested.

Improving the metadata was recognized as a requirement, both to support the content analysis and to enrich the Europeana dataset. As the Europeana Foundation was keen to continue building up its relationship with its network of aggregators, the choice was made to involve them in providing the dataset descriptions, with the aim of eventually using the descriptions to enrich the Europeana dataset. In reaching out to the Network, it was necessary to ensure that all providers would describe their datasets in a consistent manner, using the same vocabulary, allowing for proper analysis of the content. It was therefore decided to create a survey with pre-prepared vocabularies so that the respondent's answers would be uniform and consistent.

The vocabularies for resource type and subject were developed by Europeana, with the vocabulary used in a 2010 content report produced by Europeana as a starting point.[123] The vocabularies were designed using existing vocabularies (e.g. from the Getty, DBpedia and the Library of Congress), all containing linked open data. The vocabularies were reviewed by colleagues external to WP1 (but still linked with the Europeana Cloud project) at DANS, University of Gothenburg, Digital Curation Unit in Athens and The European Library. As one of the goals of the content analysis was to match the datasets in the Europeana portal to the humanities and social sciences research communities, respondents were asked to indicate subject strengths in their datasets.

The members of the network were asked to provide this information via a survey.[124] As per normal Europeana practice, communication with Europeana's data providers was through the aggregators. In a pilot, 26 aggregators of Europeana were targeted and ten responses were received. This was seen as an acceptable response rate so the survey was further distributed to 150 aggregators, who in turn were asked to distribute the survey to some 2,300 data providers. Unfortunately, this approach did not have the expected result: in two months only 62 responses came in.
The decision was made to close the survey and look for a different approach. Various options were explored.

---

[121] Valentine Charles and Cécile Devarenne, Europeana Foundation, *Europeana enriches its data with the Art and Architecture Thesaurus*; http://www.pro.europeana.eu/web/guest/home (9 September 2014)
[122] See paragraph 2.2
[123] See Appendix A, Europeana Data Survey
[124] Ibid.

An attempt was made to match the Description of Works (DoWs) of previous Europeana projects with an ingestion component (and the final ingestion reports of these Europeana projects)[125] to the vocabulary developed for the survey. This did not generate sufficient useful or meaningful data for a content analysis. The projects are heterogeneous in scope, size and meaning, making comparison difficult. In addition, the ambitions expressed in the DOWs did not always correspond with project outcomes and the content and size of the datasets actually loaded were not always recorded in sufficient details to support a thorough analysis.

Further tests were made to automatically match the subject vocabulary from the survey to datasets in the Europeana portal. This approach also failed due to a lack of consistent and sufficient subject indexing.

4.2.1 The chosen approach

After careful consideration, the best method to gain insight to the contents of the datasets in the Europeana portal seemed to be to manually map the datasets in Europeana to the survey. Earlier in the project, it was estimated it would take one person working fulltime a couple of months (considering 30 minutes per dataset) to analyse all the datasets in the portal.[126] It was therefore decided by the project partners in Work Package 1 to map the 100 largest datasets in Europeana against the survey, thereby indirectly mapping the data against its potential user communities. These 100 datasets were expected to be a representative sample of all the datasets in the portal, as together they contain over 22 million items, or two-thirds of the 33 million items in the portal.

This approach would provide clues as to the resource types and subjects of the datasets, as well as the timespan of the content and its geographical spread. Datasets delivered by libraries were excluded, as they are covered by collection descriptions in The European Library portal analysed above in Chapter 3. Excluding the datasets for which a survey had been filled in (the 62 initial responses) proved to be impossible because a very limited number of responses included a dataset number – thereby making it impossible to determine whether any of the 62 were among the 100 largest datasets.

Trinity College Dublin explored the 100 datasets via the Europeana Portal and mapped them against the vocabulary of the survey (see Appendix A for a link to the survey and the responses).

The largest dataset reviewed contained 1,179,249 items and the smallest 63,150 items.[127] The data survey allowed for the mapping of the datasets on date (year) and the item type, both at a broad level (e.g. text, image, video, sound, 3D) and at a more detailed level (e.g. photograph, ceramics, manuscript, folio, documentary). Furthermore, the academic discipline covered by the datasets could be indexed, on a broad (literature, philosophy, law etc.) and a more detailed level (e.g. social history, drama, cubism, agriculture).The original survey was designed as a questionnaire to be send to the aggregators, who in turn should send it to the data providers. For the mapping of the 100 largest datasets to the vocabulary used in the survey, the survey was modified in order to capture extra information such as the collection number, data provider and number of files within the

---

[125] http://pro.europeana.eu/projects (10 September 2014)
[126] Estimate made by staff of Europeana and The European Library
[127] You can search the Europeana Portal (www.europeana.eu) per dataset if you enter europeana_collectionName:[number]* (so, for example europeana_collectionName:09404*)

dataset. Mandatory fields were removed, as the required information could not in all cases be filled in. In these cases, an explanation for the omission is provided.

4.2.2 Mapping the 100 largest datasets to the Dataset Information Form[128]

In this paragraph, the actual mapping of the datasets (as viewed through the Europeana portal) is described in detail. The first few questions of the data survey can be answered by reviewing the filters on the left hand side of the Europeana portal (e.g. languages, providers, object type, date and country of origin). These filters were used to get a wider understanding of each dataset. Where there seemed to be anomalies among the filters, the individual objects were investigated. For example, where a year looked out of place compared to the other years represented in the dataset, it was checked if either a mistake had been made or if the item was legitimately listed. Not all items in each dataset could be reviewed, as the datasets were much too large to do this. It was therefore decided to take a sample of each dataset, on average to review 10% of the items. For most datasets this was feasible, though for the largest one this would equate to well over 100K items. Therefore, a different approach was considered, using the 'page view option' in the Europeana portal.

The 'page view' option in Europeana allows for viewing of a great number of items at once on a single page in order to reduce the overall number of pages to be viewed was. At the most, 96 items could be reviewed in a single page. Taking the total number of pages, we then determined how many pages needed to be viewed in order to have looked at 1% of the collection (Total pages / 100). This was tried out on the first dataset, which had 123 pages to review. Closely viewing 2-3 items on each page took approximately one hour and a half to complete. The next two datasets also took over an hour to review. As most items seemed to be homogeneous and the overall number of pages to be reviewed still remained large, it was decided to look for a more efficient and less time-consuming approach. This is described using the Irish Manuscript dataset as an example.

For the Irish Manuscripts dataset it was soon clear that it only contained records from the 1911 Irish Census. A check was therefore carried out in 100 page intervals for the first 1000 pages. After this, a check was done in intervals of 1000 (total pages was 9447), until the last 447, which were once again looked at in increments of 100. This approach was continued for the first 30 datasets, after which the datasets had less than 5000 pages, so every 500 pages were checked. When the datasets had less than 2000 pages, every 200 pages were checked.[129]

Depending on the variety of items within the datasets, more or less pages were reviewed. If the variety was large, more pages were reviewed, if a dataset contained homogeneous object types, for example, only object type text, fewer pages were reviewed. If a dataset contained two or more file types, the reviewing started by first filtering on file type and reviewing the pages containing that specific type before moving to the next file type. This approach increased the overall number of pages to be reviewed, but if the file type within a dataset covered the same or similar subjects (for example, plant specimens or photographs of buildings in the post-war era), the level of reviewing did not need to be very detailed.

---

[128] Called Dataset Information Form, see Appendix A

[129] This collection has since been removed from the Europeana portal for reasons related to rights and data quality

For the review of most datasets, using the thumbnails of the items provided enough information. If the items looked unusual in a dataset, they were reviewed in closer detail. This allowed for a reasonable quick review of the datasets. Problems arose when a dataset did not have thumbnails attached to its items or if the items were described in a language unfamiliar to the reviewer. In these cases, more than 5 or 6 items per page were reviewed to get a better understanding of the datasets. In some cases it was impossible to review a dataset because links to the items from the Europeana portal to the original items on the page of the data provider were broken. This was noted during the mapping in the responses.

During the content matching, issues were encountered which complicated the analysis:

- Repetition of collections and/or items: collections in this case refer to: 'a set of metadata that has been intentionally-defined at an institutional or thematic level'.[130] In some cases items from one collection appear in another collection from the same organisation.[131]

- Broken links and access to original items: several items have broken links. An inability to access the original item or even to be able to visit the website of the cultural heritage institution that houses the item, poses a problem for researchers. Europeana acknowledges this issue and estimates about 3% of the links in the portal are unreachable on any given day. Europeana strongly appeals to its data providers to use persistent identifiers.[132]

- Lack of thumbnails: as Europeana argues itself, thumbnails are important to users. Items with a thumbnail are 8 times more likely to be clicked on than items without a thumbnail.[133] Thumbnails help users to determine what an item is, especially if the metadata is not very descriptive or is sparse. Europeana encourages providers to supply thumbnails but this is not mandatory. In 2013, 40% of the items in the portal lacked a thumbnail.[134]

- Multilingualism of the metadata: The Europeana portal allows for the automatic translation of a specified search term in six different languages.[135] This translation function goes some way to addressing language barriers but often the translation is too literal to make any sense in the target language, or it doesn't recognise all of the words (e.g. if something has been abbreviated or misspelled). If data providers use abbreviations in their metadata, they will probably do this in their natural language. The translator will attempt to translate these literally and will probably fail. On the other hand it cannot be expected from the data providers to revise their metadata cataloguing practices. This is where a thumbnail would help the user with more information before deciding if they want to click through to the

---

[130] http://pro.europeana.eu/glossary (11 September 2014)

[131] In at least one instance a collection is repeated within a dataset, see collection number C03915, which is divided into two collections nominally (one listed as 'Archaeology" and the other as 'MHR')

[132] Europeana, 'Technical Requirements' (for data providers), http://pro.europeana.eu/technical-requirements (11 September 2014

[133] Ibid.

[134] Annette Friberg, Harry Verwayen, *Europeana version 2: Collections and Metadata Analysis, Strategy and Plan 2013*, The Hague: Europeana Foundation, 29 March 2013, p. 5
http://pro.europeana.eu/documents/866067/1660973/D3.8%3A+Content+Strategy+Update_(15 September 2014)

[135] David Haskiya, '*Improving search across languages*' http://blog.europeana.eu/2014/08/improving-search-across-languages/ (14 August 2014)

content on the provider's site. The European Library portal also features an automatic translation function. These translation tools go some way to playing to the strength of both datasets, bringing together a multi-cultural, and multi-lingual treasure trove as they do (an aspect that Google Books is often commented upon unfavourably).[136]

4.3 Visual representation of the content analysis

In the previous analysis of The European Library portal (Chapter 3), the numbers referred to items available online (e.g. a link to a digital item or to bibliographical metadata). In this section, the numbers refer to metadata records and thumbnails (called 'items') or the datasets they belong to, since the Europeana portal does not hold content. As in the European Library analysis, the content analysis focuses on the relative distribution of the categories resource type, country, period and subject. Only a relative distribution over these categories can be produced, as datasets may contain or cover more than one resource type, country, period and subject. Therefore, the figures below may be represented in more than one category.

4.3.1 Relative distribution of resource types in the Europeana portal

	IMAGE	TEXT	SOUND	VIDEO	3D	Total
All	20,408,320	12,709,888	498,782	231,222	15,319	**33,863,531**
%	60%	37,5%	1.47%	0,68%	0,045%	

Table 6 Total content of Europeana by type (May 2014)[137]

This table from a recent content report issued by Europeana indicates the relative distribution of resource types linking to digital items. Images make up 60% of resources in Europeana, followed by text at 37%. This differs slightly with The European Library portal, where text and image are almost equally divided. Video, sound and 3D together make up a small part of the total content.[138] As archaeologists often engage with 3D objects, their information needs will probably not be met by searching in the portal.[139]

As both humanists and social scientists value text highly as an essential tool for research (see Chapter 2, p. 4), the relatively low percentage of text might diminish the usability of the datasets in the portal for their research purposes. On the other hand, in an earlier stage of the Europeana

---

[136] Deliverable 1.2 'State of the art report on digital research practices, tools and scholarly content use' (July 2013), p. 15 ; http://pro.europeana.eu/web/europeana-cloud/results/-/document_library_display/p6BV/view/2240207 (28 August 2014)

[137] From the Content Report made by Europeana staff

[138] With 3D is meant objects which have been technically rendered to resemble their complete 3D form

[139] Deliverable 1.1 'Research Communities Identification and Definition Report', p. 13

Cloud project it was concluded the portal has research potential for humanities scholars, certainly for the ones active in the subject domain of history.[140] This may be less true for social scientists, as earlier research in the project indicated having access to digital images is of less importance to them than having access to digital text content.[141] Both humanities scholars and social scientists emphasize the need for consistent and enhanced metadata, to comply with their needs and requirements.[142]

The Europeana portal currently holds small amounts of sound and video content, which can be seen as a gap that should be filled, as this content is perceived by both researchers in the humanities and social sciences as offering 'novel search possibilities' and 'highly valued among historians'. Europeana already started adding this content by way of the projects Europeana 1989 and Europeana 1914-18.[143] If the sound and video content grows significantly in the Europeana portal, it will add to the attractiveness of the Europeana portal for researchers, assuming the portal also offers opportunities to capture, present and cite the material.[144] It will then provide a way to attract researchers with a distinctive and different kind of content, from many languages and European countries. Content other digital resources may not be able to provide.

Figure 8 provides a further analysis of the resource types by depicting the distribution of the top 19 resource subtypes to be found in Europeana.

---

[140] Deliverable 1.1 'Research Communities Identification and Definition Report', p. 13

[141] Deliverable 1.2 'State of the art report on digital research practices, tools and scholarly content use', p. 17

[142] Deliverable 1.5 (2) 'Expert Forum Tools and Content for Humanities Research Report', p. 22

[143] Deliverable 1.2 'State of the art report on digital research practices, tools and scholarly content use', p. 18
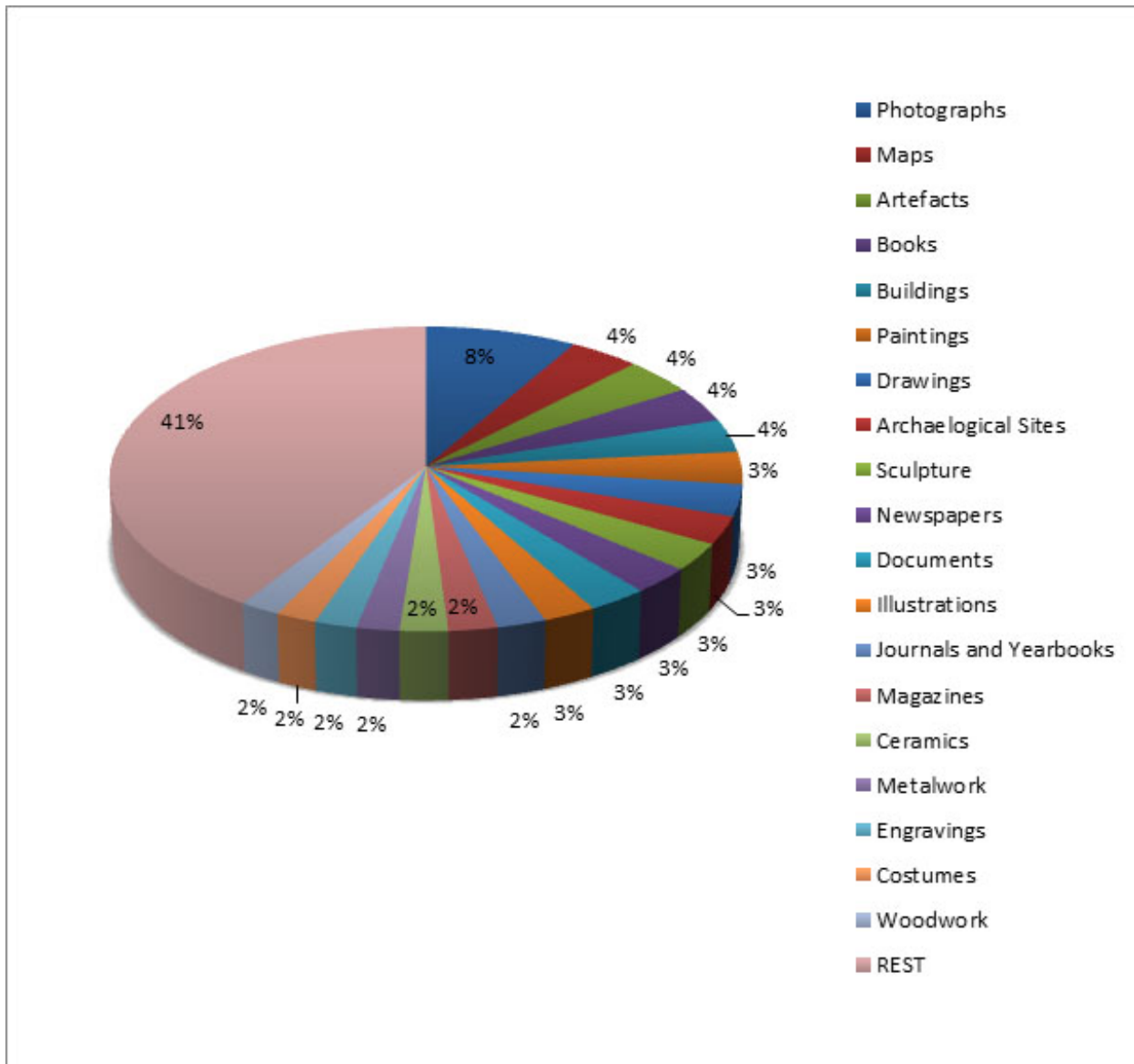
[144] Ibid., pp. 16-17

Fig. 8 **Resource Subtypes**: 19 most commonly recurring resource subtypes relative to the rest of the resource subtypes

This figure depicts all of the resource subtypes which cover at least 3 million items (1% of the total amount of items available in Europeana) – 19 in total. Breaking down the resource types of text, image, sound, video and 3D into subtypes shows the large diversity of material in the Europeana portal. Looking at the 59% subtypes divided over the categories photographs until woodwork, we can conclude that 14% is text (books, newspapers, documents, journals and yearbooks, magazines) and 45% is image or 3D. This complies with the earlier depicted distribution of resource types in table 6.

- Photographs account for the largest part of the total items, possibly offering relevant research material for one-third of humanists.[145]

---

[145] In Deliverable 1.2 'State of the art report on digital research practices', p. 12, it was stated 36% of the humanists would be interested in photographs

- Artefacts (4%) account for another possible relevant source of research material for almost 20% of humanists.[146]
- The remaining resource subtypes are almost equally divided over the portal, each containing 2-4% of the total amount of items.

The fact that most resource subtypes represent just 2-4% of total content illustrates the large diversity of material in the Europeana portal. This is an advantage in the sense that it provides humanities and social sciences scholars with a broad choice of material but also a disadvantage because it is difficult for researchers to know what to expect when searching the portal.
This runs contrary to the needs of researchers. Studies done by Europeana Cloud found that humanities scholars want to know what they can expect to find in Europeana. Additionally, the relatively small numbers of available items in each category makes the chance of finding resource types relevant to a particular research topic quite low. Providing more contextual information to the users about the content in the portal may reduce this problem, as will the implementation of a content strategy.

4.3.2 Geography: relative percentage of total dataset referring to specific countries

The next figures provide an overview of geographical distribution of items in the Europeana portal. Here we examine the countries which the collections and items describe and *not* the countries which contributed the data. As described above, every xth item was visually inspected. Based on this, an indication of the geographical area (or areas) referenced in the collection was recorded. The country of origin of the data is indexed in the Europeana portal, though Europeana acknowledges that 'accurate geospatial information is often lacking'.[147] Unlike the data's country of origin, the information to which the items refer to is not indexed, and has not been analysed before.

---

[146] Deliverable 1.2 'State of the art report on digital research practices', p. 13
[147] Annette Friberg, Harry Verwayen, *Europeana version 2: Collections and Metadata Analysis*, p. 31
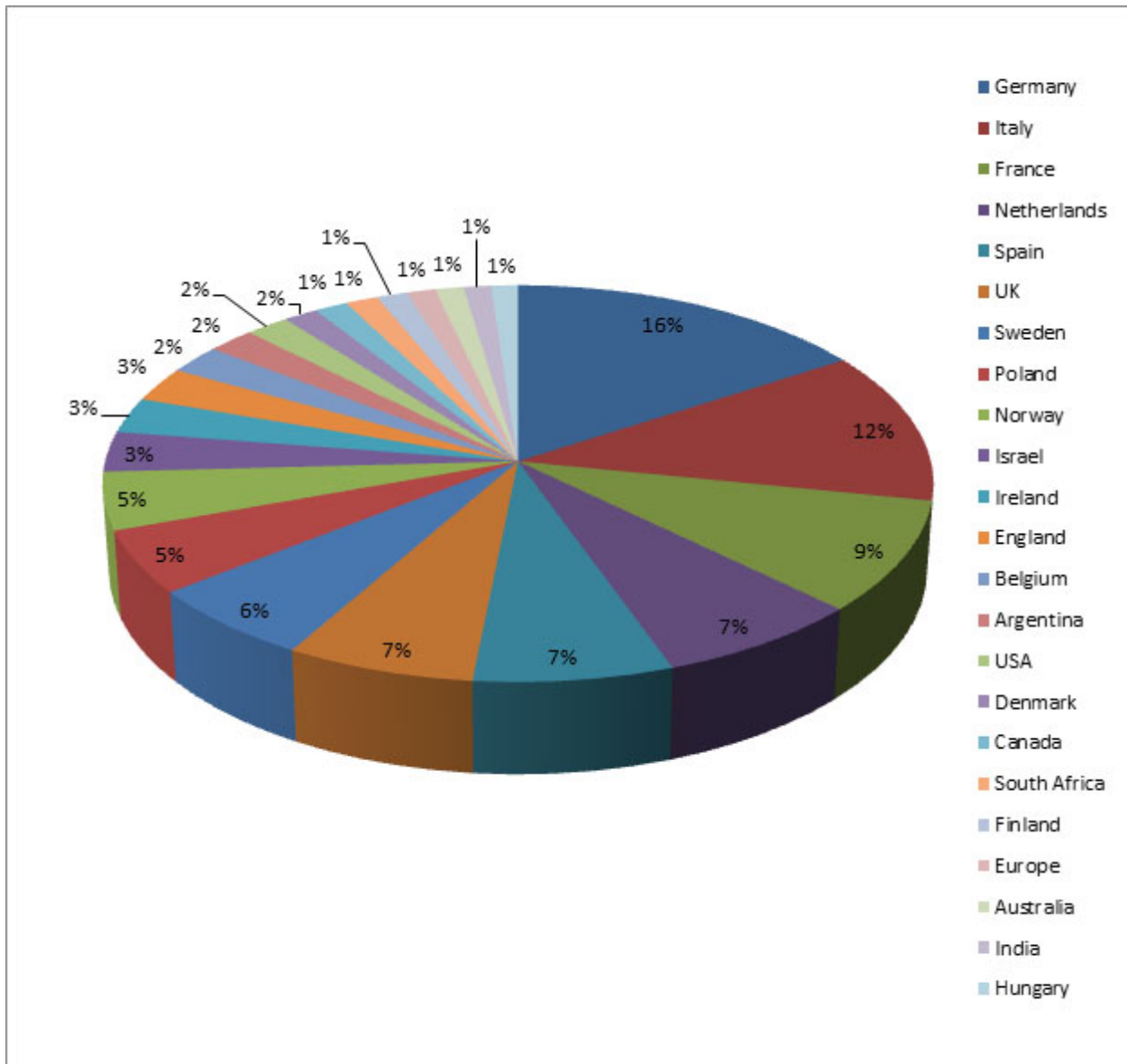
Fig. 9 **Geography**: Top 23 number of items per country (coverage)

The Europeana portal references over 23 countries. This figure provides an overview of the 23 countries which are covered by at least 400,000 digital items each. As Germany, Italy and France are among the largest European countries, it is not surprising that they form the top three. The fourth place for the Netherlands is, despite its much smaller size, is perhaps more surprising, as is the fifth place for Sweden.

To explore whether most countries provide items about their own country and fewer about other countries, the geographic coverage of the items per country will be compared to the Europeana Top 15 by country of provider (fig. 10).
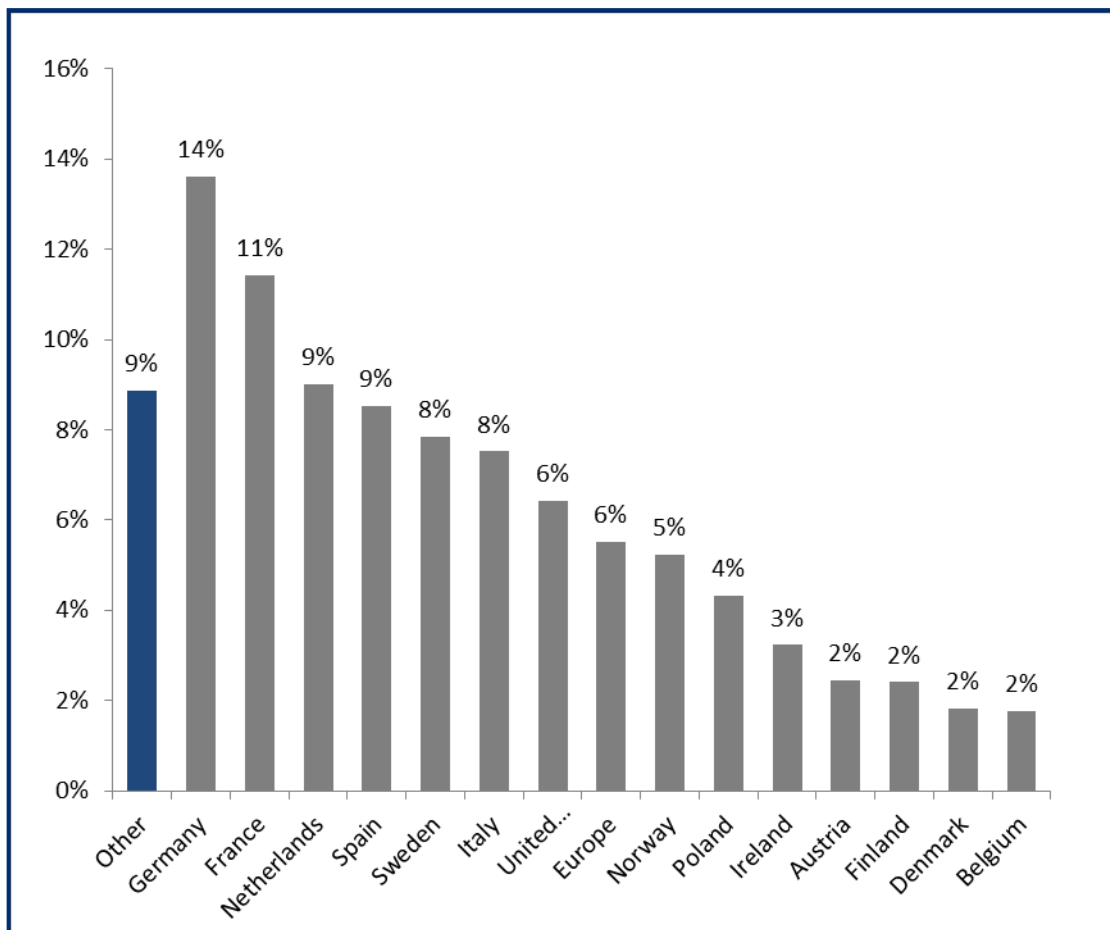
Fig. 10 Europeana Top 15 by country of provider[148]

From comparing figures 9 and 10, we can see the Netherlands and Sweden have, by provided a relatively large amount of items about their own countries.[149] The 10th place for Norway in figure 9 is striking, as Norway is not part of the European Union and as such has no obligations to the EU regarding the delivery of cultural heritage content. Israel and Argentina also feature in the top 15 in figure 9, but they do not appear in figure 10, which can be explained, as both countries are not part of the network of data providers to Europeana.

Striking in both figures is the absence of culturally-rich European countries such as Russia, Switzerland and Greece and almost all eastern European countries (except Poland and Hungary). On the other hand, as Europeana focuses on contributions from EU member states, it is not surprising Russia and Switzerland are no longer providing significantly to the portal. The European Commission advised in 2011 that countries like Bulgaria, Czech Republic, Greece, Latvia, Romania and Slovakia should deliver more content to the Europeana portal. Europeana stimulates this by supporting these countries with any aggregation issues and developments.[150]

---

[148] From: Europeana, Content report from 19 May 2014, p. 4

[149] Actually, by doing this they have already surpassed their recommended content contribution to the Europeana portal set by the European Commission for 2015: see Annette Friberg, Harry Verwayen, *Europeana version 2: Collections and Metadata Analysis,* pp.11-12

[150] Annette Friberg, Harry Verwayen, *Europeana version 2: Collections and Metadata Analysis,* pp. 11-14, 33

This may lead to a more balanced future division of items between West and Eastern Europe. A small step in this direction has been made by Estonia, Lithuania and Latvia, who have contributed much more content in 2013 than before. In general, material from eastern European countries is more difficult to access than in the rest of Europe and having this material available in the Europeana portal will therefore add greatly to Europeana's potential for the research communities.

4.3.3 Chronology: temporal distribution of online available items over the datasets

The next figures provide a chronology: an overview of the temporal distribution of items in the Europeana portal. The distribution of Europeana content over the centuries is depicted in two different graphs (fig. 11-12), comparing the content in March 2013 to the current situation in October 2014. The data is based on Europeana's Collection Metadata Analysis[151] and on an analysis done by CERL on recent data provided by Europeana staff.

Determining a chronology which presents an accurate temporal distribution of online available items in the Europeana portal is challenging because of issues with the dating of the items. Two-thirds of the items in the portal hold 'year indications', according to Europeana, but they are not always accurate.[152] (mainly caused by incorrect or inconsistent metadata provided by the contributing institutions).[153]

A document from 1856, for example, might be marked '1856' (the date it was originally created) or as '2005' (its digitisation date). Other issues with the metadata include incorrect punctuation or typing mistakes (e.g. '2310'; or '-1800', possibly indicating a date BC). Some items depict a date range (e.g. 1986-1987) or a date referring to a period of time (e.g. second part of the 15th century).[154] It is impossible to indicate which percentage of the records has an incorrect year indication.

We must keep these issues in mind when looking at the temporal distribution of items in Europeana.

---

[151] Ibid.

[152] Ibid, p. 20

[153] Ibid, p. 20

[154] This last issue was also raised by Annette Friberg and Harry Verwayen in: *Europeana version 2: Collections and Metadata Analysis,* p. 20
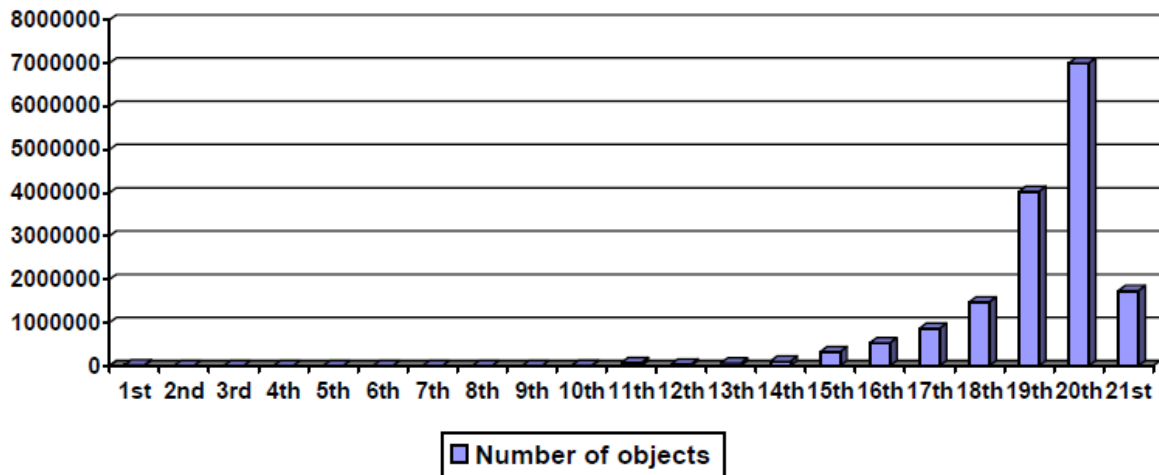
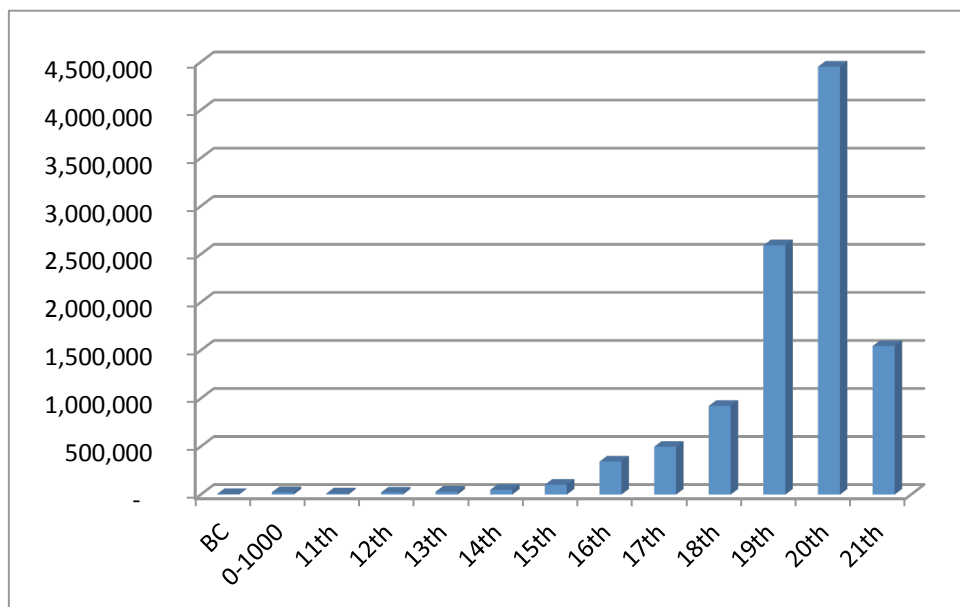Fig. 11 Distribution of Europeana content over the centuries (2012)[155]

Fig. 12 Distribution of Europeana content over the centuries (October 2014)

Both graphs show the distribution of Europeana content over the centuries. Figure 11 (2012) is based upon the date information gathered for 16.3 million objects.[156] The total amount of objects containing date information has lowered over the past two years so figure 12 (October 2014) is therefore based upon date information gathered for 10.6 million objects. The difference of about 6 million objects may be explained by the fact the Europeana Foundation announced in November 2012 it would bring all its metadata under the terms of the Creative Commons CC0 1.0 Universal Public Domain Dedication (CC0).[157] This means 'that anyone can use Europeana metadata for any

---

[155] Annette Friberg, Harry Verwayen, *Europeana version 2: Collections and Metadata Analysis,* p. 21
[156] Ibid., p. 20
[157] Breandán Ó Nualltáin*, Europeana opens up full dataset for re-use,* 9 November 2012;

http://pro.europeana.eu/web/guest/pro-blog/-/blogs/europeana-opens-up-full-dataset-for-re-use (4 November 2014)

purpose - creative, educational or commercial - with no restrictions'.[158] Not all cultural heritage institutions could or wanted to comply with this new rights statement declaration and some subsequently withdrew their metadata from the portal.

The figures seem to belie the notion of the 'Great Twentieth Century Hole'[159]. However, considering the output of text, image, video and sound in both analogue and digital form in the twentieth and twenty-first centuries, the spike should be much higher. One of the prime reasons for the absence of this material is copyright protection. Another reason is the expense of digitising modern media (such as film and video tape) and the fact that digitisation efforts have been geared more towards (the preservation of) paper records and art. [160]

Europeana aims to fill this gap in modern materials by targeting content from these decades with projects like Europeana Fashion, Europeana Newspaper and Europhoto, Europeana Photography and Digitising Contemporary Art, and by attracting user-generated content through projects such as Europeana 1914-1918. The aggregation of this kind of unique content will increase the research potential of Europeana for both humanist and social scientist researchers. As earlier research in the project has found, social scientists are mostly engaged in qualitative research using contemporary data. As much of this data is locked away in trusted digital repositories for ethical reasons, the afore-mentioned projects and possibly others to come will add a welcome amount of potential research data for social scientists.[161]

4.3.4 Subject coverage: distribution of subjects among digitally available items in the datasets

One of the goals of the Europeana survey was to match the content of the portal to the humanities and social sciences research communities. One of the ways to investigate this was to determine the area of study the datasets would be most useful for, the so-called subject strengths of the Top 100 of the datasets in the European portal. The top five subject strengths of each dataset were determined using the controlled vocabulary of the survey. For the full list of subject terms see Appendix A.

---

[158] Ibid.

[159] Great Twentieth Century Hole: resulting in a vicious circle: the absence of a vast amount of material dating from the 20th century, so a low availability of research material, leads to few 'large-scale data driven projects' in Digital Humanities studying twentieth-century phenomena. A low demand for 20th century material from digital humanities researchers in turn discourages digitisation efforts by cultural heritage institutions...
https://availableonline.wordpress.com/author/xcia0069/ (12 December 2014)

[160] Ibid, pp. 35, 37, 38

[161] Deliverable 1.1 'Research Communities Identification and Definition Report', p. 12
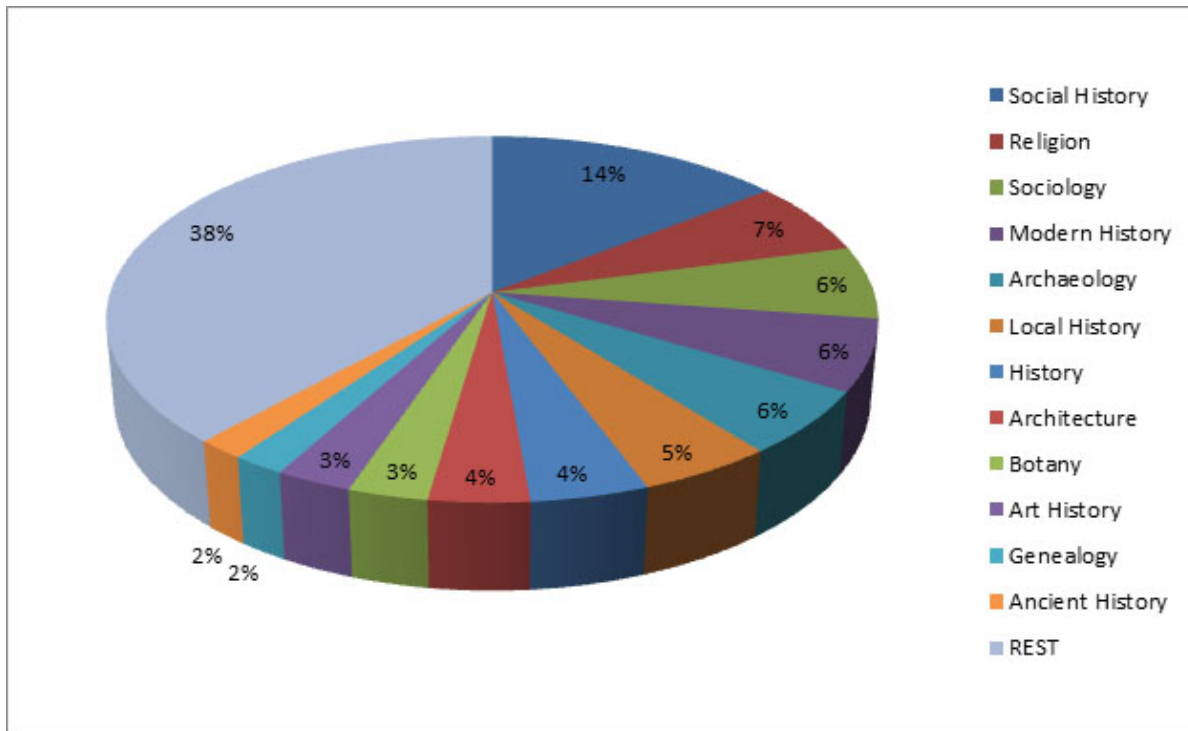
Figure 13 The top 12 of most suitable subjects for research compared to the rest of the subjects

The figure above depicts the 12 most suitable subject terms for research, based upon estimates done by Trinity College Dublin.

- The indicated research potential for social scientists seems reasonable, with the subjects Social History (14%) and Sociology (6%) accounting for 20% of Europeana content.
- The research value for Sociologists is harder to determine. The subject categories Religion (representing 7% of content) and Genealogy (2%) might be of interest to sociologists.
- The research potential for humanists is easier to determine, as the subject domains are clearer. The subjects Modern History, Archaeology, Local History, History, Art History and Ancient History together make up 28% and humanists (like social scientists) might find relevant research material in the datasets covering the subject Religion.

It is difficult to determine the boundaries between humanities and social sciences, as was explained in Europeana Cloud Deliverable 1.1.[162] Even in this report, the different definitions will become apparent.

Until now, our analysis has focused on subject domains as identified in Deliverable 1.1. In the example below (fig. 14) we use the division made for the survey for the content analysis of the datasets in the Europeana portal, based upon established vocabularies. To clarify which terms

---

[162] 'The idea of what constitutes a discipline within the broader field of the Arts and Humanities has always been rather hazy, and largely depends on different institutional structures in various schools, universities and countries'. Deliverable 1.1, p. 5 http://pro.europeana.eu/documents/1414567/1b9abb08-9e3c-485a-aca6-bccaff9a29e8 (13 November 2014)

from the survey are part of either humanities or social sciences, a full break-down is included as Table 7.

Figure 14 shows the amount of items divided over the humanities, the social sciences and the overlap. From this illustration we can see that a considerable part of the datasets in the Europeana portal hold research potential for humanities scholars. The research potential for social sciences is harder to determine and seems somewhat lower. The boundaries between both disciplines are diffuse, as is indicated by the large overlap (the red part) in Figure 14.
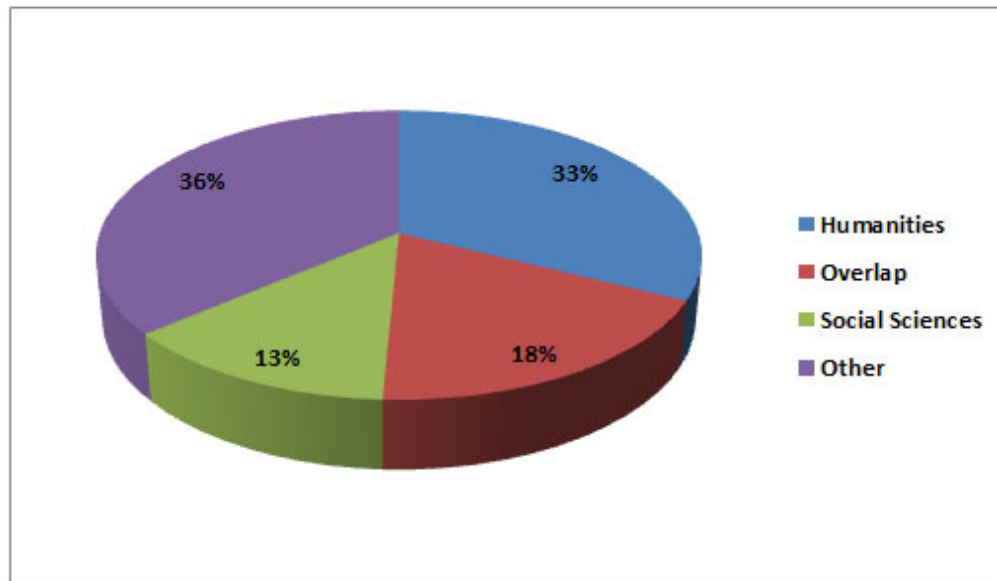

Fig. 14 Subject strengths divided over the humanities and social sciences

Table 7 Subject terms used to determine the subject strengths

Humanities	Social Sciences	Overlap
Archaeology	Anthropology	Social History
Modern History	Gender Studies	Local History
Archaeology	Political Science	Totalitarianism
History	Sociology	World Politics
Art History	Policy	
Ancient History	Genealogy	
Medieval History	Gender Studies	
Military History	Political Science	
Philosophy	Propaganda	
European History	Social Sciences	
Linguistics		
Spanish Civil War		
World War 1939-1945		
Musicology		
World War 1914-1918		
Political History		
Digital Humanities		

Source: Dataset Information Form (see Appendix A).

This content analysis of the datasets in the Europeana portal gives insight into the strengths and weaknesses in the content of the portal, while matching it against the target audience for Europeana Research. In Chapter 2 these communities were defined and their scholarly content use was explored. In the next chapter, Europeana Research will be described in more detail. The last chapter of the report focuses on the recommendations which can be extracted from the content analyses of the portals of The European Library and Europeana.

5. Recommendations for Europeana Research

This chapter provides a short introduction of Europeana Research and then goes on to present conclusions from this report in the form of further recommendations for the content strategy and the future development of Europeana Research.

5.1 Europeana Research

One of the objectives of the Europeana Cloud project is to develop a new service labelled Europeana Research. Europeana Research will be created to discover and re-use content from the project partners The Europeana Foundation, The European Library and Poznan Supercomputing and Networking Center.[163] The service will not contain the generic, pre-defined search and browse functions currently available in the portals of The European Library and Europeana, the content of the Europeana Cloud will be access via an API, instead. Europeana Research will be used to actively engage with the designated communities, which will be a highly flexible audience serviced by a highly flexible service.

Based on this report, and in the context of the wider Europeana discussion on its Content Publishing Strategy, the content strategy for Europeana Research will be developed next in the project, resulting in Deliverable 1.6 'Content Strategy Report'. It will be based upon user requirements (with actual development of course delimited by the confines of technical possibilities) and identification of potential users and user scenarios. The recommendations outlined in the next section are to provide a useful framework for the development of the content strategy with regards to target audience, content and tools, rights labelling, user requirements and user expectations and user satisfaction.

5.2 Recommendations

5.2.1 Target audience
A first step in the development of a content strategy for a certain service is a clear definition of its target audience, allowing for the provision of tailored services which satisfy their needs and match their expectations. As is stated in the publication *Business model innovation cultural heritage* by DEN:
'It is not the quality of the information provided that determines the success of the service, but rather whether or not the customer feels well served'.[164] In this respect, it was recommended in Expert Forum 2 that Europeana's landing page should be reconfigured to become an attractive entrance point for all targeted user groups (including API-developers) and offer examples for research, queries and case studies.[165] This holds true for Europeana Research. Even though it is not conceived as a portal, its points of engagement with the intended end-user will have to be transparent about what is on offer, and how it may be used effectively.

---

[163] Competitiveness and innovation framework programme 2007-2013: Annex 1: Description of Work, p. 7

[164] DEN Foundation, Knowledgeland, Ministry of Education, Culture and Science, *Business model innovation cultural heritage*, (Amsterdam, The Hague, 2010), p. 33
http://www.den.nl/art/uploads/files/Publicaties/BusModIn_eng_final.pdf (12 April 2012)

[165] Deliverable 1.5 (2) 'Expert Forum Tools and Content for Humanities Research Report', 2013, p. 23

Europeana's landing page should be reconfigured to become an attractive entrance point for all targeted user groups (including API-developers) and offer examples for research, queries and case studies.[166]

As, with Europeana Research, the Europeana Foundation has chosen to service a new group of end users, the definition of the exact user group and their expectations and needs is of eminent importance. In Deliverable 1.1 a start was made in exploring who might be the end user of Europeana Research. There, the target audience for Europeana Research is defined as 'a group of researchers working in a common subject domain that is either part of the humanities or social sciences, who use similar computational methods to create, analyse and disseminate a certain resource type of a digital resource'.[167] These researchers in turn can be subdivided over specific disciplines, which are the prime candidates for using this service.[168]

Both sets of research communities are demanding, as they are accustomed to using digital resources such as Google as the starting point to locate relevant research materials. They expect other resources to be as easy and convenient to use and access as Google.[169] And both communities have their own peculiarities, making it even harder to service both of them well simultaneously.

Humanities scholars differ in information needs and search behaviour compared to researchers in other disciplines.[170] 'They do not have homogeneous information seeking behaviour or needs'.[171] This complicates the determination of the service level to be provided by Europeana Research. And for social science researchers, if the choice is made to offer the best research material possible, Europeana Research should not be geared towards the aggregation of qualitative data only, as social sciences researchers prefer to use both quantitative and qualitative data for their research.[172]

The identification and definition of the end users of Europeana Research clearly needs to be further refined. The exact user profile for Europeana Research is currently under development, and a start was made by defining proto-personas (or 'ad-hoc persona').[173] Creating proto-personas helps to capture our assumptions about an end-user prior to discovering and validating her true nature with research.[174]

---

[166] Ibid., p. 23

[167] Deliverable 1.1 'Research Communities Identification and Definition Report', p. 4.

[168] See paragraph 2.1

[169] Deliverable 1.2, 'State of the art report on digital research practices, tools, and scholarly content use', p. 17 and elaborated on in: '10 Things That Humanities Researchers Want', *Europeana Libraries*, (2011) http://www.europeana-libraries.eu/web/guest/home/-/asset_publisher/5HLm/blog/10-things-that-humanities-researchers-want;jsessionid=95E7BF4FE37829C1FB2F034AE4461244 (11 November 2014)

[170] C. Ross, M. Terras, V. Motyckova, 'Measuring impact and use: scholarly information-seeking behaviour', in L.M. Hughes (ed.), *Evaluating and measuring the value, use and impact of digital collections*, (London, UK: Facet Publishing, 2012), p. 87.

[171] Ross, Terras, Motyckova, 'Measuring impact and use', p. 87

[172] Deliverable 1.1 'Research Communities Identification and Definition Report', p. 14

[173] Alastair Dunning, 'Drafting priorities for Europeana Research', 30 July 2014) http://pro.europeana.eu/web/europeana-cloud/blog/-/blogs/trackback/drafting-priorities-for-europeana-research (20 November 2014)

[174] Chelsey Delaney, LeanUX NYC 2014 '*Enhancing proto-personas with characterization*' (workshop) http://www.slideshare.net/chelsed/leanux-nyc-2014 (16 December 2014)

-Future users for Europeana Research might very well be defined as a digital humanist and tech-savvy social scientist like the ones presented in Chapter 2, 2.3 'User profile of the digital humanities researcher and social scientist researcher'.[175] A helpful instrument for the further definition of the target audience may be ISO 9241 Ergonomics of Human Computer Interaction; addressing the familiar aspects of usability, effectiveness, efficiency and satisfaction in the context of use.[176]

Until now the Europeana Foundation, as instructed by the European Union, was focused on the aggregation of as much digital heritage of European origin as possible. This led to an aggregation policy oriented on quantity and diversity (quota suggested by the EU) of the content entered in the Europeana portal. Engaging with a different user audience places new demands on the aggregation of content and asks for a targeted and public content strategy, as is outlined in the next section.

5.2.2 Content strategy

Europeana Research should have a content strategy clearly stating the goals of the service and the plans for its future development. A content strategy literally implies: 'planning for the creation, delivery, and governance of useful, usable content.'[177] Such a content strategy helps the user to understand what will or will not be included in the service in relation to user needs and expectations.
The choices made in a content strategy are closely related to the definition of the target audience.

The content strategy should also state the selection policy and criteria, i.e. which subject matter will be aggregated and made available in what date format and data type. A recommendation to this effect was formulated at the Expert Forums: it should prioritise on a limited number of areas in order to have the highest possible quality, and metadata comprehensiveness, of the existing content.[178]

Elements that therefore need to be addressed in the content strategy are:
- background information on selection policy, future development and definitions of what is meant by cultural heritage and what is considered 'European'; (that is, if Europeana Research will be focused on European content);
- contextualisation: an exact definition of what (type of) content and metadata the service will contain.[179] Specifically for this audience providing provenance information is important as well as information about the availability of annotations;
- provide information about the availability of deep links; if links are provided to items hosted at a contributor's site, the user should directly be referred to them instead of being redirected to the homepage (or worse, a search page) of the contributor;
- selection criteria used to select the most rewarding data types and data formats;

---

[175] Section 2.3
[176] http://www.iso.org/iso/catalogue_detail.htm?csnumber=63462; reported in: L.M. Hughes (ed.), *Evaluating and measuring the value, use and impact of digital collections*, (London, UK: Facet Publishing, 2012), p. 82
[177] http://bussolati.com/contentstrategy-definition/ (16 December 2014)
[178] Europeana Cloud Deliverable 1.5 (3) 'Expert Forum Tools and Content for Social Science Research', p. 17
[179] Deliverable 1.5 (2) 'Expert Forum Tools and Content for Humanities Research Report', 2013, p. 23

- how to strike a balance between making available bibliographic records and providing (direct) access to digital objects.

If all partners can ingest content for use in Europeana Research and the content offering is therefore flexible and variable, it will be difficult to know what it exactly is available at any given time for a specific user. This has repercussions, because for researchers their research has to be repeatable so that they themselves or others may replicate the research.

5.2.2.1 Contextualisation and selection policy

In order to help manage user expectations, it is paramount to be specific about the type of content and metadata that is on offer. The selection policy should state which resource types are available and why and how they can be used by the researchers. Which resource types are considered the most useful to researchers or whether the service is more effective if some forms of data are given prominence over others is elaborated on in paragraph 5.2.4. The selection policy should also state whether Europeana Research will hold user-generated content; i.e. will it be a two-way street offering access to research data injected by researchers themselves? All content (individual items and entire research collections) should be equipped with good quality metadata that is consistent and enhanced to comply with the needs and requirements of the target audience.[180] It may also be necessary to add new elements to the metadata in addition to those currently provided, such as the quality of data (resolution of images, level of correctness of OCR, number of citations, etc.) and the provenance of the data (who contributed what, who enhanced what, etc.). In Europeana's Metadata analysis document (2013), a similar notion is included: 'managing both content and metadata effectively will lead to improvements in the user experience'.[181]

A related next step is to determine if the service will offer access to primary sources or also to secondary material.[182] And whether besides the already provided qualitative data also quantitative data will be provided, as social scientists mostly engage with this type of data.[183]

One of the conditions for researchers to use Europeana Research is knowing in what types the resources are made available to them and in which shape. As the Europeana portal only holds metadata and links to the digital objects on the site of the data provider this is more or less clear. The datasets in The European Library portal are more diverse, ranging from digital objects, direct and indirect links to digital objects and metadata, to access to bibliographic records only. In that case, providing clear information about the format of the data is important. If the data is in the form of digital objects, providing thumbnails is very helpful, as these help users to determine what an object is (especially if the metadata is not very descriptive or lacking altogether).

If the service will offer collection descriptions such as they appear in The European Library portal, it should provide precise descriptions of the collections or part of collections that are online/digitally

---

[180] Chapter 2, section 2.2; D1.5 (2) 'Expert Forum Tools & Content for Humanities Research Report', p. 22

[181] Annette Friberg, Harry Verwayen, *Europeana version 2: Collections and Metadata Analysis*, p. 8

[182] In: Deliverable 1.1 'Research Communities Identification and Definition Report', definitions for primary and secondary material are provided. Primary Data is defined as: computerised data sets, rare books and manuscripts, maps and charts, photographs, artefacts, moving image or sound recordings; 'secondary' material as: journals, newspapers, books and electronic publications accordingly. (…)[182]

[183] Deliverable 1.1 'Research Communities Identification and Definition Report', p. 14

available via the service, to inform the user about the context of the collection. A user will become wrong-footed if a description of an entire collection is provided when only a small part of this collection is available in the portal. And if textual collection descriptions from The European Library are to be included in Europeana Research, they should be indexed on keyword, in order to be searchable by researchers. The European Library portal allows for a 'discovery by discipline', which might be a helpful tool that could be adapted for use in Europeana Research.

If available, full-text documents or direct links to digital items should be provided to comply in the best possible way with the researchers who wish to have fast, convenient and immediate access to this type of information. If full-text documents or direct links are not available, or not in large quantities, user expectations should be managed by explaining why and provide an indirect link to the item or metadata (bibliographic record). If direct links are available and provided, the data providers should be persuaded to use persistent identifiers. The use of persistent identifiers makes it easy for the target audience to cite the digital resources they have used and for other researchers to replicate the research.[184]

5.2.4 Content divided by a. Resource Type, b. Geography, c. Chronology and d. Subject

a. Resource type

In order for researchers to be able to assess whether they need or want to use this type of data for their research, detailed information on the types of data that are available should be provided. Currently, neither The European Library portal nor Europeana contain much detailed information about the type of data that is on offer for the researcher to rely upon.

Text

As text is valued as the 'most significant source' for research by older and younger (digital) humanists and social scientists,[185] adding textual resources is important. Earlier research has shown that both digital humanities and social research communities prefer full-text.[186]

The datasets in The European Library portal hold more textual resources than the datasets in the Europeana portal and are possibly of greater interest to the target audience than the datasets in the Europeana portal.

Images

A large corpus of image content is available, especially in the Europeana portal (in 60% of the total datasets) but also in The European Library portal (in 40% of the total datasets).[187] This is likely to appeal to the target audience, and in combination with the available textual data may provide good research possibilities.

---

[184] '10 Things That Humanities Researchers Want', *Europeana Libraries*, (2011)
http://www.europeana-libraries.eu/web/guest/home/-/asset_publisher/5HLm/blog/10-things-that-humanities-researchers-want;jsessionid=95E7BF4FE37829C1FB2F034AE4461244 (11 November 2014)

[185] Chapter 2, section 2.1

[186] Chapter 3, section 3.3.1

[187] Chapter 4, section 4.3.1 ; Chapter 3, section 3.3.1

It has proven difficult to assess the research value of the resource type image as the information behaviour of the target audience using images differs for each sub-discipline.[188]

Sound and video
Adding more sound and video and moving image data to Europeana Research will add to the attractiveness of the Europeana portal for researchers. The target audience values these resource types as innovative and offering 'novel search possibilities'; they are especially 'highly valued among historians'. Offering this kind of content provides a way to attract researchers with a distinctive and a different kind of material compared to other digital resources. This is content that may not be available elsewhere, at least not in such quantities and in so many languages.[189]

3D
Neither The European Library portal nor the Europeana portal hold much content of this type.[190] It would be important to identify potential providers of this type of material. Particularly as archaeologists are keen to engage with 3D objects.

b. Geographical distribution of the content

Bringing together material about diverse geographic regions of Europe (current and past) will be considered an asset of European Research by the target audience. Actively aggregating material from eastern European countries, which is generally more difficult to access, will enhance Europeana Research's potential for the designated audience. This recommendation is based upon the assumption that the service will boost its handling of multilingual resources and metadata.

The datasets in both portals offer a large diversity of material, The European Library portal contains references over 32 geographic regions, and the datasets in Europeana cover over 23 countries. The Europeana Foundation actively stimulates the aggregation of material from eastern European countries, thus offering fresh and interesting research material. This may lead to a more balanced future division of items between Western and Eastern Europe.[191] And it certainly fits in with the current academic focus on the study of transnational phenomena.

c. Chronology of the content

Aggregating content from the 19th and 20th century, for example publications commissioned by public companies and governmental bodies, will stimulate research in this era. A targeted approach to ingesting digital repositories of publishers, actively sourcing research datasets and digital archives of companies and governmental bodies would ensure Europeana Research offers unique research material, not readily available in other digital resources, thus making it more attractive to its potential users.

Sourcing 20th century material is challenging because of what is sometimes called the '20th century black hole'; primarily due to copyright protection.[192] Part of this 'black hole' can be filled in by aggregating user-generated content such as Europeana did with the *Europeana 1914-1918* project.

---

[188] Deliverable 1.2 'State of the art report on digital research practices', p. 17
[189] Chapter 4, section 4.2.2
[190] Chapter 4, section 4.3.1 ; Chapter 3, section 3.3.1
[191] Chapter 3, section 3.3.2 ; Chapter 4, section 4.3.2
[192] Chapter 3, section 3.3.3

d. Content by subject[193]

As we have seen, Europeana Research aims to service groups of researchers, and defines them as a flexible audience requiring a flexible service. Each of these groups will require a tailored, homogeneous and fit-for-purpose corpus of content. This has implications for the sourcing of content.

Instead of providing many datasets covering diverse subjects with sometimes few datasets, which is how the portals are currently build up, large and tailored research corpora focused on specific subjects will need to be brought together. If the future target audiences of Europeana Research are focused on defined sub-disciplines, which is very likely, the current gaps in the content in The European Library and Europeana portals have to be filled in for each new research group that presents itself and for which not sufficient useful data is available.[194]

For the humanities the gaps in the content are, at the moment, in the following subject domains:

Archaeology: content in 3D format. The Europeana portal does hold enough research material on the subject Archaeology, but not in this format; The European Library portal does not provide this kind of content.

Linguistics: very little material is available in either portal covering this subject domain; it is not part of the 'subject strengths' as defined by TCD for the 100 largest datasets in Europeana.[195]

Law: another gap; makes up 1% of the total datasets in The European Library portal providing a medium-sized collection of datasets with large variety; not much content available in the Europeana portal; it is not part of the 'subject strengths' as defined by TCD.[196] As this subject domain is defined as one of the 'prime examples' of digital humanities[197], it is important for Europeana Research to acquire content covering this subject. These humanists do not require content of the types moving image and 3D.[198]

Readily available content for humanities:

History: both portals seem to offer interesting research material for historians, depending on their chosen subject. The subject categories 1914-1918; Modern History, Art History, Medieval History, Local and Military history will offer a large and diverse dataset. The Europeana portal also offers research material on Ancient History; this is one of the 12 subject strengths in the portal. As, according to the D1.1 Research Communities and

---

[193] For the full analysis and interpretation see sections on subject coverage: 3.3.4 and 4.4.4

[194] The analysis and interpretation of the top 12 most suitable Subject terms for research in the Europeana portal is based upon estimates made by Trinity College Dublin, see section 4.4.4

[195] Section 4.4.4 Subject coverage: distribution of subjects among digitally available items in the datasets

[196] Section 4.4.4 Subject coverage: distribution of subjects among digitally available items in the datasets

[197] Deliverable 1.1 'Research Communities Identification and Definition Report', p. 13

[198] Ibid., p.13

Identification Report, the subject domain 'history' has developed as a 'prime example of the digital humanities'[199], servicing these research communities can be rewarding for both parties.

Music(ology): musicologists interested in music (as opposed to material about music as a resource type) will find a reasonable research corpus in The European Library portal in their field of interest. Depending upon the point of view of the musicologist, he or she could experience the relative absence of the resource type music as a gap in the collection of datasets in the portal. 'Music' was defined as one of the subject domains in the Europeana portal of interest to the research communities.[200]

Philosophy: for humanists interested in the history of philosophy, some diverse research material (a medium-sized dataset of about 80,000 online available items) is available in The European Library portal. 'Philosophy' was defined as one of the subject domains in the Europeana portal of interest to the research communities.[201]

For social scientists the relevance of the described datasets is not always easy to determine, it has proven difficult to figure out what search terms social anthropologists[202] and sociologists would use. As sociologists and social anthropologists deal 'with all types of human endeavours and cultural expressions' they might find interesting material among the datasets in both portals.[203]

For the social sciences the gaps are in the following subject domains:

Social/Human/Economic/Political/Cultural Geography: a small quantity of research data is available in both portals, needs to be built up further.

Gender Studies: a small quantity of research data is available in both portals, needs to be built up further.

Economic History: a small quantity of research data is available in both portals, needs to be built up further.

---

[199] Ibid., p. 13

[200] Ibid., p. 12

[201] Ibid., p. 12

[202] A search for the Subject term Social Anthropology offers no separate datasets about this subject but searching the portal for the same term generates around 6,500 items. This is little compared to the material available on other subjects.

[203] Stefan Ekman of the University of Gothenburg (UGOT) has consulted several colleagues active in the subject domains social anthropology and sociology to determine what search or subject terms they would use to search for material in The European Library portal. The outcome was 'there is no limit to the search terms they would use, as they interested in all types of human endeavours and cultural expressions'.

Readily available content for social scientists:

Political Science**:** together with World Politics, Ideology and Propaganda a considerable amount of research data is available in The European Library portal; few data available in the Europeana portal: a gap in this portal.

Social history: a large quantity of research material with a large variety is available in both portals; this subject domain forms one of the subject strengths in the Europeana portal.

Sociology: a large quantity of research material with a large variety is available in the Europeana portal; it hard to determine if relevant research material is available in The European Library portal.[204]

(Popular) culture: sociologists interested in (popular) culture may also find relevant content, as The European Library portal contains a considerable research corpus covering the subjects Music, Poetry, Art and Drama.

Religion: sociologists interested in religion may also find relevant content, as both portals hold a considerable research corpus covering this subject.

Genealogy: is one of the 12 subject strengths of the Europeana portal, albeit in the 11th position. May also provide a considerable research corpus. In The European Library portal, this is a gap.

5.2.5 Tools

Some general remarks need to be made first regarding tools. As is stated in D1.2, 'tools very often do not feature prominently in research papers due to the fact that they are a means to an end to the researcher.'[205] And: 'the latest tool is not synonymous with the state of the art', in other words: 'the newest tools are not necessarily the best tools'.[206] Another point to consider is that researchers often use multiple tools on one dataset/content/collection, this needs to be enabled and made easy to engage this demanding target audience.[207]

The tools need to be tailor-made for the researchers; generic tools will proof to be not very useful, as was concluded in earlier deliverables and earlier research. Research done by LAIRAH 'Build it and they will come' found 'tools are evaluated and selected by researchers to achieve a specific goal therefore using a generic approach risks creating tools that are not useful to anyone'.[208] Van

---

[204] See section 3.3.4 Subject coverage

[205] Deliverable 1.2 'State of the art report on digital research practices, tools and scholarly content use', pp. 50-51

[206] Ibid.

[207] Ibid., p. 52

[208] Warwick, C., Terras, M., Huntington, P., Pappa, N. 'If You Build it Will They Come? The LAIRAH Study: Quantifying the Use of Online Resources in the Arts and Humanities through Statistical Analysis of User Log Data' *Literary and Linguistics Computing* 1 (2008), pp. 85-102

Zundert reached a similar conclusion in his research, stating that 'highly heterogeneous humanities digital data need 'highly distributed web services'.[209]

The case of Europeana Research the users will very likely build the tools themselves, based on data made available there. Europeana Research data will be accessible via an API, which raises two issues. Firstly, using an API effectively will require specific skills from the user/researcher, possibly restricting the user group. To achieve the maximal benefit of this approach, potential users of the service will be required to appreciate and become comfortable with working in an environment driven not by generic search and browse functions (as in the Europeana portal) but driven by the Europeana API as an enabler of advanced segmentation and download of Cloud data. This shift carries with it a substantial risk: although the target audience may be regular consumers of data provided to them over commercial or open source API's developed by others, the common perception of API development would be as a programmer's, rather than a historian's or linguist's, field of expertise.[210]

Secondly, as has become clear from this analysis report, it is difficult to fully appreciate the extreme complexity and variety of the content of both portals. For any user to successfully engage and develop tools with Europeana Research data through the API, they will most likely require access to staff at the Europeana office, who may help them to interpret the Europeana Research offerings. Providing an online tutorial or manual on the API might be a good starting point.[211]

Thirdly, considerations need to be made if new tools will be developed or if existing tools will be used. An advantage of the use of existing tools is that researchers are already familiar with them.[212]

From WP1 work with the Expert groups it has become clear that:
- Offering researchers the possibility to view relevant content based on previously viewed objects would enhance user satisfaction and promote use of the service
- Having data in different export formats would enable and enhance quantitative analysis by researchers[213]
- Enabling the export of records to popular reference management systems such as Mendeley and Zotero, helps the target audience to develop and reshape their e-research.[214]
- Providing consistent and sufficient metadata will improve trust in the database. Installing a user-ranking system or star-ranking system enabling researchers to evaluate the completeness and quality of the metadata would be useful.[215] As well as the development of metadata enrichment tools.[216]

---

[209] J. Zundert, "If You Build It, Will We Come? Large Scale Digital Infrastructures as a Dead End for Digital Humanities", *Historical Social Research*, vol. 37, no. 3, (2012), p. 165

[210] Citing Jennifer Edmond, Trinity College Dublin during a meeting: Alastair Dunning, 'Drafting priorities for Europeana Research', 30 July 2014) http://pro.europeana.eu/web/europeana-cloud/blog/-/blogs/trackback/drafting-priorities-for-europeana-research (20 November 2014)

[211] Suggestion added by Kees Waterman, DANS

[212] Deliverable 1.2 'State of the art report on digital research practices, tools and scholarly content use', p. 51

[213] Deliverable 1.1 'Research Communities Identification and Definition Report', p. 14

[214] D1.2, p. 14

[215] Chapter 2, section 2.2

[216] D1.5 (1) 'Expert Forum Case Studies Report', 2013, p. 13

- If multilingual content becomes part of Europeana Research a thesaurus or translation tool is needed which automatically links multilingual metadata enabling researchers to find all items associated with a search term.[217]

5.2.6 In conclusion

With Europeana Research, The Europeana Foundation aims to attract a different and specific user audience, which has far-reaching consequence as we have seen above. Until recently, the Europeana Foundation, as instructed by the European Union, was focused on the aggregation of as much digital heritage metadata and content of European origin as possible. This led to an aggregation policy oriented on quantity and diversity (quota suggested by the EU) of the materials entered in the Europeana portal. This policy is reflected in the Foundation's mission: 'To create new ways for people to engage with their cultural history, whether it's for work, learning or pleasure'.[218] From this mission and the policy we can gather that the focus lies on the aggregation of as much content as possible for a rather unspecified audience.

The initial definition of Europeana Research seems to implicitly assume that it would be sufficient to create an audience specific layer of functionality on top of the existing Europeana collection. Our research has shown that this is not the case. To create a useful resource for humanities scholars and social scientists, both the functionality layer and the underlying collection will significantly differ from the standard Europeana and The European Library portals. For the content this means that more detailed metadata is required, and that a significant part of the content included in Europeana is less useful in Europeana Research as it adds 'noise' and not much value. New content, perhaps not traditionally aggregated by neither The European Library nor Europeana, needs to be sourced. Desirable data for aggregation probably needs to be determined in conjunction with each new group of scholars that wishes to engage with Europeana Research. Whether or not this content will also be of interest for the broader Europeana audience as serviced by the Europeana portal very much remains to be seen.

The Foundations' choice to focus on a different user audience in creating Europeana Research asks for a different approach to the aggregation of content and to the policy and mission of Europeana Research. Instead of a focus on quantity and content, the customer will have to be the base around which the service is shaped. And this new target user group, the humanities and social sciences researchers, is a demanding audience used to tailored digital resources and tools. This is a user community which needs to be actively engaged with and reached out to in the further development of Europeana Research, which should be shaped in a way which best fits into their workflow and the research methodologies they use.[219] The need for such close liaisons was also clearly signalled in the Expert Forum Meetings earlier in the project.[220]

---

[217] Ibid., p .9

[218] http://pro.europeana.eu/foundation (15 May 2014)

[219] '10 Things That Humanities Researchers Want', Europeana Libraries, (2011) http://www.europeana-libraries.eu/web/guest/home/-/asset_publisher/5HLm/blog/10-things-that-humanities-researchers-want;jsessionid=95E7BF4FE37829C1FB2F034AE4461244 (11 November 2014)

[220] Deliverable 1.5 (2) 'Expert Forum Tools & Content for Humanities Report', p. 18

Research methodologies, subjects and patterns in digital humanities and social sciences change rapidly, and Europeana Research will therefore have to be highly flexible by design. Another essential condition for the success of Europeana Research will be to stay in close connection with the target audience, not only by the building and maintaining of an active user community, but also by being prepared to engage with the user communities to assist in their efforts to shape their tools and make the most of the rich offerings of Europeana Research.

References

Articles/Blogs/Books/Reports

'10 Things That Humanities Researchers Want', Europeana Libraries, (2011)
http://www.europeana-libraries.eu/web/guest/home/-/asset_publisher/5HLm/blog/10-things-
that-humanities-researchers-want; jsessionid=95E7BF4FE37829C1FB2F034AE4461244 (11
November 2014)

Greta Bahnemann, 'Collection Analysis of the Minnesota Digital Library's'
Minnesota Reflections, June 1, 2012 http://www.mndigital.org/reflections/analysis.pdf (3 June
2014)

Hein van den Berg, Gonzalo Parra, Anja Jentzsch, Andreas Drakos, and Erik Duval. 'Studying the
history of philosophical ideas: supporting research discovery, navigation, and awareness'. In:
*Proceedings of the 14th International Conference on Knowledge Technologies and Data-driven
Business* (i-KNOW '14). ACM, New York, NY, USA, , Article 12 , 8 pages
DOI=10.1145/2637748.2638412 http://doi.acm.org/10.1145/2637748.2638412)(20-11-2014)

British Library, From Stored Knowledge to Smart Knowledge: The British Library's Content Strategy
2013-2015
http://www.bl.uk/aboutus/stratpolprog/contstrat/british_library_content_strategy_2013.pdf

Julie Carpenter, 'Researchers of Tomorrow: the research behaviour of Generation Y doctoral
students', *Information Services and Use* 32(1) (2012), pp. 3-17
http://www.jisc.ac.uk/publications/reports/2012/researchers-of-tomorrow (31 October 2014)

Valentine Charles and Cécile Devarenne, Europeana Foundation, Europeana enriches its data with
the Art and Architecture Thesaurus http://www.pro.europeana.eu/web/guest/home (9 September
2014)

Lynn Silipigni Connaway and Timothy J. Dickey, The Digital Information Seeker: Report of the
Findings from Selected OCLC, RIN, and JISC User Behaviour Projects, (Bristol, UK: JISC, 2010)
http://webarchive.nationalarchives.gov.uk/20140702233839/http://www.jisc.ac.uk/publications/re
ports/2010/digitalinformationseekers.aspx (9 January 2015)

Chelsey Delaney, LeanUX NYC 2014 '*Enhancing proto-personas with characterization'* (workshop)
http://www.slideshare.net/chelsed/leanux-nyc-2014 (16 December 2014)

Alastair Dunning, 'The Great Twentieth-Century Hole Or, what the Digital Humanities Miss', DH
Benelux, June 2014 https://availableonline.wordpress.com/author/xcia0069/ (26 August 2014)

Education for Change Ltd, SIRU, 'Researchers' use of libraries and other information sources:
current patterns and future trends' (Brighton: The Research Partnership; University of Brighton,
2002), pp. 1-84. http://www.rslg.ac.uk/research/libuse/LUrep1.pdf (23 October 2014)

Europeana Communication Team *Essentials: Facts & Figures*, April 2014
http://pro.europeana.eu/about (15 May 2014)

Annette Friberg, Harry Verwayen, *Europeana version 2: Collections and Metadata Analysis, Strategy and Plan 2013*, The Hague: Europeana Foundation, 29 March 2013
http://pro.europeana.eu/documents/866067/1660973/D3.8%3A+Content+Strategy+Update (15 September 2014)

K. Spark Jones and R. Bennett, '*E-resources for research in the humanities and social sciences: a British academic policy review*', (London : British Academy, 2005), pp. 45-46
http://www.britac.ac.uk/policy/eresources/e-resources.cfm (12 January 2015)

L.M. Hughes (ed.), *Evaluating and measuring the value, use and impact of digital collections,* (London, UK: Facet Publishing, 2012), pp. 63-114, 125, 156-181

T. Huyghe, *Waardecreatie door diensten in productiegeoriënteerde ondernemingen.* (Gent : Universiteit Gent, Faculteit Economie en Bedrijfskunde, 2003)
http://lib.ugent.be/fulltxt/RUG01/000/790/291/RUG01-000790291_2010_0001_AC.pdf (4 September 2012)

Maurice Lévy, Elisabeth Niggemann, and Jacques Decker, *The New Renaissance*. Brussels: European Commission, 2011. https://ec.europa.eu/digital-agenda/sites/digital-agenda/files/final_report_cds_1.pdf

Jan Molendijk, *Metadata=better data*, 20 December 2012 http://pro.europeana.eu/pro-blog/-/blogs/1442749/maximized?p_p_auth=y1367oWQ (11 September 2014).

David Nicholas, et al. *E-journals: their use, value and impact,* (London: Research Information Network, 2009), p. 14 http://www.informationr.net/ir/reviews/revs337.html (12 January 2015)

Breandán Ó Nualltáin, Europeana opens up full dataset for re-use, 9 November 2012
http://pro.europeana.eu/web/guest/pro-blog/-/blogs/europeana-opens-up-full-dataset-for-re-use (4 November 2014)

Research Information Network, *Reinventing research? : information practices in the humanities* (London: Research Information Network, April 2011) http://www.rin.ac.uk/our-work/using-and-accessing-information-resources/information-use-case-studies-humanities

Warwick, C., Terras, M., Huntington, P., Pappa, N. 'If You Build it Will They Come? The LAIRAH Study: Quantifying the Use of Online Resources in the Arts and Humanities through Statistical Analysis of User Log Data' *Literary and Linguistics Computing* 1 (2008), pp. 85-102

Karin Michelle Wickett, *Collection/Item metadata relationships: dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Library and Information Science in the Graduate College of the University of Illinois at Urbana-Champaign*, 2012
https://www.ideals.illinois.edu/bitstream/handle/2142/42198/Karen_Wickett.pdf?sequence=1 (23 June 2014)

Oksana L. Zavalina, 'Free-text collection-level subject metadata in large-scale digital libraries: a comparative content analysis', _International Conference on Dublin Core and Metadata Applications, DC-2011, The Hague_, 21-23 September 2011, pp. 147-157
http://dcevents.dublincore.org/IntConf/dc-2011/paper/download/50/19 (3 June 2014)

J. Zundert, "If You Build It, Will We Come? Large Scale Digital Infrastructures as a Dead End for Digital Humanities", _Historical Social Research_, vol. 37, no. 3, (2012)

Websites
Europeana
http://pro.europeana.eu/foundation (15 May 2014)

http://pro.europeana.eu/projects (10 September 2014)

Europeana Cloud
Europeana Cloud: a factsheet for aggregators, September 2013
http://www.slideshare.net/Europeana/europeana-cloud-factsheet-final (8 January 2015)

Competitiveness and innovation framework programme (CIP) 2007-2013: Annex 1: 'Description of Work' : 'Europeana Cloud: Unlocking Europeana's Research via The Cloud'
http://pro.europeana.eu/web/europeana-cloud/results/-/document_library_display/p6BV/view/1861926 (8 January 2015)

Europeana Cloud deliverables can be found at: http://pro.europeana.eu/web/europeana-cloud/results/-/document_library_display/p6BV/view/2240207 (8 January 2015)

Europeana Cloud Deliverable 1.1 'Research Communities Identification and Definition Report', 2013

Europeana Cloud Deliverable 1.2 'State of the art report on digital research practices, tools and scholarly content use', 2013

Europeana Cloud Deliverable 1.3 'User requirements analysis and case studies report', 2013 Link to Expert Forums: http://pro.europeana.eu/web/europeana-cloud/results/-/document_library_display/p6BV/view/1861918 (9 January 2015)

Europeana Cloud Deliverable 1.5 (1) 'Expert Forum Case Studies Report', 2013

Europeana Cloud Deliverable 1.5 (2) 'Expert Forum Tools and Content for Humanities Research Report', 2013

Europeana Cloud Deliverable 1.5 (3) 'Expert Forum Tools and Content for Social Science Research', 2013

Europeana Cloud Deliverable 3.1 'Personas, scenarios and use cases', 2013

Europeana Cloud Deliverable 3.3 'Evaluation report on integration of tools with Europeana', 2013

Europeana Cloud Deliverable 6.1 'Stakeholder Engagement & Infrastructure Plan', 2013

Europeana Cloud Milestone for Task 1.3.5 'Evaluating Use of Innovative Tools in Research', 2014

Europeana Cloud: blogs
Alastair Dunning, 'Drafting priorities for Europeana Research', 30 July 2014
http://pro.europeana.eu/web/europeana-cloud/blog/-/blogs/trackback/drafting-priorities-for-europeana-research (20 November 2014)
Friedel Grant, 'Visualising The Architecture Behind Europeana Cloud', 24 February 2014
http://pro.europeana.eu/web/europeana-cloud/blog/-/blogs/visualising-the-architecture-behind-europeana-cloud (14 November 2014)

The European Library

The European Library http://www.theeuropeanlibrary.org/tel4/ (19 May 2014)
Friedel Grant, Q&A: Partner Profile - The European Library
http://pro.europeana.eu/web/europeana-cloud/blog/-/blogs/2175228 (19 May 2014)
http://www.theeuropeanlibrary.org/tel4/discover (3 December 2014)

Paper 8: Update report on developing a content strategy for The European Library, Management Committee teleconference 13th December 2010.

The European Library - 2013 Annual Report & Accounts, June 2014
http://www.theeuropeanlibrary.org/tel4/news (10 June 2014).

General
http://www.cendari.eu/about-cendari/ (6 August 2014)

http://enumeratedataplatform.digibis.com/figures/ (6 August 2014)

http://en.wikipedia.org/wiki/Secondary_research (3 December 2014)

Unpublished Sources
Europeana Cloud Research Community Web Survey T1.3.4, 2013

Europeana, 'Europeana Content Analysis Summer 2010',
http://www.pro.europeana.eu/documents/869571/874106/Content_Analysis_-_Summer_2010.doc, 2010, pp. 1-42 (14 October 2014)

Appendix A List of surveys

The European Library portal

The TEL Collections Descriptions Form represents the overview of the collections in TEL:
https://docs.google.com/spreadsheet/ccc?key=0ArFeVeAoD0YBdG5zWUhTVjVxSkhPQ0p1Sl9fSlhpS
FE&usp=drive_web#gid=0

Link to the responses:
https://docs.google.com/a/cerl.org/spreadsheet/ccc?key=0ArFeVeAoD0YBdHJaOWdhZzNIbXR2Y3B
qanlfOU5CN0E&usp=drive_web#gid=0

Europeana portal

TCD Content Survey:
https://docs.google.com/a/cerl.org/forms/d/1HzQgz-
lblrz6I42nSH5tkQ1neVrEUmABYwXo9ot8GKE/edit

TCD Content Survey responses:
https://docs.google.com/spreadsheet/ccc?key=0Au8WbVWw0GTRdDl6b2xLSWhFYWtRLXVDRVlOM
GhtNGc&usp=sharing

TCD eCloud Wp1 'Europeana Content Survey' Methodology notes
https://docs.google.com/a/cerl.org/document/d/1kon0_zkW8SsnbPvz-3ugu-
WuW8tshsuFnPcin8O8yZ8/edit

Original Europeana Survey:
https://docs.google.com/a/cerl.org/forms/d/12LMyQxQTGmVN9qqxdwDGRvPWfp_iHgf_d3TzNwtB
pV4/viewform?edit_requested=true

Responses to original survey:
https://docs.google.com/a/cerl.org/spreadsheet/ccc?key=0ArFeVeAoD0YBdE1YSkJIT2hfMjZQR285
QUxhdGxVbVE&usp=drive_web#gid=0