D 7.2 Design and Sustainability Plan for an Open Humanities Data Platform

Stefan Buddenbohm, Maaike de Jong, Jean-Luc Minel, Mike Priddy, Nicolas Larrousse, Yoann Moranville

D 7.2 Design and Sustainability Plan for an Open Humanities Data Platform – Concept for a Data Deposit Recommendation Service

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

Date: 24-05-2017
Version: 1.1

Project funded under the Horizon 2020 Programme

Grant Agreement no.:	675570
Programme:	Horizon 2020
Project acronym:	HaS-DARIAH
Project full title:	Humanities at Scale: Evolving the DARIAH ERIC
Partners:	· DIGITAL RESEARCH INFRASTRUCTURE FOR THE ARTS AND HUMANITIES · CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE · KONINKLIJKE NEDERLANDSE AKADEMIE VAN WETENSCHAPPEN – KNAW · GEORG-AUGUST-UNIVERSITAET GOETTINGEN STIFTUNG OEFFENTLICHEN RECHTS
Topic:	INFRADEV-3-2015
Project Start Date:	01-09-2015
Project Duration:	28 months
Title of the document:	D7.2 Design and Sustainability Plan for an Open Humanities Data Platform - Concept for a Data Deposit Recommendation Service
Work Package title:	Open Data Infrastructure
Estimated delivery date:	31.12.2016
Lead Beneficiary:	UGOE-SUB
Author(s):	Stefan Buddenbohm (buddenbohm@sub.uni-goettingen.de) Maaike de Jong (maaike.de.jong@dans.knaw.nl) Jean-Luc Minel (jean-luc.minel@u-paris10.fr) Mike Priddy (mike.priddy@dans.knaw.nl) Nicolas Larrousse (nicolas.larousse@huma-num.fr) Yoann Moranville (yoann.moranville@dariah.eu)
Quality Assessor(s):	Claudia Engelhardt (claudia.engelhardt@sub.uni-goettingen.de) Claudio Leone (leone@sub.uni-goettingen.de)
Keywords:	open data, open humanities data platform, research data, research infrastructure, sustainability, repository registry, recommender, data deposit

Revision History

Version	Date	Authors	Beneficiary	Description
0.1	31-08- 2016	Stefan Buddenbohm Jean-Luc Minel Maaike de Jong	UGOE-SUB CNRS DANS-KNAW	First draft
0.2	21-09- 2016	Stefan Buddenbohm Jean-Luc Minel Maaike de Jong	UGOE-SUB CNRS DANS-KNAW	Second draft
0.3	30-09-2016	Stefan Buddenbohm Maaike de Jong Jean-Luc Minel Nicolas Larrousse Julius Peinelt	UGOE-SUB DANS-KNAW CNRS CNRS DARIAH-EU	Third draft to be discussed at the project meeting in Ghent
0.4	28-11-2016	Stefan Buddenbohm Maaike de Jong	UGOE-SUB DANS-KNAW	Stable revised edition to be discussed and agreed upon in the project consortium
1.0	24-05-2017	Stefan Buddenbohm Maaike de Jong Mike Priddy Yoann Moranville	UGOE-SUB DANS-KNAW DANS-KNAW DARIAH-EU	Final report
1.1	04-05-2018	Yoann Moranville Stefan Buddenbohm	DARIAH-EU UGOE-SUB	Chapter 6 modified following the review report and recent developments of the DDRS.

Table of contents

Executive Summary

The Data Deposit Recommendation Service (DDRS) intends to help the user to identify suitable research data repositories depending on case-specific requirements. As an added value service, the DDRS offers the initiation of the ingest and communication process between user and repository by forwarding a deposit request along with a structured description of the research data to the appropriate point of contact.
About the nature of this document: This deliverable follows the report "Deliverable 7.1 State of the Art Report on Open Access Research Data for the Humanities"[1] in the Humanities at Scale (HaS) work package 7 "Open Data Infrastructure". It forms the concept for the "D 7.3 Open Data in the Humanities Platform" and prepares the implementation of the service. The Open Data Platform has been refined to a Data Deposit Recommendation Service (DDRS). The reasoning behind this process and the concept for the service are described in this document[2].

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

Disclaimer

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

[1] The report is freely available under: https://halshs.archives-ouvertes.fr/halshs-01357208

[2] This report is freely available under: https://halshs.archives-ouvertes.fr/search with the keyword Humanities at Scale

1 Introduction and scope of the report

Humanities at Scale (HaS) is an offspring project of DARIAH-EU[3]. The project functions as catalyst activity for the already existing Digital Humanities resources, networks, research data, services and infrastructures at the European level. Whereas many initiatives have been developed at a national level, for instance DARIAH-DE[4] in Germany, Huma-Num[5] or Open Edition in France, CLARIAH and EASY[6] in the Netherlands, they have been designed with a transnational and European perspective in mind. The main goals of the HaS project[7] are:

- Scaling up the DARIAH community by integrating new research, more research data and methods and new regional communities into the DARIAH cosmos.
- Improving the sustainability and usage of funding for digital humanities but also exploring alternative funding models.
- Expanding the knowledge within the digital humanities by means of a pan-European training programme and by summer and winter schools, particularly in regions without a longstanding tradition in the digital humanities.
- Developing core services that allow better access to the DARIAH contributions from different member states.
- Supporting research in the digital humanities with basic infrastructure services to facilitate better integration of digital humanities projects with DARIAH, meaning that researchers can easily use the infrastructure and services.
- Facilitating open access in the domains of research data (open data) and methods (open methods).
- Informing the stakeholders in the digital humanities and other research communities of its results.

With the HaS project, DARIAH-EU seeks to connect with the open access movement in the European Union. The project aims to intensify the collaboration with open access initiatives and will support the implementation of corresponding services within the arts and humanities. The goal of HaS Work Package (WP) 7 'Open Data Infrastructure' is to develop an Open Humanities Data Platform where communities in DARIAH can develop their understanding of open data, locate suitable repositories and can promote their data. An Open Humanities Data Platform can facilitate one or more aspects of the research data life cycle (Figure 1). The main aim of the platform as defined in the HaS WP7 Description of Work is to develop a registry for collections and research data. This way, scholars can access a suitable repository to deposit their research data and then promote it within the DARIAH community for others to discover, reuse and enrich it.

[3] http://dariah.eu/

[4] https://de.dariah.eu/

[5] http://www.huma-num.fr/

[6] https://easy.dans.knaw.nl

[7] http://has.dariah.eu/?page_id=7

This report is the deliverable for Task 7.2, and presents the design and sustainability plan for an Open Humanities Data Platform that was developed during this task. Task 7.2 is a joint effort of the University of Göttingen – State and University Library (UGOE-SUB), Data Archiving and Networked Services (DANS) and Centre National de la Recherche Scientifique (CNRS). In this report, we start by outlining stakeholders and users in the context of an open data platform in Chapter 2. This is followed by an overview of existing humanities and general open data platforms in Chapter 3. We then present the different possible functional scenarios and implementation approaches for a platform in Chapter 4, while discussing factors such as usefulness to the community and sustainability aspects. Chapter 5 presents the concept and use cases of the chosen platform approach, which is following by Chapter 6 which details the technical implementation of the platform. Finally, Chapter 7 discusses the possibilities for future developments and the sustainability of the platform. This plan will serve as a guideline for the implementation of the platform during Task 7.3, the final task of WP7.

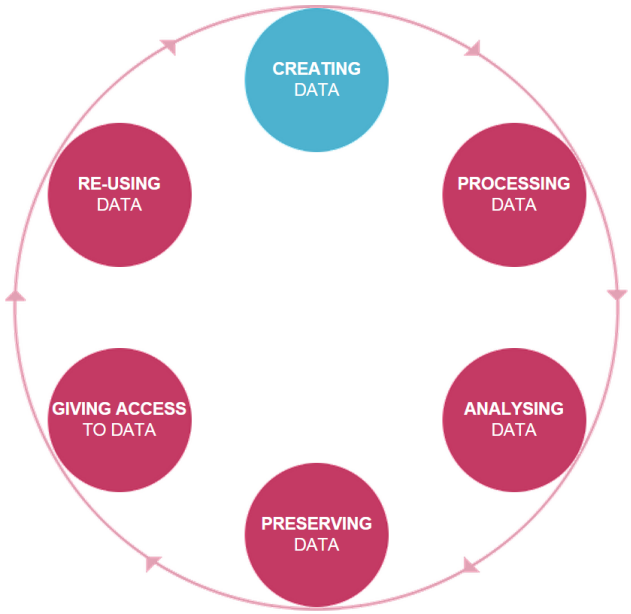

Figure 1: Research data lifecycle - (c) UK Data Archive[8], starting with the creation of research data

Figure 1: The Research Data Lifecycle. The data lifecycle begins with a researcher developing a concept for a study; once a study concept is developed, data is then collected for that study. After data is collected, it is processed for distribution so that it can be archived and reused by other researchers. Once data reaches the distribution stage of the lifecycle, it is stored in a location (i.e. repository, data archive) where it can then be discovered by other researchers. Data discovery leads to the repurposing of data, which creates a continual loop back to the data processing stage where the repurposed data is archived and distributed for discovery.[12]

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

2 Stakeholders and users of an Open Humanities Data Platform

A key factor to consider in the choice and design of an Open Humanities Data Platform is which stakeholders are involved and what their particular interests or requirements are. Below we give an overview of this for the main stakeholders and users in this context.

- Researchers (and associated research institutions) are the core users of an Open Humanities Data Platform: they are the main data producers as well as consumers of digital research data. As data sharers, they need to trust that their data is preserved, accessible, and useable in the long term. As data users, the main concerns are the ability to find the data, and the authenticity and quality of the data. An Open Humanities Data Platform could facilitate work for researchers in all stages of the research data lifecycle: data management best practice, finding, reusing, depositing and publishing/ promoting research data.
- Digital repositories make data findable, accessible, and usable in the long-term, by e.g. using sustainable file formats, and providing persistent identifiers and informative descriptive data (metadata). Related to this are online data platforms that do not store data, but bring together metadata of research datasets, making them findable for data users.
- Galleries, libraries, archives and museums (GLAMs) are important data providers in the humanities. Their main concerns lie in preservation of their collections and making their resources available to the general public, and secondarily in providing support to researchers.
- Other digital infrastructures (national/international): Other national or international infrastructures are relevant to HaS in terms of a possible cooperation, concerning e.g. the integration or reuse of components within HaS services, interoperability issues or extensions. One main aspect is to design cooperation with mutual use or benefit and to foster synergies in the field of providing (data) services, relevant information and recommendations to relevant target groups. This includes also the use and enrichment of already existing databases.
- Other service providers, such as data curation experts, or providers of training in digital methodologies, or higher education in digital humanities. Training and education - although not at first sight integrated within the development - form an important space for dissemination, feedback and stimuli for the improvement of an infrastructure or service.
- Research funding agencies benefit from promoting the optimal use and reuse of data in which funds were invested. They can do this by encouraging good data practices, investing in data infrastructure and raising data awareness. Funding agencies, both at the European and national level, increasingly require the research data (and publications) resulting from funded research projects to be

published open access. For example, the EU obliges researchers funded by Horizon 2020 to publish their research data as open data[9].

- Policy makers, i.e. national governments, and the EU, increasingly have Open Access on the political agenda and are driving research data publishing top down by adapting science policies, often implemented via the national and EU funding bodies (see above).
- Academic and other publishers: Academic publishers impose requirements on the availability of data connected to submitted and/or published papers, and provide identifiers to cite papers and link to related data. Non-academic publishers (for example societies) are also important in the humanities, however, for these the availability of data connected to publications is often less clear.
- Humanities data consumers: These can include e.g. education practitioners, journalists and the general public. These users can access source data, research findings and educational tools through an open data platform in the humanities. This also applies to educators and teachers interested in humanities, as well as NGOs and humanitarian organisations. The general public is also increasingly involved in producing data through e.g. involvement in citizen science.

Table 1 gives an overview of the stakeholder groups (description and/or examples), their interests in an Open Humanities Data Platform, and their relative importance when considering the functional requirements of the platform. It is clear that the main users will be (in order of importance): researchers, data content holders (in particular digital repositories), and other research infrastructures (in case these have functional links with the platform). Other stakeholders will use the platform less frequently (e.g. education practitioners, journalists, general public) or only have an indirect interest (e.g. research funders, policy makers).

Stakeholder group	*Description / examples*	*Interest in platform*	*Importance*
Researchers	Academics and other researchers in the arts and humanities, regardless their institutional affiliation	Services and tools for finding, reusing and depositing data; data management; data publishing/promoting; information on current trends and standards in RDM (partly funder-driven)	Very high (particularly academic researchers)
Data content holders	Digital repositories; galleries, archives, libraries, museums (GLAM)	Enhancement of visibility and usability of their collections	High
Other digital research	Depending on the platform type, e.g. registries of services	To collaborate, improve services, content and expertise;	Medium

[9] http://europa.eu/rapid/press-release_IP-16-1408_en.htm

infrastructures	such as Re3data.org	increase user base; perspective for added value services	
Other research service providers	Depending on the use proposition, e.g. analysis of research data practices or potential new functions	To collaborate, improve services, content and expertise; increase user base; perspective for added value services	Low
Research funders	International (e.g. European Commission) and national funding bodies	Better use of funding through reuse of data; Improved data management practices; usage statistics	Low
Policy makers	Governments; EU; national/transnational research frameworks	Advancement of Open Access, Open Data; similar objectives like the research funders because of institutional overlaps	Low
Publishers	Academic publishers (e.g. Elsevier); Non-academic publishers; Open publishing platforms	Improved availability and findability of data connected to publications	Low
Humanities data consumers	E.g education practitioners; journalists; the general public; currently not relevant target groups	To find and use humanities (research) data; all possible future uses of humanities data beyond the realm of research	Low

Table 1. An outline of the main stakeholder groups, their description and examples, their main interest in an open humanities data platform, and their relative importance for the design of the platform.

3 Overview of existing platforms

3.1 Open humanities data platforms

Before the conception of any service or infrastructure one has to gain an overview of the already existing landscape in order to identify gaps or avoid redundancies and in general, to get a sense of the competitors. This applies especially to the scientific domain as usually public money is spent. The DARIAH context and the humanities specific research infrastructure are the field to look at in this respective context. There are plenty of projects, infrastructure initiatives and in general, the strive towards standards and common infrastructures. This landscape study has already been largely conducted in the deliverable 7.1 State of the Art Report on Open Access Publishing of Research Data in the Humanities[10] and will therefore not be repeated here. Instead, only a concise overview of

[10] The report is freely available under: https://halshs.archives-ouvertes.fr/halshs-01357208

the most relevant infrastructure players in the field is given here. The label 'most relevant' is partly subjective and related to the DARIAH context of the project HaS.
In Germany, the Netherlands and France, several national open data platforms for the humanities have already been or are about to be developed. These provide effective and relevant open data services for arts and humanities researchers and largely cover the needs of an open data platform in these disciplines:

- DARIAH-DE Repository: DARIAH-DE is developing a research infrastructure in support of service and research data as well as materials for research and teaching in the digital humanities. DARIAH-DE is the German national contribution to the European research infrastructure "DARIAH-EU - Digital Research Infrastructure for the Arts and Humanities" within the framework of ESFRI[11]. One major service pillar of DARIAH-DE will be the DARIAH-DE repository, able to ingest research data from the arts and humanities[12]. This function is particularly interesting for the platform concept we present later in this report.
- Huma-Num: Huma-Num[13] is in charge of the Very Large Infrastructure (VLRI) dedicated to social sciences and humanities operating at a national level in France. Huma-Num coordinates French national contributions to the European research infrastructure "DARIAH-EU - Digital Research Infrastructure for the Arts and Humanities" within the framework of ESFRI national roadmap. Huma-Num offers a range of services for research data for their preservation and reuse. The two main services being NAKALA[14] and ISIDORE[15]. NAKALA is a repository able to ingest research data from the arts and humanities in order to share data and metadata using Semantic Web technologies and OAI-PMH. NAKALA also provides a PID in order to make data citable. ISIDORE is an aggregator that harvest more than 4000 sources. ISIDORE process metadata which are enriched, classified and aligned with common LOD (Linked Open Data) repositories entries like the one from BNF (French National Library): the main goal is to disseminate data to make them discoverable and "unforgettable" to facilitate reusability using Semantic Web technologies.
- EASY repository: the online archiving system EASY[16] is hosted by DANS (Data Archiving and Networked Services), the Netherlands institute for permanent access to digital resources. This repository offers access to thousands of datasets in the humanities, the social sciences and other disciplines. EASY is a DSA- and WDS-certified Trusted Digital Repository – giving an indication of quality, preservation and accessibility of data. DANS also provides access to thousands of

[11] https://de.dariah.eu/dariah-de-english

[12] https://de.dariah.eu/repository

[13] http://www.huma-num.fr/

[14] See HAS D7.1 https://halshs.archives-ouvertes.fr/halshs-01357208/document (p. 42)

[15] See HAS D7.1 https://halshs.archives-ouvertes.fr/halshs-01357208/document (p. 43)

[16] https://easy.dans.knaw.nl

scientific datasets, e-publications and other research information in the Netherlands via NARCIS[17], the national portal for scientific information.

- At the European level, PARTHENOS[18] is a H2020 project dedicated to cultural heritage data and DARIAH-EU participates in the project. It covers the specific area of standardisation in the areas of documentation of primary data and sources, reference resources and procedures and protocols. PARTHENOS also addresses interoperability and semantics which involves defining a common semantic framework, the integration of multi-lingual reference resources and designing resource discovery and will define the technical development of the tools and services that are required to create the desired trans-humanities research infrastructure. In addition, training material will be provided along with best practice and documentation guides. The approach is mainly distributed among the various services in the consortium[19] and promotes the FAIR principles.

3.2 General open data platforms

In addition to open data platforms with a specific focus on the humanities, there are various generic and international open data platforms: including Dataverse, Dryad, EUDAT, FigShare, Mendeley Data, and Zenodo (descriptions of the repositories are retrieved from the Registry of Research Data Repositories, except for Mendeley Data):

- Dataverse[20] : Dataverse, a repository software, has been developed by Harvard University. 'The Harvard Dataverse is open to all scientific data from all disciplines worldwide. It includes the world's largest collection of social science research data. It is hosting data for projects, archives, researchers, journals, organisations, and institutions.' There are many communities that work together in platforms of Dataverse, for example the Dutch DataverseNL, a cooperation of nine institutions using the Dataverse platform.
- Dryad[21]: DataDryad.org is a curated general-purpose repository that makes the data underlying scientific publications discoverable, freely reusable, and citable. Dryad is an international repository of data underlying peer-reviewed scientific and medical literature, particularly data for which no specialized repository exists. The content is considered to be integral to the published research. All material in Dryad is associated with a scholarly publication.
- EUDAT[22]: The EUDAT project aims to contribute to the production of a Collaborative Data Infrastructure (CDI). The project´s target is to provide a pan-European solution to the challenge of data proliferation in Europe's scientific and

[17] www.narcis.nl

[18] http://www.parthenos-project.eu

[19] http://www.parthenos-project.eu/consortium

[20] http://dataverse.org/

[21] http://datadryad.org/

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

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

- Figshare[23]: Figshare allows researchers to publish all of their research outputs in an easily citable, sharable and discoverable manner. All file formats can be published, including videos and datasets. It offers an optional peer review process. Figshare uses creative commons licensing. Figshare also contains research data in humanities.
- Mendeley Data[24]: The platform allows researchers to upload the raw data from their research, and give it a unique identifier (a versioned DOI), making that research citable. For partnering journal websites, the article links to the research dataset on Mendeley Data, enabling readers to quickly drill down from a research article to the underlying data; while the dataset also links to the article. Researchers can also privately share their unpublished data with collaborators, and make available multiple versions of the data relating to a single research project, creating an evolving body of data.
- Re3data.org[25]: re3data.org is a global registry of research data repositories that covers research data repositories from different academic disciplines, funded by the German Research Foundation (DFG). It presents repositories for the permanent storage and access of data sets to researchers, funding bodies, publishers and scholarly institutions. re3data.org promotes a culture of sharing, increased access and better visibility of research data. Some publishers and journals like Copernicus Publications, PeerJ, Springer and Nature's Scientific Data refer to re3data.org in their Editorial Policies as a tool for the easy identification of appropriate data repositories to store research data. The use of re3data.org is also recommended in the European Commission's "Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020"[26].

[23] https://figshare.com/

[24] https://mendeley.com/

[25] http://www.re3data.org/

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

- Zenodo[27]: Zenodo builds and operates a simple and innovative service that enables researchers, scientists, EU projects and institutions to share and showcase multidisciplinary research results (data and publications) that are not part of the existing institutional or subject-based repositories of the research communities. Zenodo enables researchers, scientists, EU projects and institutions to: a) easily share the long tail of small research results in a wide variety of formats including text, spreadsheets, audio, video, and images across all fields of science; b) display their research results and get credited by making the research results citable and integrate them into existing reporting lines to funding agencies like the European Commission; c) easily access and reuse shared research results.

As outlined in the paragraphs above there is a diverse landscape of relevant infrastructures and services. This has to be taken into account when conceptualising the HaS Open Humanities Data Platform. The concept has not only to pay attention to the service and functions as such but also to the sustainability issue which is of particular importance within the DARIAH context as the resources to sustain infrastructures are sparse and have to be justified by clear benefits, for example in terms of usage, interoperability.

4 Design and sustainability scenarios for an Open Data platform

This chapter discusses the different types of approaches and functions that could be implemented through an Open Humanities Data Platform, considering the fit with DARIAH's scope, the usefulness to the community, the potential for possible future service extensions, and importantly, the sustainability aspects. First, we give an overview of different functional scenarios that are possible within the broad concept of an open data platform. We then describe two contrasting approaches to the possible technical, software and organisational architecture of the platform, a centralised and a distributed approach, and discuss which is the most suitable approach for the Open Humanities Data Platform.

4.1 Functional scenarios for the platform

In this design and sustainability plan, we consider the following functional scenarios with distinct services that an Open Humanities Data Platform could offer and that could be developed during the HaS project, taking into account factors such as technical implementation, already existing services, and sustainability aspects (summarised in Table 2). Against this background, the specific implementation approach has to be chosen.

- Knowledge Base: In this scenario, the platform would function as a knowledge base for information on research data, such as standards, recommendations, definitions, reports, best practices, tools, and training materials. Such a platform,

[27] https://zenodo.org/

which is essentially a website with information, can be implemented quite easily. An obvious point related to the sustainability is the challenge of keeping the information provided by such a platform current. Sustainability for knowledge bases is not primarily a question of technical maintenance and security, but the main costs and challenges lie in the necessity of sustained editorial supervision and fostering of the platform. Existing relevant knowledge bases include Isidore[28] (French, English and Spanish language), forschungsdaten.org (German language), and Open-Access.net.

- Brokering or recommendation function: The platform can also function as a brokering or recommendation hub for expertise and consultation. Different from the knowledge base-function, human resources are in the foreground in this scenario. In this regard brokering is to be understood as connecting researchers with an individual consultation demand with experts/data curators who are competent in the respective area and subjects. This function could make use of the existing DARIAH community. The implementation of brokering or recommending external resources can be conducted in phases. At first a recommendation tool could - for example - point to registries or repositories. In a second step this recommendation could be enriched with cost estimations. The technical implementation of the brokering function is more complex than that of the knowledge base. A database with a search interface could initially serve as proof-of-concept, however, a recommendation function would provide more benefit for the user. This recommender would require a more complex database solution that guides the inquirer through a set of questions and then presents its results. The long term-maintenance depends on the complexity of the platform function. If only a database with a query form is implemented, regular maintenance and content updates are needed at a smaller level. If a more complex recommender function is implemented that brokers experts, not only the database functionalities need to be more sophisticated but also the underlying mechanism of accounting the services need to be covered. This function could serve as a hub for processing DARIAH in-kind contributions. This means a research institution could register experts for certain topics and the "deployment" of these experts to inquirers could be invoiced via the in-kind contribution scheme. An example of services developing in the direction of user support/recommendation may be found at CLARIN[29], where the user gets a proposal for a research data management plan after completing a short questionnaire. Another example is provided by the German GFBio-project[30] which is developing a recommender tool for depositing research data from the Biology.
- Access to research data: An Open Humanities Data Platform as access point to research data could either only link to other registries, data centres or

[28] https://www.rechercheisidore.fr/

[29] https://www.clarin-d.de/en/preparation/data-management-plan

[30] https://www.gfbio.org/

repositories, or ingest and disseminate research data into its own repository. In the first case the platform could be integrated with a third-party service that specialises in access to research data, for example an established harvesting service. The technical challenge here lies in the broad range of formats of research data. Whereas it is relatively easy to set up such a service for publications - as they are fairly standardized - research data comes in broad and heterogeneous formats. The technical implementation of this service depends on whether the research data would be ingested into and disseminated from the platform. If the platform requires a research data repository at the core of the services, the technical implications are considerably larger than those of a service linking to existing external sources of data. This is likely a service that would not be implemented by the HaS project itself but rather by a third party which HaS is allowed to use. It is to be decided in which direction a demonstrator for this function should be developed. Depending on the specific implementation, the effort and therefore costs for harvesting and harmonizing of metadata, describing research data, could be considerable. A good example of a discovery service pointing to research publications from all research disciplines is Base.net[31]. Another example, more aiming at repositories than research data, is Re3data.org.

- Registry of Tools and Services: The Open Humanities Data Platform could serve as registry or catalogue for services and tools which can, for example, be utilised with research data in the humanities, or to create research data. These services and tools can be entire infrastructures such as repositories, research data centres or registries, workflows of data transformation tools, or they can be smaller services like data management plan support tools, cost calculators, or mapping tools for metadata schemes. This function could be connected to the HaS WPs 6 and 8 which also work on tools and services. In case of WP6 this involves basic services for humanities research, while WP8 specifically works on a decentralised registry that harvests and displays metadata on DH tools and services that are implemented by RDFa directly in the websites of the providers[32]. A registry would - similar to the brokering function - consist of a catalogue which can be queried by the user. Individual search results should point to useful resources for the user. Generally, the tools will be third party services. From the sustainability perspective, ongoing editorial support is necessary after implementation. If the function is integrated with the knowledge base, the user community of the platform could also be involved in keeping the information up to date. Existing relevant services are Re3data.org, an already well-established registry for data

[31] www.base-search.net

[32] To put it simple: the service is based on harvesting RDFa code snippets from external websites and merging them on the service's website in a structured way for further use.

repositories, DARIAH Collection Registry[33], ROAR Registry of Open Access Repositories[34], and Dh-projectregistry.org[35].

	Description of function	*Possible technical Implementation*	*Sustainability issues*	*Examples*
Knowledge Base	information and recommendations on standards, best practices, tools and training	set up of a website / wiki platform (e.g. Confluence) with information	High relevance of long term editorial supervision and fostering (main factor causing costs)	Isidore Open-Access.net forschungsdaten.org
Brokering Function	bringing together researchers and experts/data curators for individual consultation including existing DARIAH community	database able to narrow down to recommendations and brokering to experts (including accounting)	basic solution: regular maintenance and content update; DARIAH in-kind contribution: institutions registering experts	CLARIN GFBio
Access to Research Data	access point linking to other deposit infrastructures (basic solution); ingest and dissemination of research data via an own repository (optional)	web instance linking resources or harvesting metadata from data providers; platform could in a later building phase be an access point to the research data itself	service likely implemented by a third party HaS is allowed to use; harvesting and harmonising of metadata and describing research data (great effort)	Re3data.org DANSdatajournal.nl
Registry of Tools and Services	registry or catalogue for services and tools to handle research data in the humanities or used methodologically to create research data	catalogue which can be queried by the user pointing to useful resources (mainly third party services)	ongoing editorial support will be necessary; involve the community to keep the information up to date (integration with the knowledge base)	Re3data.org ROAR Dh-projectregistry.org

Table 2: Overview of functional scenarios for the Open Humanities Data Platform

4.2 Distributed vs. centralised approach

The attributes centralised vs. distributed relate particularly to the development, deployment, maintenance and management of services and infrastructure. In a

[33] https://colreg.de.dariah.eu/colreg/

[34] http://roar.eprints.org/

[35] http://dh-projectregistry.org/

centralised scenario, the development work, the hosting of services and the sustainability of the overall framework is the responsibility of one main infrastructure provider. Such a scenario almost necessarily involves a centralised governance, not least because of administrative and institutional constraints. A distributed approach means that various functions are performed by different existing services in different national and/or international infrastructures. On the one hand, this scenario requires considerable attention to interoperability between existing services, and on the other hand, a workflow must be developed in order to dispatch the information between different existing services.
Table 3 presents an overview of the strengths and risks of each conceptual approach for the purpose of comparison. In reality hybrid forms are more likely to occur.

	Centralised	*Distributed*
STRENGTHS	• Simplicity of a central management structure. • Branding and dissemination issues under control. • Strategic development of infrastructure under full control. • Full control on potential added-value services.	• Distributed responsibility. • Division of labour, greater effectiveness. • Resilience at a failure of a partner. • Reusing of existing tools possible. • Gradual improvement. • Lower costs of development compared to a centralised approach. • Embedding in existing communities, standards and infrastructure networks is relatively frictionless compared to a centralised approach.
RISKS	• Dependent of the hosting institution or necessity for a hosting policy. • High costs of development. • Risk of technological stalemates. • Higher dissemination efforts necessary as a completely new service has to be published.	• Management interoperability. • Coordination issues. • Conflict management. • Transaction costs and business model. • Interoperability issues as not all areas are under control. • Software maintenance & management overheads are higher. • Gaps occur or appear in the overall provision.

Table 3: Comparison of the centralised and the distributed approach

4.3 Reasoning for the distributed approach

The HaS project is connected to manifold initiatives by universities, infrastructure providers and projects to foster the use of digital research infrastructures and dissemination as well as the reuse of research data. The project works from the premise that the management of research data has moved into the focus of university libraries and data centres which apply their experiences in building data repositories and related services to support the research community. Consequently, it is not only embedded in the

broader context of DARIAH, which is geared towards the Digital Humanities, but also in the research data management activities of various institutions such as libraries or data centres. One of the aims of HaS is therefore to consider a service approach for the community, meaning to refine and employ existing standards, infrastructures and best practices to meet the researcher in their current working environments.

A centralised approach would have the benefit of being a newly developed platform and therefore not having a historically grown technical burden. Also, since it would be developed in-house at one place, DARIAH-EU would be free to choose the technical solutions and make other decisions directly without the need to discuss with other partners, which would reduce management issues during development. However, these advantages are opposed by several downsides. The maintenance costs of a large and centralised platform would be concentrated on DARIAH-EU alone, the design and development of such a complex platform is very time consuming, and after releasing the platform it would need to be introduced to the community to gain users.

The preferred alternative is to use a distributed approach by leveraging already existing platforms and connecting them. This implicates that several parties will be involved and DARIAH-EU will not have a direct, overall influence. Instead, part of the planning and development will include compromises with partners and providers of connected platforms. This should be seen as a virtue since these partners know their users and the problems they are facing. Related to this, a distributed approach can build on the existing user base of connected platforms and therefore can answer to the needs of different user groups. Another major benefit of using a distributed solution is that the basic functionalities are provided by the partners and do not need to be developed from the ground up. Only the connection between the features needs to be designed and implemented, which means that the resulting platform can be released faster to users, is easier to extend and can grow in functionality over time. Therefore, rather than developing a new platform, it seems preferable to aggregate data produced by these existing platforms (see the table above) in order to increase their access of the research community in the arts and humanities. This also applies to services and tools or to expertise, reflecting the nature of DARIAH as a network and community. Each platform node has its own specialisms and together make up a diverse network which facilitates services to a broader spectrum of researchers and thus increases its user-base and impact. In the humanities one size does not fit all. If any single platform provider decides to end its provision then the remaining platforms are not affected and the majority of the service continues. Moreover, it may be possible for one of the remaining platforms to take over the provision of the service from provider who is stopping, thus maintaining the sustainability of the distributed service.

5 The Data Deposit Recommendation Service

5.1 Function and concept

Considering the open humanities data landscape outlined in the previous chapters, and the limited available resources in the HaS project, we decided to go forward with a pragmatic and trim approach to the open humanities data platform. An important motivation behind WP7 was to address the specific character of DARIAH and therefore pay particular attention to the interoperability of the platform and its service. Summarising the discussion on the specific nature of the platform, we came to the conclusion to:

- focus on one or at least only a few functionalities.
- to base the service from the very beginning on using third party services.
- to keep it interoperable especially with regard to research data repositories.

These deliberations are reflected in the platform concept described below: the Data Deposit Recommendation Service (DDRS).

The DDRS is geared towards researchers and research projects from the arts and humanities, especially from the digital humanities. It addresses the question of how and where to deposit research data, an issue increasingly gaining importance as reuse of research data becomes more common and more funders require (open) publishing of data taking into account the aspect of reproducibility of research. The user experience of the service should be kept as straightforward as possible. Through a guided concise questionnaire, the system recommends the best suited data repositories for the individual case.

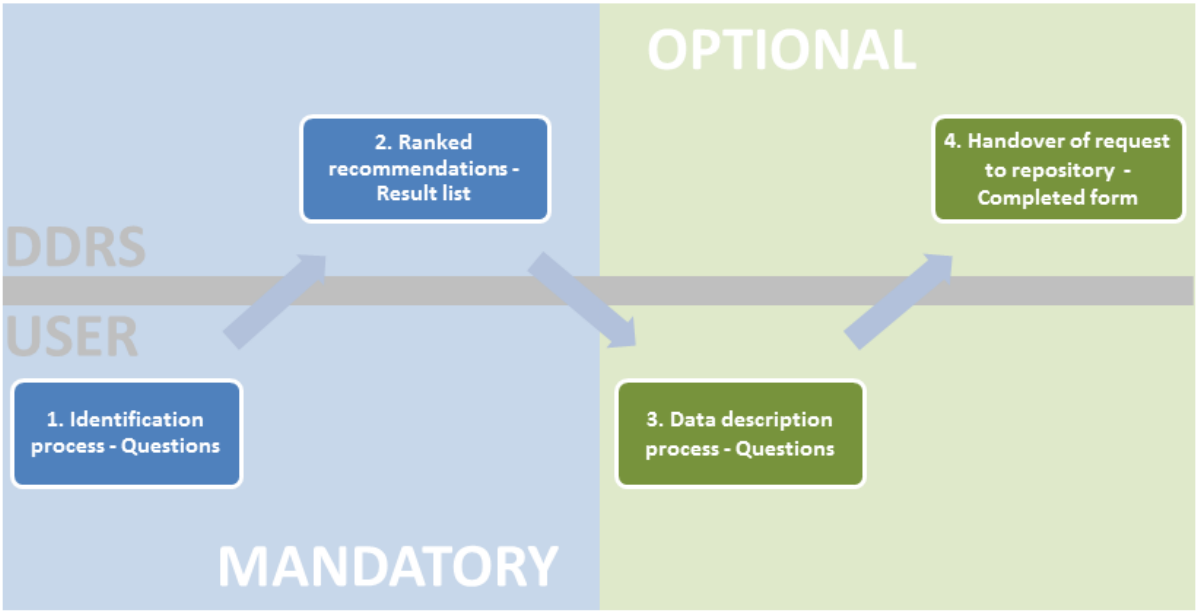

Figure 2: Concept of the DDRS as two-tier-concept

Beyond this main objective, the DDRS also aims to increase visibility of research data repositories and to improve collaboration and interoperability between such services. An area with large potential in this regard will be the use of re3data.org database for identifying suitable repositories for the researcher.
The DDRS is intended to be as easy to use as possible. Usability and clarity are of utmost importance in the process of identifying suitable research data repositories. Considering this premise, the concept of the DDRS changed in the discussion from a questionnaire service covering all relevant areas to a two-tier service as seen in Figure 2.
The first tier aims at identifying suitable repositories for the user with requesting answers to no more than a handful of questions. The user receives a ranked list of repository recommendations. The ranking is based on an internal but simple mechanism. For instance, a repository that is able to offer certain services or functionalities will be ranked higher than a simple repository only able to store file-based research data. These areas do not have to be answered by the user because the ranking can be undertaken quite simply on the service side and because most users, although they probably care about areas as licensing, metadata schemas or long-term preservation, may not be able to verbalise these issues in information science terminology.
In the second tier the user may - if they wish so - describe their specific case, i.e. the research data to be deposited. The research data concerned is described by the user along a few standardised categories, such as format, data volume, licences and so on. The aim of this description is to allow the repository an overview of the specific ingest case and to prepare for the communication with the researcher. This information, along with personal contact information, flows into a form that can be forwarded to the preferred repository at the instigation of the user. The second tier is optional, in other words, the user should have useful information about a suitable repository for their Data Management Plan after the first tier.
As long as a widely used and established infrastructure for the deposition of research data (as for publications) is not available, a conventional service like a repository registry can be useful in boosting the growth of archived research data. It contributes to lowering the inhibition threshold of the researcher to deposit their data on the one hand and it may be useful to standardise information on the data repositories as an incentive for interoperable services on the other hand.

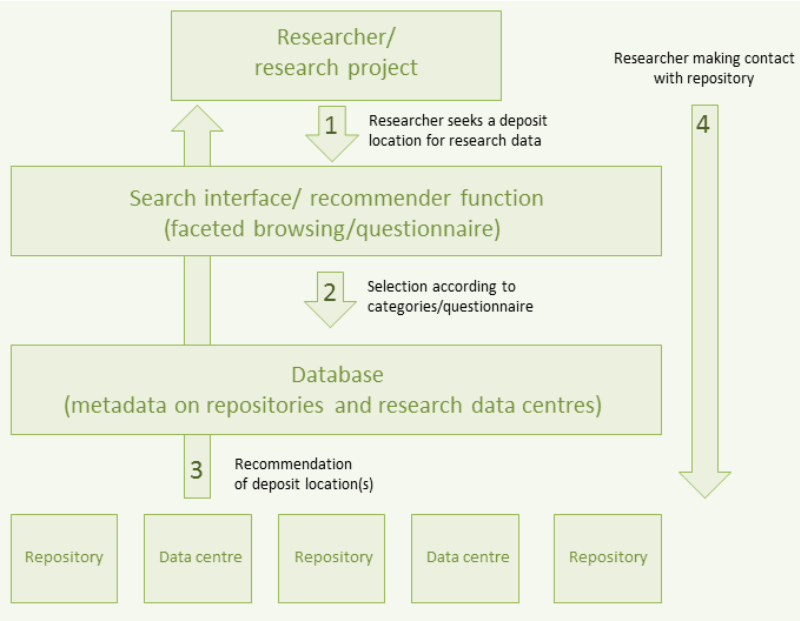

Figure 3: Workflow of the registry/recommender function

In the beginning, the service will be a type of registry of suitable deposit services. It catalogues repositories and research data centres and delivers standardised metadata on them, allowing the researcher to choose which repository may be the most qualified one for his or her case. The collection of the data will initially be done manually but future automated procedures can be developed to harvest data from repositories. It is important to keep in mind that for file-based data repositories will likely be the preferred infrastructure type but for more complex forms of research data – as mentioned above – other infrastructures have to be considered that go beyond conventional repositories.

The registry for deposit infrastructures takes into account the heterogeneity of data and the often compartmentalised research structures in the humanities. Obviously, it also takes into account already established services such as re3data.org but functions – as a main distinction – on a discipline-specific level. The service will connect researchers who search for a deposit service on the one hand and the repositories and data centres on the other hand which have a mandate to acquire content and it aims at establishing incentives for both sides to deposit and ingest research data in the humanities.

The repository registry and recommender function will initially be implemented on a simple technical level. With progress in coverage and usage, the service will become more sophisticated. The current workflow of the user through the two tiers - repository recommendation and data description/ contact with repository - is illustrated in Figure 3 above. The service will guide researchers and research institutions to the most qualified services for their individual depositing concern. What kind of research data can be deposited in which repositories or data centres, what requirements do they have to meet, how does the ingest process work, who can be consulted, what are the costs and

necessary service level agreements[36]? These kinds of questions will initially be answered only rudimentary as the researcher browses through a catalogue of research data repositories and data centres. A good visualisation of such a faceted browsing could look like re3data.org[37] has implemented the browsing by subject. The underlying technical solution uses a database with repositories tagged with classifications of the covered content. The user browses through this metadata and gets results very quickly, leading him to the suitable repositories.
A similar solution is at hand for HaS. Depending on the detail of the metadata on the repositories that is available various use cases become possible. With a more detailed description of the metadata a more branched inquiry of the researcher's individual requirements becomes feasible allowing more individually tailored results and thereby improving the user experience. To achieve such a maturity and usefulness in service quality requires one the hand a stable technical solution, on the other hand - and possibly more important - a rich and reliable database.

5.2 Use cases

The following section describe the most relevant use cases for the DDRS. By describing the use cases in a structured way, we try to cover all necessary aspects that should be considered with the design of the service.
The DDRS offers benefit for at least the following four basic researcher driven use scenarios and six management driven use cases (summarised in Figure 4):

- (A) Identify deposit repositories: the user - a scholar or a researcher - wants to archive a set of research data and has to identify a suitable repository which should fulfil certain requirements. These requirements can be deducted from the research funder's policy or be set by the user himself and will be fixed through a questionnaire process. The questionnaire should be as short as possible, requiring maybe not more than five questions. The DDRS should be not only able to suggest the best suited repository, or a list of ranked repositories, but also be able to initiate the contact between the user and repository/-ies. One desirable feature of the DDRS would be to build up a growing memory of "requests/decisions" to improve or accelerate the identification process.
- (B) Collect specific information for a DMP: the user - a scholar or researcher - has to collect information for a project specific data management plan. The necessary information comprises - amongst other things - information on the deposit repository and some of its specifications such as access policy or discipline coverage. The process for collecting this kind of information could basically be the same as the above described one for the identification of research data repositories.

[36] The SLAs are subject to Humanities at Scale WP 6.3

[37] Browse by subject facet: http://service.re3data.org/browse/by-subject/

- (C) Collect general information on research data repositories: the user wants to inform him- or herself on the research data repository landscape. This information interest can be focused on disciplines, access policies or can be country- or language-specific. The DDRS should offer for this use case a transparent, complete and detailed browsing option to perform different searches in a row. This could be implemented similarly to the re3data-interface but with lesser categories.
- (D) Register a research data repository: the user - a repository operator - wants to register his or her service for the DDRS. This should be conveniently possible directly via the DDRS or - if we pursue the intended plan - via re3data. This use case is aimed at extending the visibility of research data repositories and/or enhancing the database quality and quantity of re3data. The DDRS could be a leverage for repositories to improve their dissemination and interoperability.

Furthermore, the DDRS system has six management use cases (summarised in Figure 4):

- (F) Language localization of interface: the service has to be designed in a way that future localizations can be incorporated as easy as possible. This demand is important for the usability of the services.
- (G) Addition of information about a repository that is not available in external sources: as in the current design state the service relies heavily on the re3data-database to identify suitable repositories for the user. As this database does not focus on the arts and humanities we have a likely risk of non-inclusion of repositories that may be relevant for the user. The gap of these "missing repositories" can be addressed at least in two ways: indexing them in the re3data-database or adding the information on the side of the DDRS. Although the latter way seems more challenging it opens the way for including other information than those included in the re3data-database. As a reminder, the re3data-database relies upon a selected set of properties summarized in the re3data-metadata schema v.2.2[38]. This schema covers all research domains and is not arts and humanities-specific. A new version of the schema is being implemented within the re3data API, version 3.0[39].
- (H) Monitoring of successful deposits: this aspect relates to the above described usage statistics. The data on successful deposits would be a main quality indicator for the DDRS. So far, the design approach does not offer an easy implementation for the monitoring of successful deposits. If a deposit is finished successfully the user will not return this result to the DDRS. Possibly this aspect can be covered during the forwarding of the ingest request to the repository. Simply spoken: the form includes our request to receive an update on a successful ingest, as some kind of brokerage fee.
- (I) Usage statistics reporting: the DDRS has to include some kind of usage statistics reporting. This is not only important to improve the quality internally but

[38] http://www.re3data.org/schema/2-2

[39] http://www.re3data.org/schema

it becomes crucial with regard to two aspects: firstly, it becomes possible to use the usage statistics as an enrichment for the identification process, i.e. to rank services along their popularity; secondly the usage statistics can be used to raise the attractiveness of the service towards repositories that so have not been included both in our DDRS or in re3data.

- (J) Changes to questions and question structure: the design of the service has to reflect a flexibility to change the set of questions in the future. This can become necessary as soon as the used database changes, e.g. gets more granular in certain areas, or as the users' perceptions of research data changes, e.g. new issues become important for them or other issues are becoming less important. This flexibility is necessary both for the questions used to identify repositories for the user but also for the data description process. Likely the latter one is easier to adapt than the questionnaire process.
- (K) Language localisation of questions: the service has to be designed in a way that future localisations can be incorporated as easy as possible. This demand is important for the usability of the services.

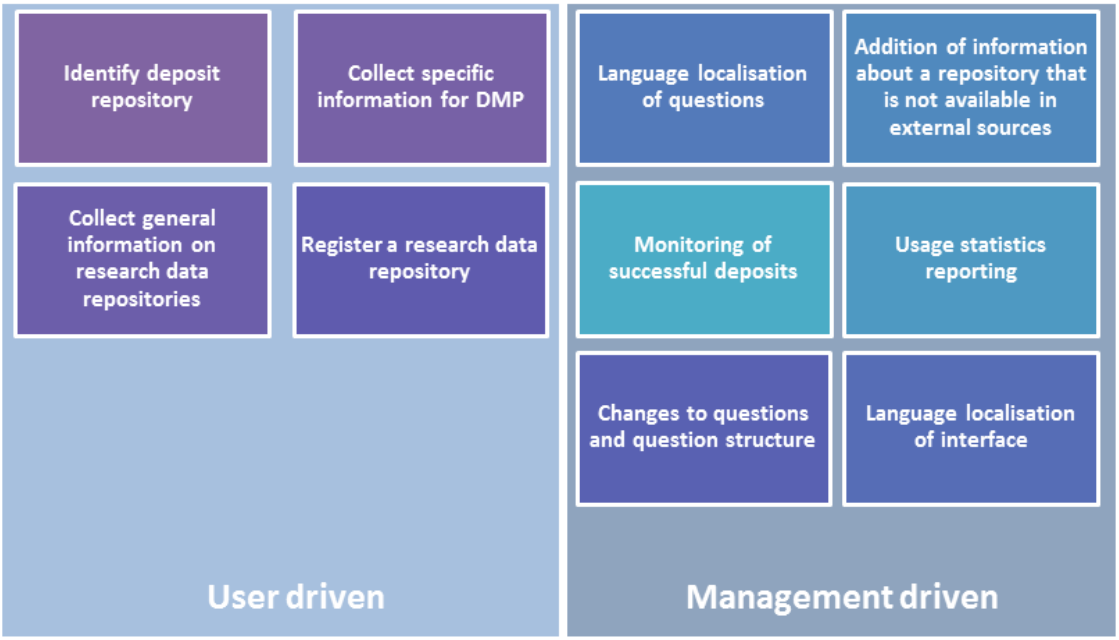

Figure 4: Use cases of the DDRS from the users' and the management's perspective

5.3 User stories

During the design process, we drafted several user stories or simulations of the user's workflow taking into account the re3data-database and its metadata schema. An example of these user stories is shown below in Table 4 (see also Figure 5).

User story 1 Scenario 4: Scholar looking for a relevant repository to deposit research data: Looking ByAnyCriteriaAlternative
Number + name UC1.SC4.Looking ByAnyCriteriaAlternative
Goal Provide a relevant list of repositories to the scholar
User input Metadata of one corpus
Persona details French scholar, in Communication Science, working on annotated corpus of tweets, good knowledge in DH.
Preconditions The user does not need a login and password, (s)he accesses the DDRS web site via the DARIAH.EU website.
Basic flow of events / scenarios 1. The user accesses the page1 of the DDRS website. This page explains briefly the next coming steps and display two choices (buttons) : **"Looking for a repository"** or **"Prepare a DMP"** 2. The user click on **"Looking for a repository"** 3. DDRS suggest **'Looking for by subjects'**, **'Looking for by keywords' and 'Looking for with specific requirements"** 4. The user click on **"Looking for by keywords'"** 5. DDRS displays an open window where the user type his/her keywords 6. DDRS checks the answer. There is at least one keyword. If not, retry step 5 7. DDRS does not find relevant repositories. 8. DDRS displays a message 'No repository found' and suggests to display a sorted list of keywords extracted from re3data.org (to clarified how to implement this on a technical level, e.g. via re3data's elastic search) 9. The user chooses one or several keywords in the list 10. DDRS finds relevant repositories and displays them. (there are repositories because the user chooses keywords from existing ones)
Postconditions Suggestions Displayed
Remarks This scenario is identical to the 'search' menu of re3data.org. Any criteria can be used (subject, country, keyword) in the query

Table 4: Initial user story based on the metadata schema of re3data and describing a repository identification and the according flowchart below

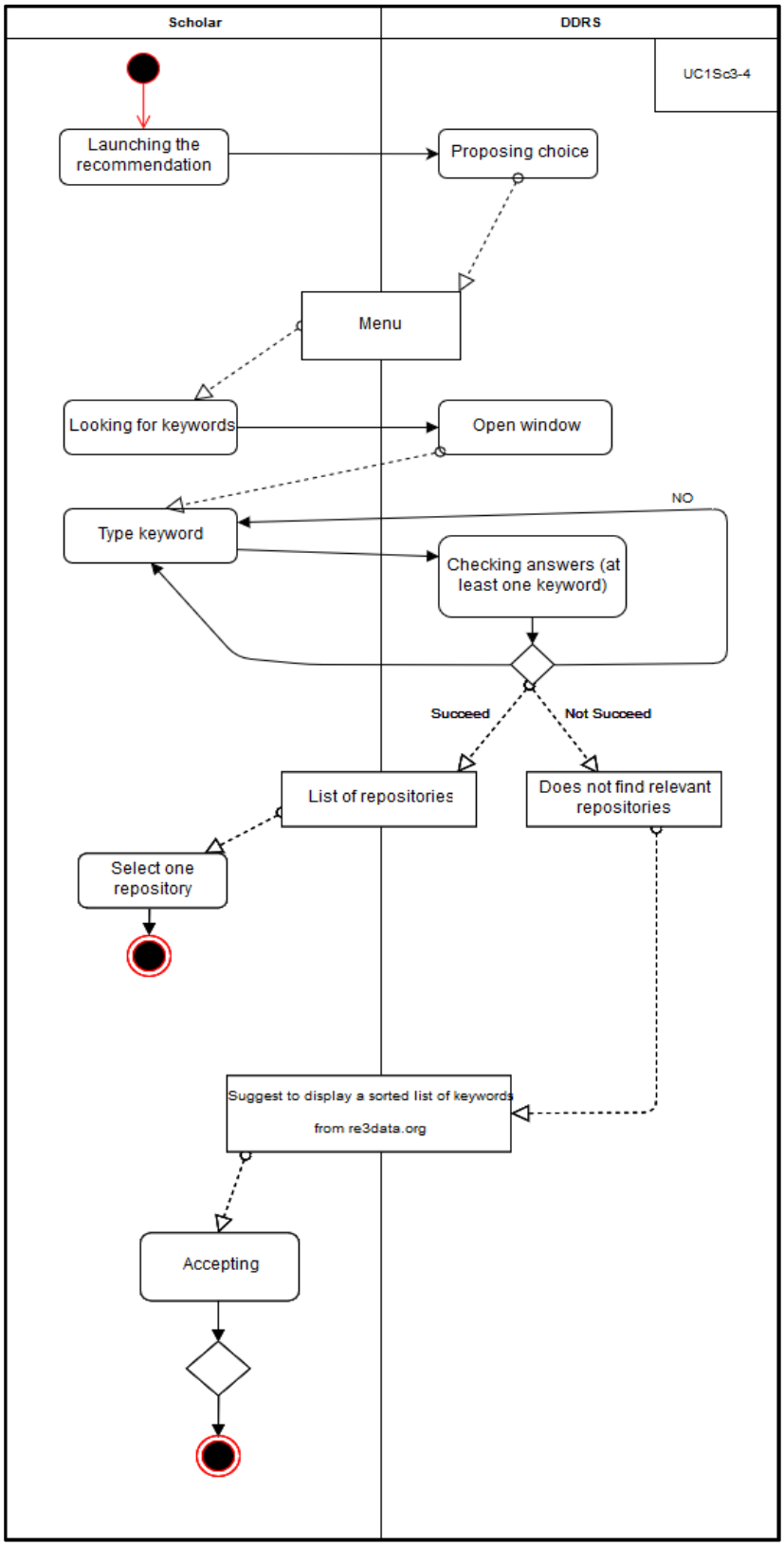

Figure 5: schematic representation of DDRS workflow

Acknowledging the limitations of this approach the re3data-metadata schema has been transformed in the next step into a simple include-exclude table to inquire how granular a search function based on their schema could be.

Table 5 lists the categories that may be subject either to the repository identification process (questions to be answered by the user) or to the data description process (filling the form to be forwarded to the repository). The categories are based on the current re3data-metadata schema version 2.2[40].

No.	*CATEGORY*	*INCLUDE*	*EXCLUDE*	*Include in data description process*	*Comment (e.g. if a category can be used to rank the results)*
1.	Subjects	X		X	
2.	Content types		X	X	
3.	Countries	X		X	possibly to be extended by affiliation
4.	AID systems		X		ranking option
5.	API		X		ranking option
6.	Certificates		X		ranking option
7.	Data access		X	X	ranking option
8.	Data access restrictions		X	X	
9.	Database access		X	X	ranking option
10.	Database access restrictions		X		possibly include in data description process
11.	Database licenses		X	X	ranking option
12.	Data licences		X	X	ranking option
13.	Data upload		X		
14.	Data upload restrictions		X		
15.	Enhanced publication		X		ranking option
16.	Institution responsibility type		X		
17.	Institution type		X		

[40] The Include-Exclude-Table will likely be the same for most scenarios. The re3data-metadata schema 2.2 is available under the following URL: http://www.re3data.org/schema/2-2

18.	Keywords	X		X	If free keywords are allowed, this option has to be explained
19.	Metadata standards		X	X	ranking option
20.	PID systems		X		ranking option
21.	Provider types		X		
22.	Quality management		X		ranking option but this category is possibly not comparable defined
23.	Repository language	X			
24.	Software		X		
25.	Syndications		X		
26.	Repository types		X		
27.	Versioning		X	X	ranking option

Table 5: Include-Exclude table for the DDRS based on the re3data metadata schema

Based on the analysis in Table 5 a survey has been sketched which demonstrates the flow of questions and tasks for the user. The survey has been created with LimeSurvey and is illustrated below in Figures 6A-D.

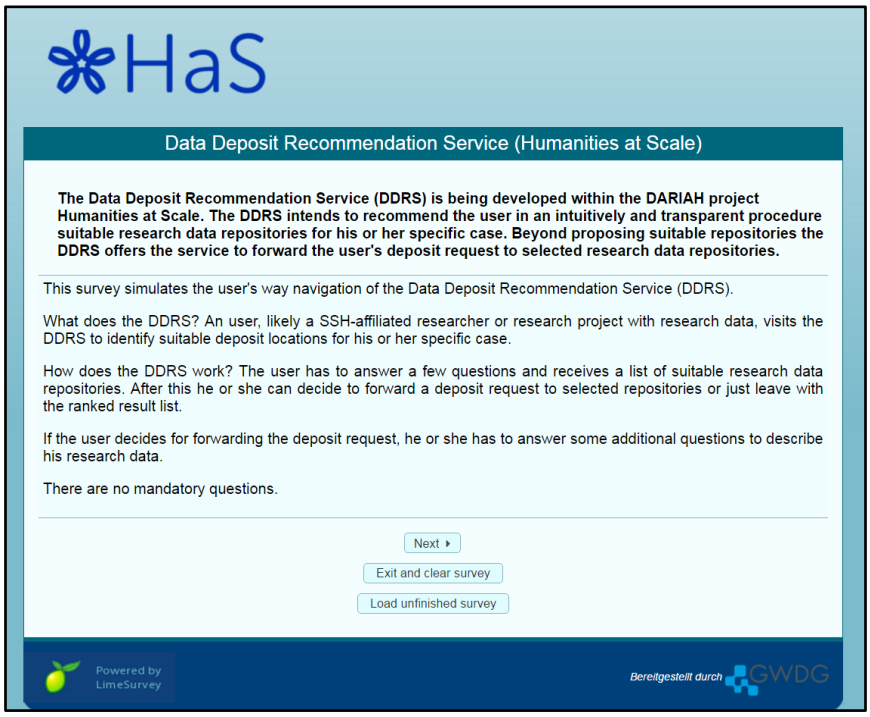

Figure 6A: Starting page of the survey demonstrating the user flow through the DDRS

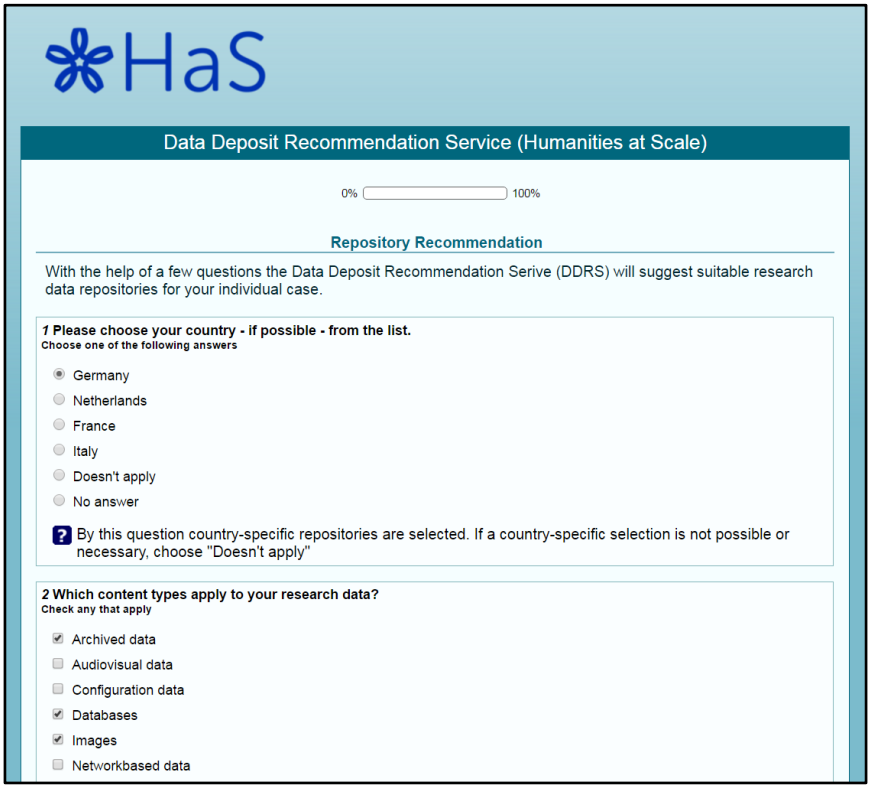

Figure 6B: By answering up to five optional questions the DDRS recommends suitable repositories

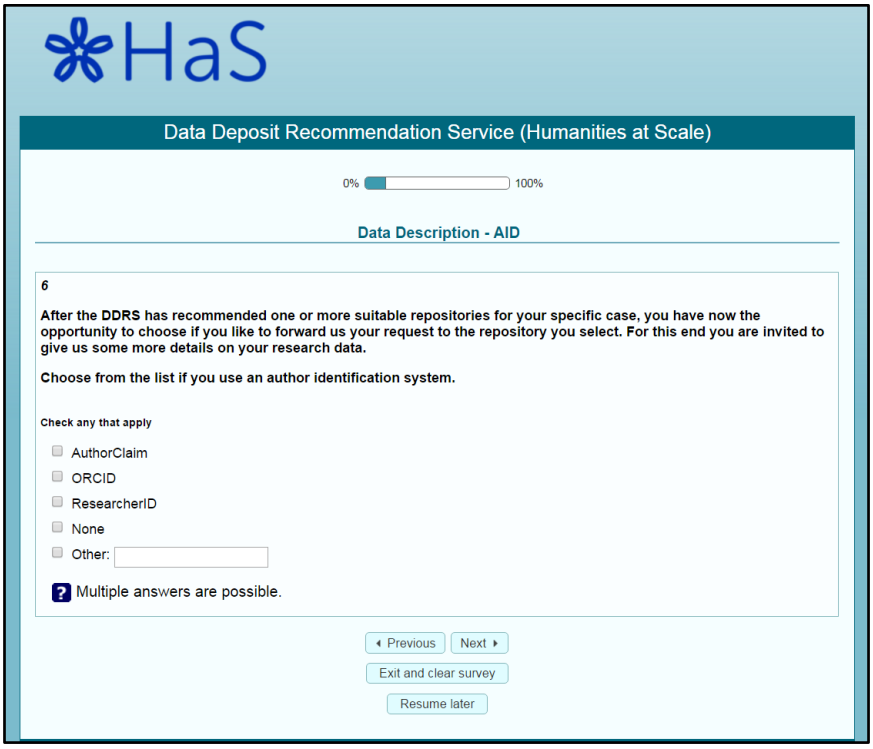

Figure 6C: After the suggestion of the repositories - any by this finishing the first tier of the DDRS - the user can decide to proceed and describe his or her ingest case

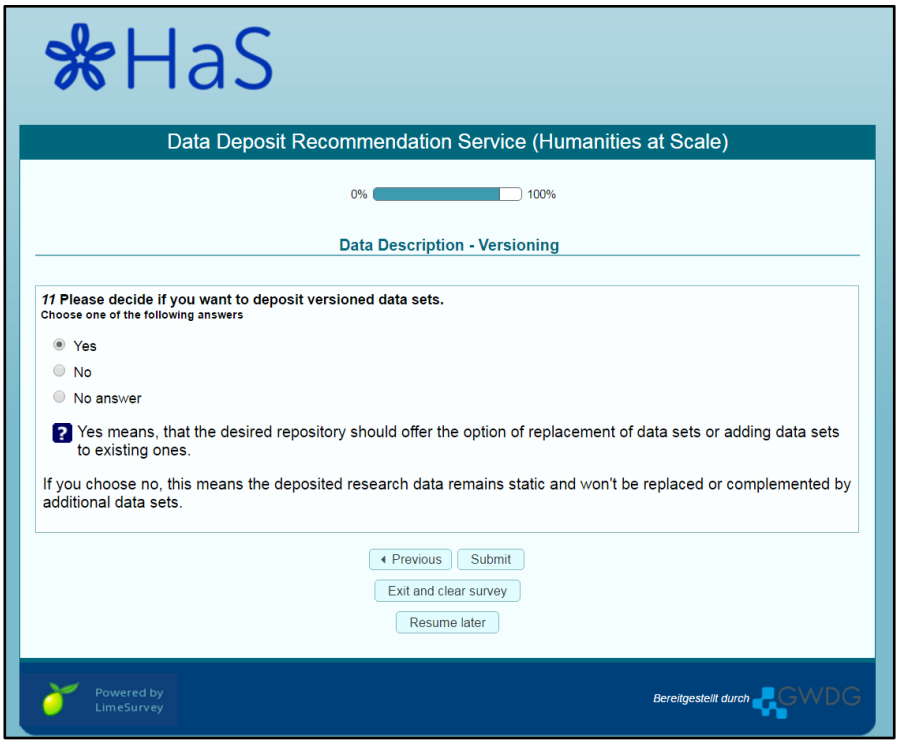

Figure 6D: In the second tier of the DDRS the user may describe his or her specific ingest case in a more detailed way. By this a structured description of the ingest case is created which can in the next step be forwarded to the selected repository, respectively the contact person

6 Technical implementation of the DDRS

This chapter has been modified following the project review and the natural development of the tool. The information provided for the technical implementation, especially the paragraph 6.3 of this report, was outdated. The version below reflects what has been achieved in D7.3.

6.1 Overall approach

This section describes the technical implementation of the DDRS within the Humanities at Scale project. It is important to distinguish between an ideal concept of the service and the actual implementation in the project. The latter one has to consider the available resources, the time horizon and the institutional context.
As a reminder: the DDRS assists the user to identify suitable research data repositories for the individual case depending on only a few criteria, like formats of the research data, language or affiliation or certain indispensable functions[41]. The result of this step will

[41] These additional criteria don't have to be indicated by the user but are shown in the detailed metadata result for the repositories. This aspect of the DDRS changed during the design phase. Initially a more comprehensive set of questions was planned to deliver

likely be a ranked list of repositories which can be used by the user as it is. The questions leading to the result list are not mandatory but the result gains quality by answering more questions. After displaying the result list the user can decide to enter the second functionality layer of the DDRS, which is about the structured description of the individual research data. Aim of this step is to gain, as easy and convenient as possible, a structured and coherent data description which serves as basis for initiating the ingest process with the repository. At this stage, the DDRS serves only as communication handler on behalf of the user, pointing his or her ingest request to the appropriate contact person.

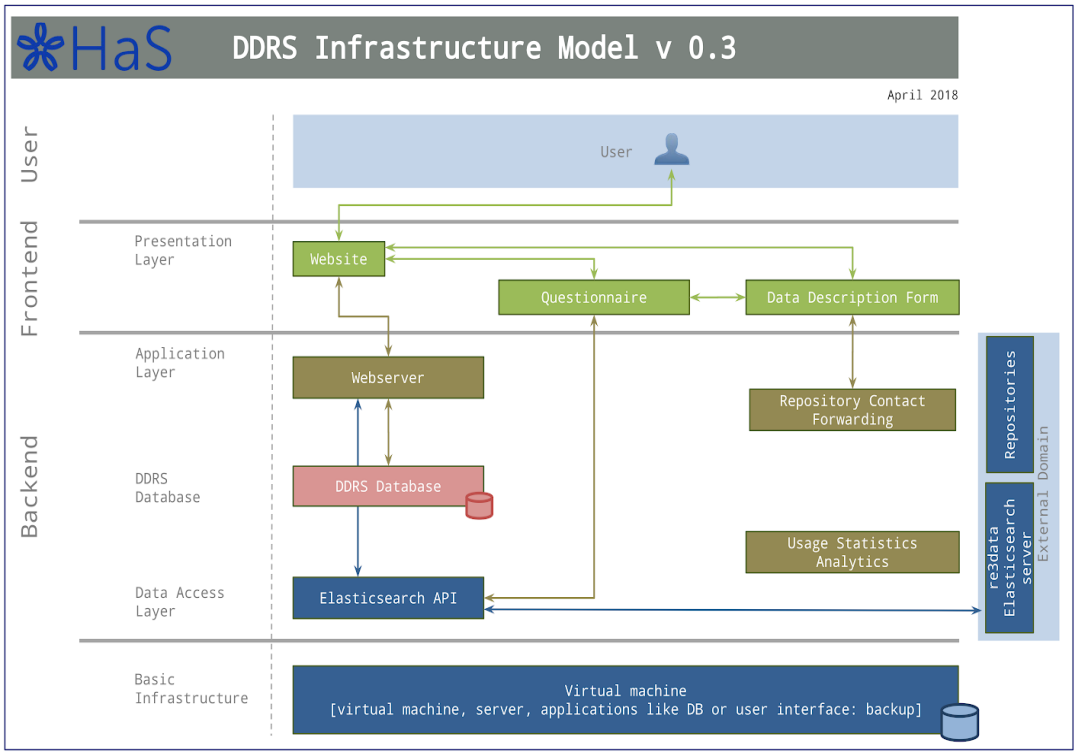

Figure 7: The DDRS infrastructure model version 0.3

Figure 7 provides an overview of the infrastructure being set up within the project. The result is a functional demonstrator, flexible to be developed further on or to be enhanced with additional functionalities. This result serves as proof-of-concept for the idea and will highlight the community's demand for such a service.

As basic infrastructure for this stage of the DDRS a virtual machine (VM), accessible via the internet are sufficient. The VM consists of all necessary applications and will initially be accessible over an IP.

It was decided that the branding of the service would be quite close to the DARIAH's one, obviously including the logo of the project in which the DDRS was created: Humanities at Scale and the logo of the underlying service which provides the data: re3data. The URL

results with more accuracy. The current practice however showed that this idea finds challenges in terms of usability and in the number of humanities-specific research data repositories. It may be the case, that this aspect will change with a more common use of research data repositories.

was also branded as DARIAH, using the following one: https://ddrs-dev.dariah.eu/ also keeping in mind that the service is in a demonstrator's phase.
The DDRS infrastructure model above illustrates the basic infrastructure layer and several components facilitating the use of the DDRS functionalities for the user. The following components are part of this infrastructure:

- A web server hosts the components described below.
- A simple website provides the user with explanatory information on the service, practices for research data in the humanities, further information sources, and displaying the results of the user requests for layer 1 (repository identification via a search) and layer 2 (data description).
- A simple questionnaire suggests the user a ranked list of suitable research data repositories for the specific use case. The questionnaire is designed in such a way that adjustments of the questions are possible in an easy way via the administration section. This is necessary as the used database for the requests - initially re3data - will likely change over time. For example new research funder mandates could be reflected in the metadata and the DDRS had to consider this.
- A web form describes the individual research data in a structured way (can be implemented in a similar way as the questionnaire). The questionnaire is also designed in a flexible way to allow further adjustments on the research data criteria that are to be described by the user. This will likely be the case as the research data practices in the humanities develop and new standards emerge.
- Currently[42] the DDRS sends queries directly from the server to the Elasticsearch of re3data. A request API conducts the requests to identify the repositories. The API sends - either filter by filter or all in one - (a) request(s) to the re3data database, displaying in the end a list of repositories fulfilling the respective criteria. On the basis of early tests of the re3data API the data quality and performance seem to be sufficient for our purpose and do not seem to trespass on the re3data API's general performance.
- A database is used to enrich the request results from re3data with contact details. This enrichment is necessary as the DDRS not only wants to suggest suitable repositories but also points the user to a competent point of contact to facilitate the ingest of the individual research data. Therefore someone with expertise in humanities research data is necessary but this information is not available through the re3data database as this is a non-disciplinary service.
- A forwarding component, basically a mail server. This components mails the completed data description form to the respective repositories.
- A usage statistics component, currently Matomo. At this point is not clear what kind of data could be collected by this service in the future. If the DDRS has a

[42] In an early phase of the DDRS development, the request API conducted the requests to identify the repositories. The API sent - either filter by filter or all in one - (a) request(s) to the re3data database, displaying in the end a list of repositories fulfilling the respective criteria. On the basis of early tests of the re3data API the data quality and performance seem to be sufficient for our purpose and do not seem to trespass on the re3data API's general performance. This had been changed to accessing directly the Elasticsearch in the course of development.

considerable user uptake in the future the usage statistics could become a valuable asset to be used for further added value services.

6.2 Information retrieval

Regarding the quality of the search results one has to consider first of all the limitations of the current approach which relies heavily on re3data's database.
Initially the design of the DDRS[43] relied on an include-exclude table which meant that the DDRS could select the search results only by applying the filters which are given by the re3data metadata schema v2.2 and its 39 main properties and according subproperties[44].
The DDRS now includes an additional database containing information on the points of contacts for forwarding the ingest request. The re3data schema contains only information on technical points of contact for the repositories but not for research data managers or information specialists. This additional database is relying on re3data's external persistent identifiers in order to keep the information always bound to the same repository - for information, a contact information can only be bound to a single repository within re3data.
The DDRS supplementary database also includes a selected set of research data repositories of generic, national or European orientation. This ensures that a user in every case will receive a result list, in case the filtering of re3data would result in zero results. Although this approach makes sense from re3data's perspective, it is not helpful with look at the DDRS' use case. Our aim is to equip each user with a selection of suitable research data repositories. To avoid a zero result upon filtering the DDRS database had been supplemented with a set of generic research data repositories suitable for humanities data and referring to the national or European level.
But, considering these limitations the decisions still came to using the re3data database. To our understanding re3data has the potential to grow in data quantity and usage and is for this end in each scenario a better choice than setting up an own exclusive database for the DDRS. Our assessment of the future development of re3data also implicates a further enhancement of their schema. With more and more established practices and growing use of research data management infrastructures in the humanities, additional properties reflecting this growth will enrich the schema and database. The current concept of the DDRS permits the integration of other databases, but not easily as we would need access to their Elasticsearch servers or any kind of APIs they are providing.
The following remarks describe in a more technical way the information retrieval of the DDRS from re3data starting with a result list after filtering for two countries affiliations (Germany, France).

[43] The complete documentation and according code of the DDRS are available at GitHub: https://dariah-eric.github.io/ddrs/

[44] See chapter 5.3.

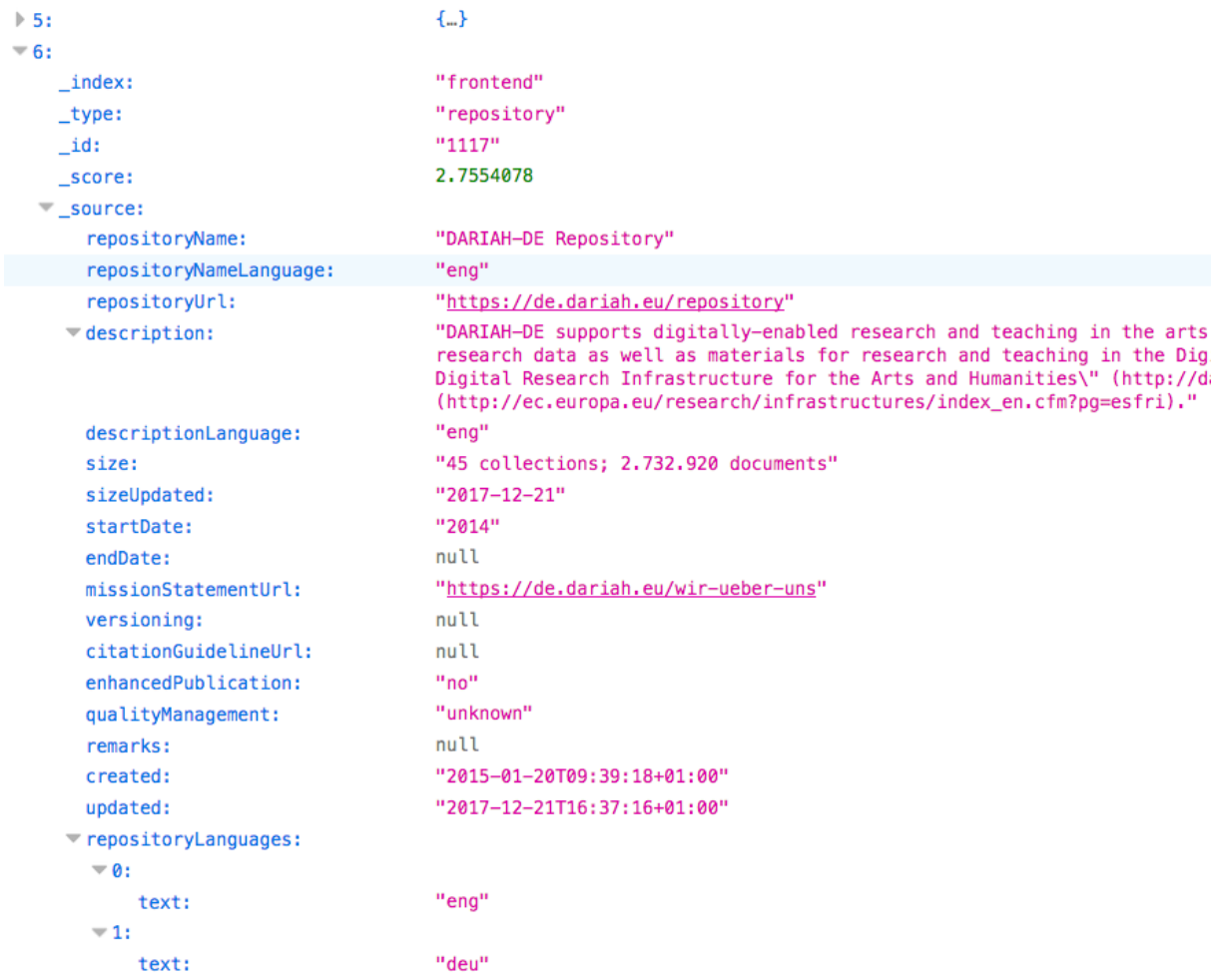

Figure 8: 58 repositories are listed as result of a query to the re3data Elasticsearch server (as of April 2018). The screenshot shows a snippet with only two repositories. The following URL leads to the complete result.

Figure 8 shows a snippet of the search result of re3data's Elasticsearch server for the following query (we can't provide the full URL as this is not a public API):

http://……/_search?q=institutions.country.raw:DEU AND subjects.text:11 Humanities

The search requests re3data to deliver all repositories with German affiliation and included in the DFG subject "11 Humanities". The aforementioned integration of additional sources like the DDRS supplementary database (or even completely different sources) poses rather a challenge in terms of information science than of technology. Different data sources merging into one result for the user requires a mapping on side of the DDRS to ensure that additional properties are associated with the concerned repository. The merging of this information is done thanks to the use of the re3data's external persistent identifiers, the ones used in their public API, such as "r3d100010677".

6.3 Presentation of search results to the user

Technically there are three concepts at hand for the information retrieval:

- Simultaneous retrievals: for each filter 2 requests are sent to the re3data Elasticsearch server (1 request to get a query's result and 1 request to retrieve the information of the saved generic repositories) and the result is displayed immediately to the user. The questionnaire used for the repository identification is in this case used as a kind of live search. With each filter applied, the list of repositories coming into questions is reduced and the user can decide after each filter if he wants to browse the results or apply another filter.
- Consolidated retrieval: the user answers all questions necessary for the repository identification in a row and after this one a request to re3data is sent and the result is displayed to the user. The main difference of the consolidated against the simultaneous approach is, that the user doesn't see a "filter history". He or she receives the results and in some case this may only be one or no repository. In terms of usability the simultaneous approach may be the better choice.
- DDRS-ranked results: multiple API retrievals of re3data are stored in the session and then ranked for the user presentation in form of a list. This concept is able to combine aspects of the two other concepts but it is technically more elaborated and possibly not useful in all cases.

In practice a hybrid has been implemented. It is a combination of simultaneous retrieval and an enrichment by the DDRS database. As the number of questions had been condensed a consolidated retrieval is currently not necessary. This could change if the questionnaire in the beginning would be extended with more questions via the administration section.

A simple example illustrating the search principles using the public API - the user searches for repositories using ARK as PIDs:

http://www.re3data.org/api/beta/repositories?pidSystems%5B%5D=ARK

and ends up with 17[45] results. But the user also wants to include the ones using DOI as PIDs in the search as the research data only needs a PID, but not necessarily one or the other:

http://www.re3data.org/api/beta/repositories?pidSystems%5B%5D=DOI

[45] All search requests described in this chapter have been retrieved in March 2018 and may have changed in the meantime, particularly in terms of the number of results.

and ends up with 504 results. After applying the filter for both PID systems at once:

http://www.re3data.org/api/beta/repositories?pidSystems%5B%5D=DOI&pidSystems%5B%5D=ARK

only 10 results are remaining. However, this last result is confusing as one would like to have all the results using ARK and all results using DOI, but not only the repositories using both ARK and DOI. Therefore, using the public API we would be forced to launch multiple simple queries in order to retrieve meaningful repositories to users. That's why we started liaising with the re3data's team in order to find a solution for this issue. They kindly provided us with a full Elasticsearch server on their private network which allows us to make easier complex queries as seen below.

http://……/_search?q=pidSystems.text.raw:ARK OR pidSystems.text.raw:DOI

This provides 518 repositories (511 using DOI, 17 using ARK but including 10 using both) which are more useful to someone looking for a repository using PIDs in general.
This issue may also be more complex when other filters are applied, for instance specific technical functionalities or metadata requirements of the repositories. The third concept would add a ranking mechanism to the results. Simply spoken the user checks five filters and the results compliant to all five filters would appear on top, the results compliant to only one filter at the bottom of the list. Additionally the ranking concept could be enhanced by weighting of criteria, for example the availability of a specific author identification system, such as ORCID, is more important than the national affiliation of the repository. This weighted ranking is more sophisticated than the simple ranking and requires a more complex questionnaire approach than the concept currently allows. The current design of the DDRS neglects this option with look at the limited number of humanities-specific research data repositories. This may change in the future.

7. Recommendations for future development and sustainability

The issue of sustainability has been a key factor of consideration in this design study for an Open Humanities Data Platform. The chosen concept, which builds upon an existing and well-established service - re3data -, requires relatively little future maintenance compared to most of the different possible architectures (Chapter 4). During the implementation phase (WP7 T7.3) the platform will be developed in such a way that it allows for adaptability (e.g. change of questions, updated repository contact information,

additional languages, etc.) and service extensions for the changing needs of the community.

A likely example of a future update requirement will be the replacement of re3data's metadata schema. At the time of conceptualising the DDRS, re3data uses the 2.2 version of their schema[46], but they already present version 3.0 on their website. The DDRS therefore has to be able to process retrievals with the new schema in a seamless way as soon as it becomes active. This case may also be true if the DDRS wants to include other providers similar to re3data in the retrieval process.

An example of a likely future service extension may be on the recommending functionality. With growing usage of the DDRS it can become useful to aggregate the usage statistics and analyse them in a way to enrich the recommendation results. Additional service extensions could cover one or more aspects of the research data life cycle (Figure 1 in Chapter 1). Our chosen platform concept facilitates long-term preservation of data: the depositing of data for humanities researchers and the curation on the side of the archives. The DDRS can also be used to help select a suitable repository for use by a researcher when writing their project DMP. A logical extension of this service would be to include more resources for data management planning, for example a registry of DMP formats for different Humanities disciplines and funding agencies, and/or tools that help with data management planning.

Another aspect closely related to depositing data is the promotion and visibility of published data content. We see two ways in which this could be implemented here. Firstly, it should be possible for the depositor to simply post links to their newly deposited dataset on social media platforms, blogs, and project websites. With a distributed data deposit network, an integrated solution for publicising datasets for reuse would require the participating repositories and archives to also indicate, in a common machine-actionable format, when a dataset is publically available, either directly to the depositor or via a functional extension to the DDRS: for instance, this can be done by a simple RSS/ATOM feeds to be aggregated in the DDRS or by more sophisticated means with common REST APIs. To improve visibility and searchability, another future possibility would be to recommend a common description of 'DARIAH datasets', which means the use of common descriptors and vocabularies like the BackBone Thesaurus.

Secondly, DARIAH could consider setting up an Open Humanities data journal. In addition to increasing the visibility of published data, and providing quality assessment of data through peer-review, data journals create an extra incentive for researchers to publish their data because it counts towards their publishing output. Examples of data journals in the Humanities are the Journal of Open Humanities Data (JOHD)[47] and the Research Data Journal for the Humanities and Social Sciences[48]. The creation of a DARIAH data journal could be facilitated by the DARIAH Virtual Competence Centres (e.g. VCC3: Scholarly

[46] The used schema can be seen here: http://www.re3data.org/api/beta/repository/r3d100011839

[47] http://openhumanitiesdata.metajnl.com

[48] http://www.brill.com/products/online-resources/research-data-journal-humanities-and-social-sciences

Content Management), for example through the organization of a DARIAH Working Group to set up and maintain such a data journal.
Other facets of the data life cycle that could be covered by an Open Humanities Data Platform are, for example, finding data and processing or analysing data. This could be met by an extension of the platform with a data search function and by offering an overview or registry of tools and services, respectively (see Chapter 4.1) for an overview of service options). However, as discussed in this report, many of such functionalities are already covered by existing services. Moreover, the more complex the platform services and functionalities, the more resources will be necessary to guarantee the sustainability of the platform.
To what extent there will be resources available to maintain the platform in the future, and extend it with new functionalities, will depend upon the integration of HaS outputs by DARIAH-EU or partner institutions. The discussion of sustainability implies that a project leaves the status of third-party-funding and enters the status of an organisation with a legal status, clear decision-making structures and cost structures[49]. At this point, the continuation or follow-up of the HaS project or the WP7 Open Humanities Data Platform is uncertain as the project is not near this point. With regard to the ESFRI-phases[50] for scientific infrastructures: preparatory phase, construction phase, and operational phase, HaS currently is in the preparatory phase with respect to the development of DARIAH. In this regard, it is also relevant to consider the DESIR (DARIAH ERIC Sustainability Refined)[51] project. This Horizon2020-funded project, which runs from the beginning of 2017 until the end of 2019, develops means to enhance the usage and awareness of DARIAH and its services within the humanities research community and thereby contributes to the sustainability of the DARIAH digital research infrastructure. The DDRS could thus benefit from DESIR in terms of usage and sustainability through its links to and possible integration with the DARIAH infrastructure.
Since the DDRS is built utilising data and services from other platforms and service providers, it requires minimal maintenance as it does not need to provide a support helpdesk service (FAQs, support documentation may suffice). Update issues notwithstanding, this service could be localised and hosted at a number of institutions. We do not anticipate high bandwidth needs as this is just a simple (http) web service. At the current stage of the HaS project it seems that the sustainability of developed infrastructure components will be established through the DARIAH ERIC context, of course only under the assumption of a functional and demanded service. But this may not be the right scale for a smaller infrastructure component like the Open Humanities Data Platform. This DARIAH-coined approach does not exclude other forms of ensuring sustainability or even a non-DARIAH-branding of the platform. As the current DDRS concept is a lightweight web service that does not need a great deal of infrastructural

[49] Neuroth, Rapp (2016): Nachhaltigkeit von digitalen Forschungsinfrastrukturen. In: Bibliothek Forschung und Praxis 2016; 40(2).

[50] ESFRI Roadmap 2016: https://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri-roadmap

[51] http://www.ghentcdh.ugent.be/projects/desir-dariah-eric-sustainability-refined-0

resources to run, it could also be hosted and maintained by one or more institutions as an in-kind contribution to DARIAH.