HAL open science

Find Research Data Repositories for the Humanities - The Data Deposit Recommendation Service

Stefan Buddenbohm, Maaike de Jong, Jean-Luc Minel, Yoann Moranville

To cite this version:

Stefan Buddenbohm, Maaike de Jong, Jean-Luc Minel, Yoann Moranville. Find Research Data Repositories for the Humanities - The Data Deposit Recommendation Service. 2020. hal-03020703v2

HAL Id: hal-03020703

https://hal.science/hal-03020703v2

Preprint submitted on 14 Jan 2021 (v2), last revised 19 Aug 2021 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

Find Research Data Repositories for the Humanities - The Data Deposit Recommendation Service

Abstract

How can researchers identify suitable research data repositories for the deposit of their research data? Which repository matches best the technical and legal requirements of a specific research project? For this end and with a humanities perspective the Data Deposit Recommendation Service (DDRS) has been developed as a prototype. It not only serves as a functional service for selecting humanities research data repositories but it is particularly a technical demonstrator illustrating the potential of re-using an already existing infrastructure - in this case re3data - and the feasibility to set up this kind of service for other research disciplines. The documentation and the code of this project can be found in the DARIAH GitHub repository: https://dariah-eric.github.io/ddrs/.

Authors

Stefan Buddenbohm, Göttingen State and University Library, buddenbohm@sub.uni-goettingen.de
Maaike de Jong, Koninklijke Nederlandse Academie van Wetenschappen, Maaike.DeJong@bristol.ac.uk
Jean-Luc Minel, Universite Paris-Nanterre, jean-luc.minel@u-paris10.fr
Yoann Moranville [Corresponding Author], DARIAH, yoann.moranville@dariah.eu

Keywords

Research data, repositories, recommendation, re3data, DARIAH, digital research infrastructure for the arts and humanities

1 Introduction

The increasing production, dissemination and re-use of humanities research data leads to a growing demand for easy-to-use discovery services for the identification of deposit services and research data repositories. The establishment of research data centres and journals[1] are shaping this trend. Although there are already some meta services[2] making research data repositories visible and searchable, a humanities specific service is still missing. To address this gap, the Data Deposit Recommendation Service[3] (DDRS) has been conceptualised within Humanities at Scale (HaS), a DARIAH[4]-affiliated project.

The DDRS offers a simple humanities-specific faceted search for research data repositories applying the already existing, reliable and discipline-spanning database of re3data[5]. The DDRS is intended to enable humanities researchers to identify suitable repositories for the deposit of their research data. For the time being it is provided as an almost fully-functioning technical demonstrator[6]. The service can recommend research data repositories by discipline and consider a national, European or even global search scope. Technically the service is designed to be adaptable for other disciplines or functions, for instance as a registry for research data collections.

The DDRS as technical demonstrator (Buddenbohm, et al., 2017) serves not only the purpose of generating lists of selected research data repositories but also shows the potential and feasibility of re-using already existing services - in this case re3data.org - or of adapting the service to other research disciplines or even other use cases such as connecting it to the process of creating research data management plans. The intention to look for already existing resources and to build upon them – in this case re3data.org –, was a conscious decision to give an example set against the often experienced un-sustainability of temporary funded projects and infrastructural undertakings.

2 The DARIAH context

DARIAH (Digital Research Infrastructure for the Arts and Humanities) is a pan-European research infrastructure for arts and humanities scholars working with computational methods, being an European Research Infrastructure Consortium (ERIC) since 2014. DARIAH serves as framework for the DDRS. It supports digital research as well as the teaching of digital research methods. DARIAH connects several hundreds of scholars and dozens of research facilities in currently 19 European countries, the DARIAH member countries. In addition, DARIAH currently has 26 cooperating partner institutions in non-DARIAH member countries, and strong ties to many research projects across Europe

[1] A few examples: DARIAH-DE repository (https://search.de.dariah.eu/search/), CLARIN-INT (https://portal.clarin.inl.nl/), DANS-EASY (https://easy.dans.knaw.nl/ui/home), Research Data Journal for the Humanities and Social Sciences (https://brill.com/view/journals/rdj/rdj-overview.xml) or NAKALA (https://www.nakala.fr).

[2] Such as Re3data (https://www.re3data.org/) or an edited list of repositories at PLOS ONE (http://journals.plos.org/plosone/s/data-availability#loc-recommended-repositories).

[3] The DDRS has been developed within the Humanities at Scale project, a DARIAH-EU affiliated research project. The project received funding within the Horizon 2020 INFRADEV 3-2015 programme for the Individual Implementation and Operation of ESFRI projects of the European Commission. The Grant Agreement number is 675570.

[4] Website of DARIAH: http://dariah.eu/

[5] We like to thank the colleagues at re3data for their cordial and substantial support in constructing the Data Deposit Recommendation Service, in particularly Robert Ulrich, KIT.

[6] DDRS demonstrator instance: https://ddrs-dev.dariah.eu/ddrs/

and is also cooperating closely with other ERICs such as CLARIN[7] or CESSDA[8]. DARIAH[9] provides digital tools and shares data as well as know-how. It organises learning opportunities for digital research methods, like workshops and summer schools, and offers training materials for Digital Humanities. The DARIAH-affiliated Humanities at Scale (HaS)[10] project offered the opportunity to think about and experiment with technological solutions for humanities-related use cases. The project functions as catalyst activity for the already existing Digital Humanities resources, networks, research data, services and infrastructures at the European level and is partly followed up by the DARIAH ERIC Sustainability Refined (DESIR)[11] project, which is tasked with exploring technological and organisational sustainability scenarios for the DARIAH research infrastructure.

3 The DDRS user experience

The DDRS is geared towards researchers and research projects from the arts and humanities. The service addresses the question of how and where to deposit research data. This is a user need, which is gaining increasingly importance as reuse of research data becomes more common and more funders require researchers to publish their research data to stimulate the reproducibility of research.

The user experience of the service should be as simple as possible. After answering one or two short questions (see figure 1), the service recommends the best suited data deposit locations considering the user-provided parameters. The user can compare the details of the recommended repositories, which also include links to the different repository websites.

Designed for utmost usability, the DDRS requires only few interactions from the user. The first tier aims to identify suitable repositories for the user by requesting answers to a limited number of questions. The user receives a ranked list of repository recommendations. The ranking of the repositories is based on a simple mechanism, falling into two steps:

- Firstly, an internal DDRS list of default repositories' identifiers with a national or European scope is checked against the user's criteria and a query is sent to re3data's server to retrieve detailed information. These instant results appear at the bottom of the result list for the user and ensure to provide a useful result, allowing the continuation of the process (see figure 1).
- Secondly, this list of repositories gets enriched with results from the re3data ElasticSearch search engine which is queried with two user-driven criteria (country, disciplinary field) and with a set of DDRS-induced criteria (PIDs[12], Open Access, European countries and humanities related subjects). This set of criteria allows us to retrieve a certain subset of the repositories of re3data. For example, we would only like to have humanities repositories and only those providing PIDs,

[7] CLARIN – European Research Infrastructure for Language Resources and Technology: https://www.clarin.eu/

[8] CESSDA – Consortium of European Social Science Data Archives: https://www.cessda.eu/

[9] For more details on the DARIAH strategic plans for the next years consult the DARIAH Strategic Plan under: https://www.dariah.eu/2019/08/19/dariah-publishes-a-strategic-plan-for-2019-2026/

[10] Project information on Humanities at Scale is available under: http://has.dariah.eu/

[11] https://www.dariah.eu/activities/projects-and-affiliations/desir/

[12] Persistent Identifier

and so on. The two result lists - DDRS and re3data – are compared against each other and duplicates are deleted.

As result, the user receives a list of suitable research data repositories in the following order: national thematic repositories > national general repositories > European general repositories.

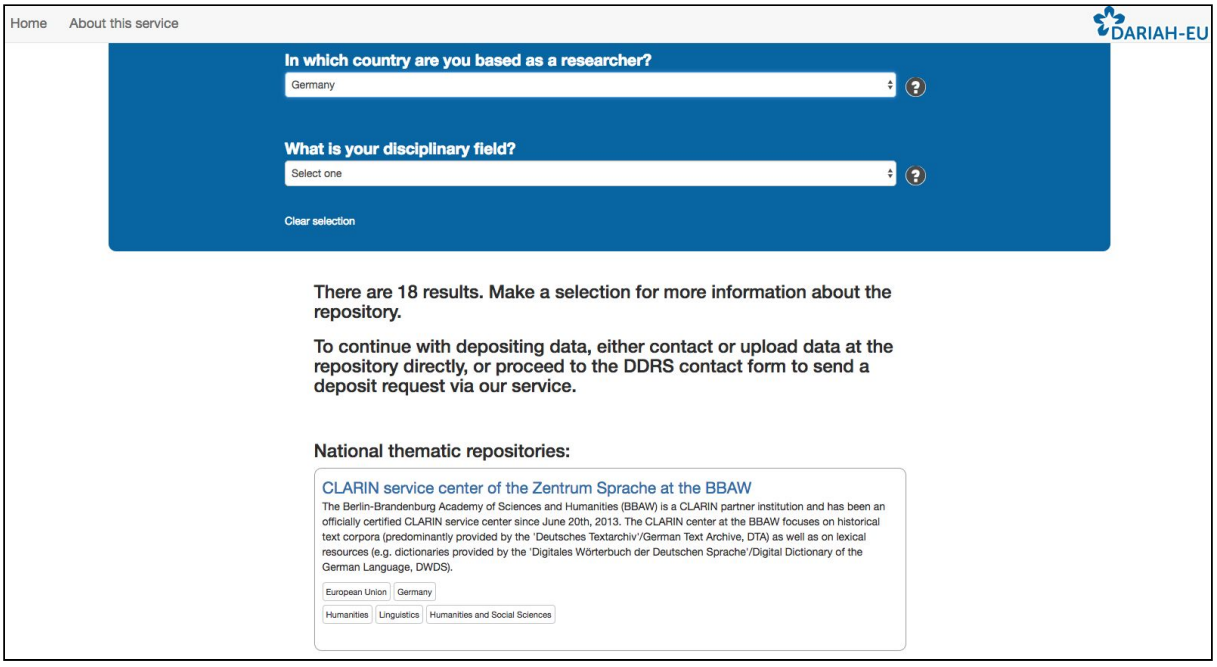

Figure 1: First instant result list after selecting a country affiliation

The described selection process may be enriched in the future with additional filters, for instance for licensing[13], metadata schemas[14] or data licences[15] or other fields that are offered by the re3data metadata schema (Re3data metadata schema 2.2.: http://doi.org/10.2312/re3.006).

At this stage, this additional filtering was not implemented for two reasons. Firstly, the effort required seemed to outweigh the benefits as currently there are a manageable number of repositories from which the user can select and quite easily assess for additional characteristics. The second reason refers to the user experience. The usability of the DDRS should be as easy as possible and for most users the currently available filters of re3data are not self-explanatory or not often needed. What would be of interest - for instance a research funder's recommendation for certain repositories - is unfortunately not available yet.

In the second tier the user may describe the specific case, i.e. the research data that shall be deposited. The research data concerned is described by the user along a few standardised categories, like 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

[13] Re3data Metadata schema 2.2, ID 22.1 dataAccessType (open, embargoed, restricted, closed)
[14] Re3data Metadata schema 2.2, ID 35.1 metadataStandardName
[15] Re3data Metadata schema 2.2, ID 23.1 dataLicenseName (for instance: CC0)

user should have useful information about a suitable repository for their data in order to fill in their Data Management Plan already after the first tier (see figure 2).
As long as a widely used and established infrastructure for the deposit of research data (as for publications) is not available, a service like a repository registry can be useful in boosting the growth of archived research data. It contributes to lowering the barrier of the researcher to deposit his 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.

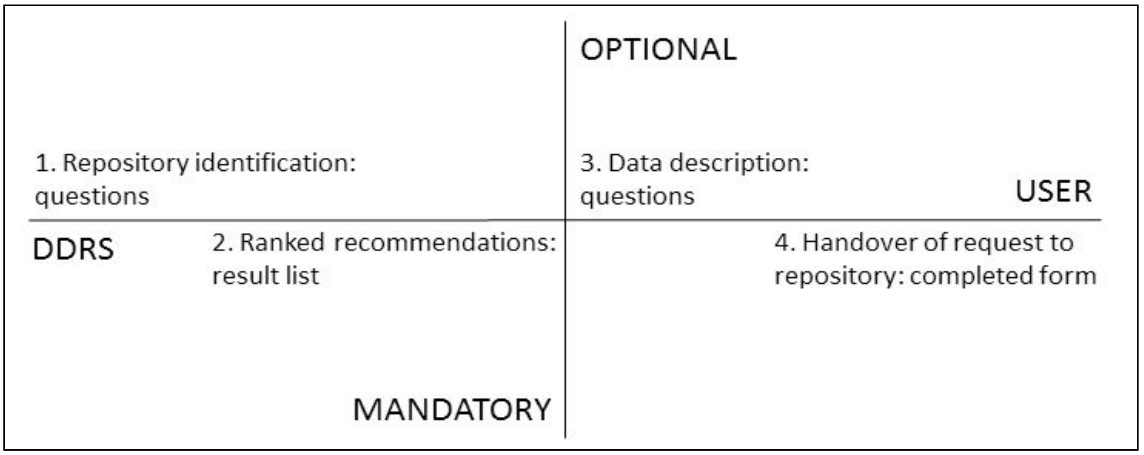

Figure 2: The DDRS as a two-tiered service

3.1 User walkthrough

Upon visiting the DDRS web page[16] the user can inform himself about the service or begin directly with the repository identification process. At least one selection has to be made, be it a geographical parameter or a selection of disciplinary field. These questions can be changed in the administration section of the DDRS. The disciplinary selection employs the DFG-Fachsystematik[17] (only its three-digit level), a widespread metadata schema also used by re3data.org which is used for further information retrieval.

[16] Website of the DDRS prototype: https://ddrs-dev.dariah.eu/ddrs/
[17] The subject areas of the German Research Foundation (DFG-Fachsystematik) as of April 2018: http://www.dfg.de/en/dfg_profile/statutory_bodies/review_boards/subject_areas/index.jsp

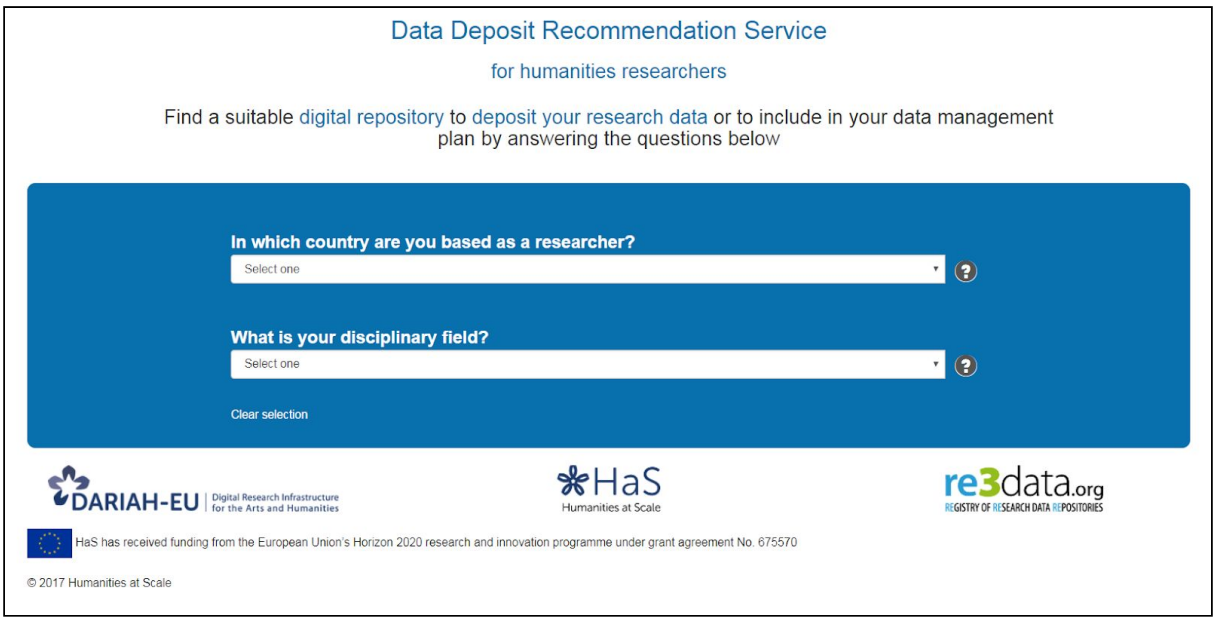

Figure 3: Landing page of the DDRS

After entering the preferences (or leaving one of the fields open, see figure 3), the DDRS displays a sorted and ranked list of research data repositories. The mechanics of the information retrieval and the ranking of the repository list are described in more detail below.

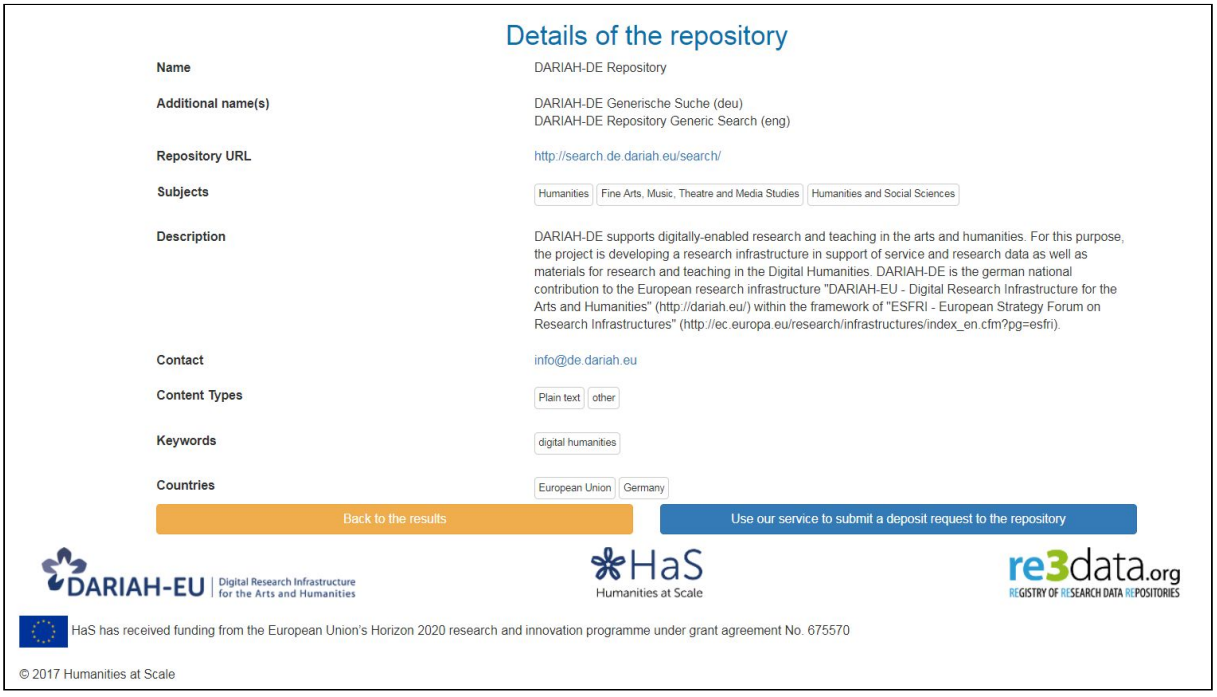

Figure 4: Detailed view of a research data repository record

Figure 4 shows the detailed view of an example repository. The displayed details are a condensed view based on the metadata schema of re3data. The most relevant properties for the DDRS users are displayed and many others - provided by re3data - are not displayed. It may be the case for many users that they are already happy with the

presented information and leave the DDRS at this stage. For all others, willing to submit a deposit request, the data description is described below.

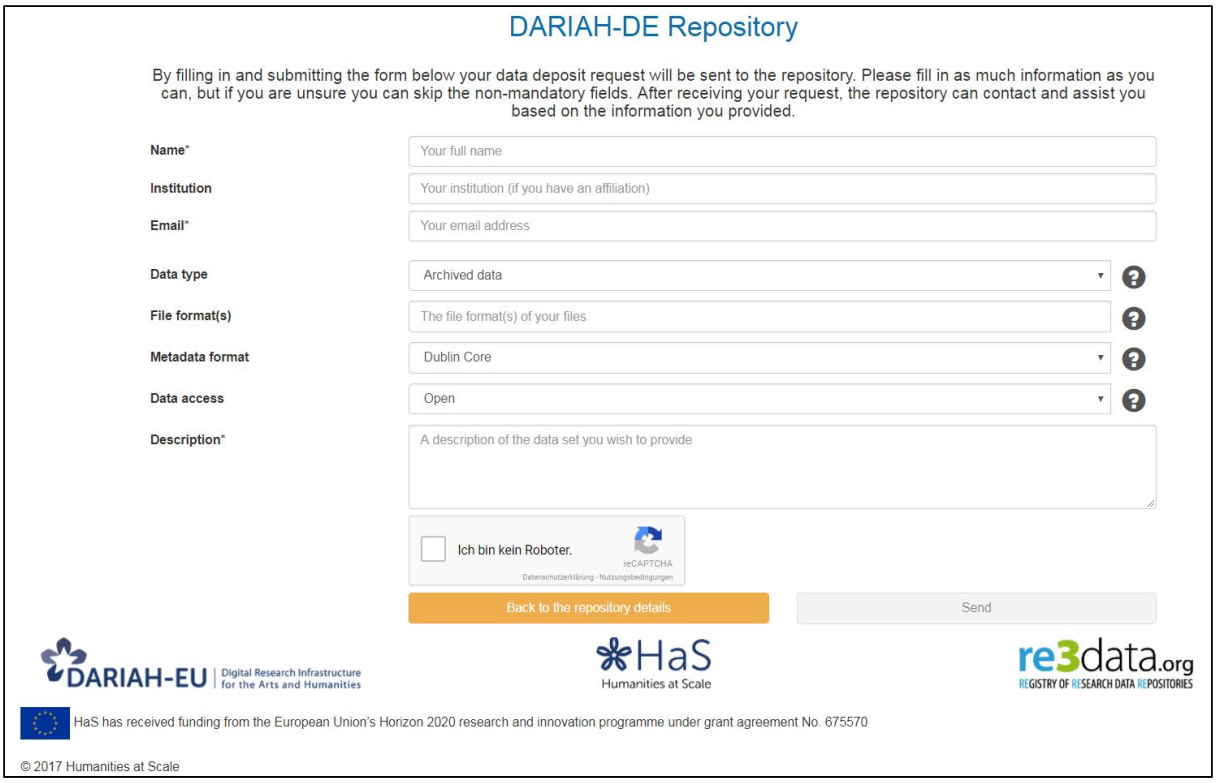

Figure 5: Data description form to submit a deposit request

Figure 5 shows the data description form that may be filled in by the user with case specific information. It is not intended to ask for a comprehensive assessment of the data, something not necessary from the DDRS' point of view and also not feasible for many users. Although they may know exactly what their research data is about and what it may be used for, they may find it challenging to complete the short description using the metadata fields provided and therefore see it as a barrier to use. This is the reason we ask to fill out only three mandatory fields, including a name and an email address for further contact. Furthermore the service stays GDPR-compliant using this approach.
Beyond this main objective of identifying suitable research data repositories, the DDRS also aims at raising awareness for research data repositories, increase transparency in their use and improve collaboration and interoperability between such services.

4 Stakeholders and users of the DDRS

Although the DDRS is aimed at humanities researchers to assist them in finding suitable deposit repositories, other stakeholders are also relevant and have been considered during the conceptualisation. The stakeholders include (compare figure 6):

- **Researchers (and associated research institutions)** are the core users of the DDRS: they are the main data producers as well as "consumers" (or re-users) of digital research data. As data sharers, they need to trust that their data is preserved, accessible, and usable 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.

- **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 holders of data 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**: other national or international infrastructures are relevant to DARIAH in terms of a possible cooperation, concerning e.g. the integration or reuse of components within DARIAH services, interoperability issues or extensions, such as CLARIN or CESSDA among others. One main aspect is to promote 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 stewards, data curation experts, or providers of training in digital methodologies, or higher education in digital humanities. Training and education - although not in the first instance integrated within the development - forms 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 demand that funded research data (and publications) is being published as open access. For example, the EU obliges researchers funded by Horizon 2020 to publish their research data as open data (European Commission 2016).
- **Policy makers**, i.e. national governments, and the EU, increasingly put 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, the availability of data connected to these publications is often less clear.
- **Humanities data consumers**: these can include e.g. educational 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.

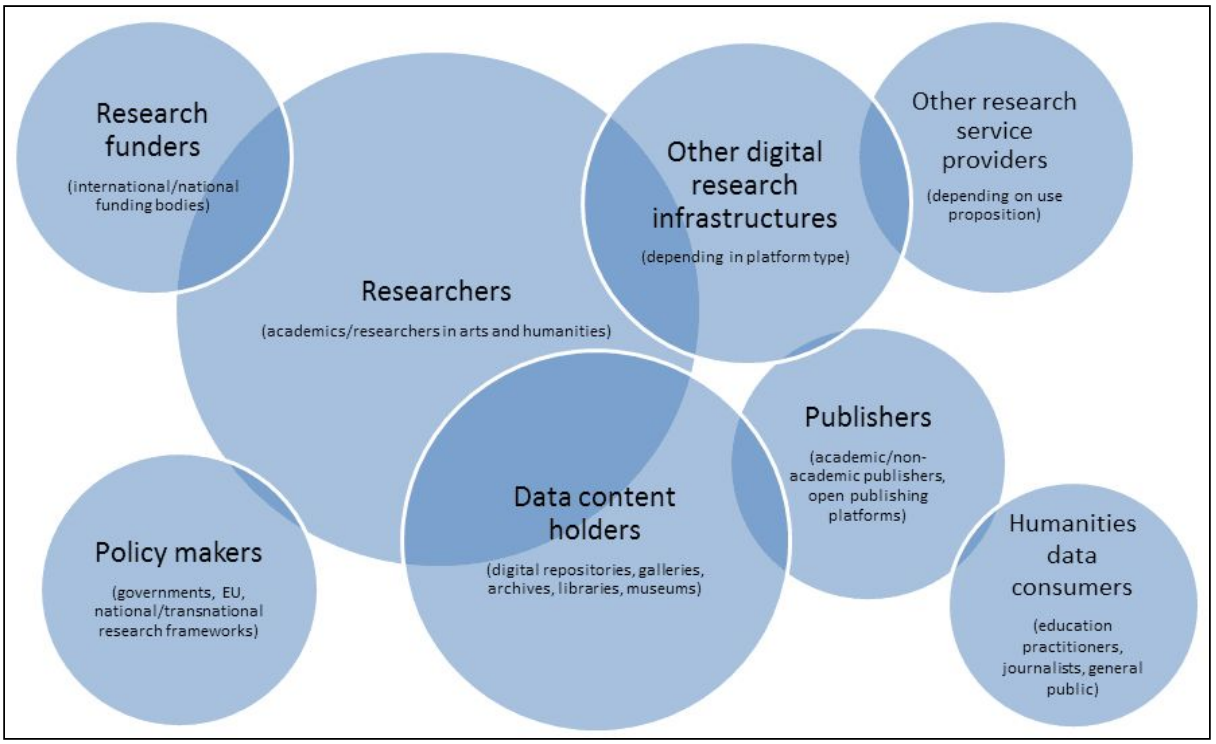

Figure 6. An overview of the main stakeholder groups and their relative importance for the design of the platform.

5 Use cases

The DDRS offers benefit for at least the following four basic researcher driven use scenarios and six management driven use cases (refer also to figure 7).
These four use scenarios are:

- **(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 deduced 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 not more than five questions. The DDRS should not only be 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 repositories. 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 Data Management Plan (DMP)**: the user - a scholar or researcher - has to collect information for a project specific data management plan. The necessary information comprises - amongst other parameters - 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 one described above for the identification of research data repositories.
- **(C) Collect general information on research data repositories**: the user wants to inform himself on the research data repository landscape. This information interest can be focussed 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 fewer categories.

- **(D) Register a research data repository**: the user - a repository manager - wants to register a service for the DDRS. This is conveniently possible directly on re3data via the DDRS. 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 leveraged for repositories to improve their dissemination and interoperability. A suggest page exists on the DDRS in order to be able to register new repositories on re3data which then become automatically available to the DDRS.

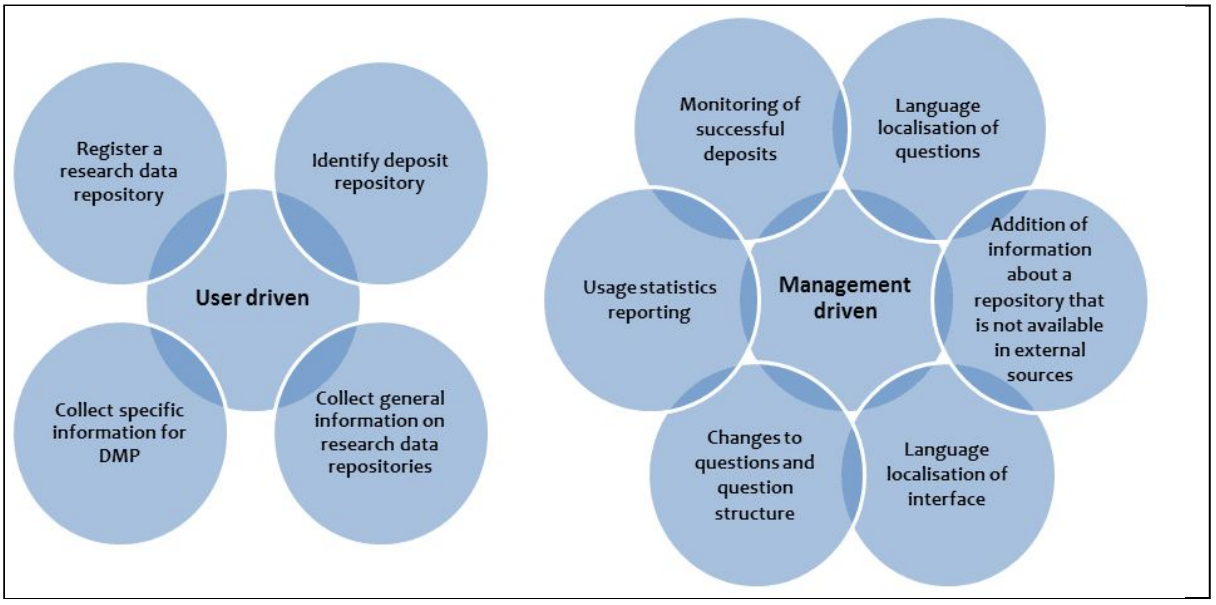

Figure 7: Use cases of the DDRS from the users' and the management's perspective

Furthermore, the DDRS system has six management use cases (refer also to figure 7):

- **(F) Language localisation of interface**: the service has to be designed in a way that future localisations can be incorporated as easily as possible. This requirement is important for the usability of the services.
- **(G) Addition of information about a repository that is not available in external sources**: in the current design state the service relies on the re3data-database. As this database does not focus on the arts and humanities there is a risk that repositories relevant for the user may not be included. 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 re3data[18]. As a reminder, the re3data-database relies upon a selected set of properties summarised in the re3data-metadata schema v.2.2[19]. 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[20].

[18] Nevertheless it is on the planned features-list to implement a re3data update function, delivering metadata on humanities-specific repositories from the DDRS to re3data. This would also reinforce the aspect of mutual benefit.
[19] Re3data's metadata schema, v.2.2: http://www.re3data.org/schema/2-2
[20] http://www.re3data.org/schema

- **(H) Monitoring of successful deposits**: this aspect relates to the usage statistics described below (see also section 1). 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. In other words: 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 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 based on their popularity; secondly the usage statistics can be used to raise the attractiveness of the service towards repositories that so far have not been included both in our DDRS or in re3data. During the demonstrator phase of the DDRS only rudimentary usage statistics are collected via Piwik/Matomo[21].
- **(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 underlying database used for the DDRS 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 become less important. This flexibility is necessary both for the questions used to identify repositories for the user but also for the data description process. The latter one is likely to be easier to adapt than the questionnaire process. Those changes can be easily done within the administration section of the DDRS.
- **(K) Language localisation of questions**: the service has to be designed in a way that future localisations can be incorporated as easily as possible. This requirement is important for the usability of the services. For now, even though it is not active by default, the interface is ready to be available in multiple languages, and has already been translated into French. The translations of the filters can also be easily made in the administration section of the DDRS.

6 Technical implementation of the DDRS

6.1 Overall approach

This section describes the technical implementation of the DDRS[22] within the Humanities at Scale project. It is important to distinguish between an ideal concept of the service and the actual implementation during the project. The latter one has to consider the availability of resources and time as well as the institutional context.

As a reminder: the DDRS assists the user in identifying suitable research data repositories for the individual case depending on only a few criteria, like formats of the research dataset, language or affiliation or certain indispensable functions[23]. The result of this step

[21] https://matomo.org/ (until January 2018 Piwik, since then Matomo)

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

[23] 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 results with more accuracy. The current practice however showed that this idea is challenging in terms of usability and in the number

is 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 proceed to the second functionality layer of the DDRS, which is about the structured description of the specific research dataset. The aim of this step is to gain, as easily and conveniently 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.

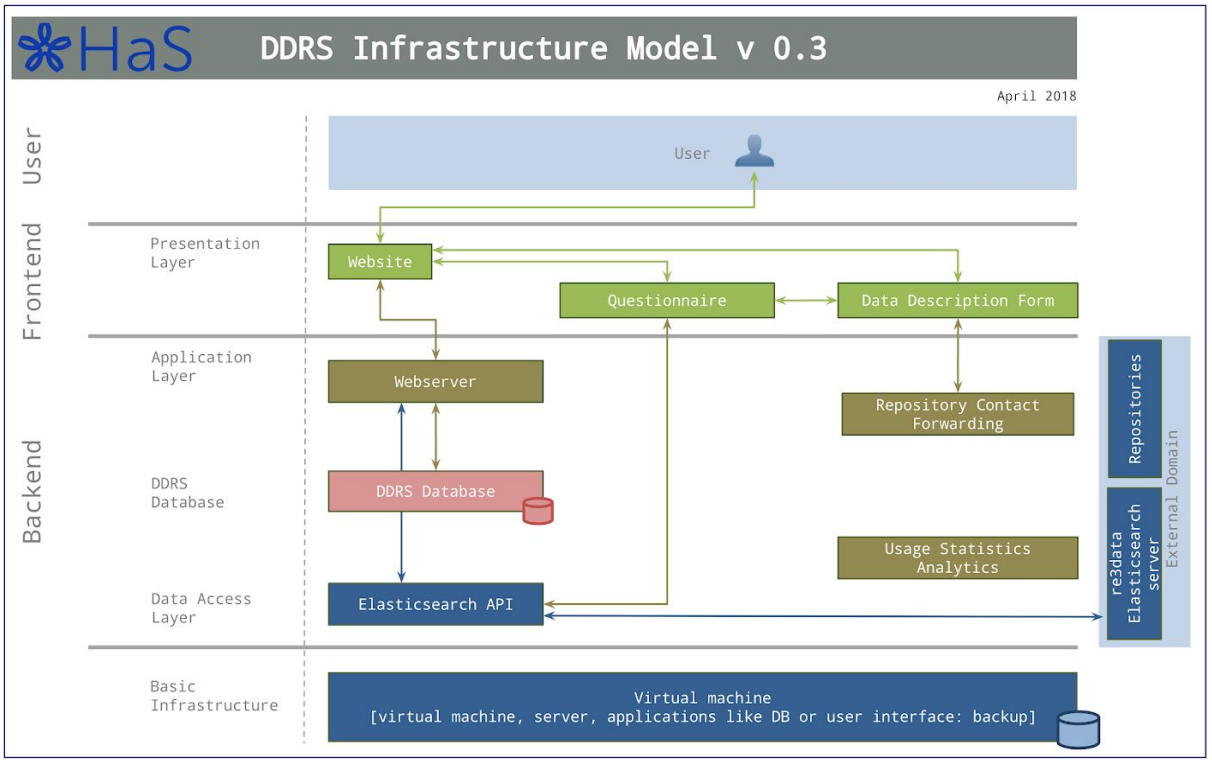

Figure 8: The DDRS infrastructure model version 0.3

Figure 8 provides an overview of the simple infrastructure which has been set up within the project. The result is a functional demonstrator, flexible to be developed further upon 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 is sufficient. The VM consists of all necessary applications and will initially be accessible over an IP address[24].

It was decided that the branding of the service would be quite close to DARIAH's, 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

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.

[24] An Internet Protocol address is used to locate the server on the network.

was also branded as DARIAH: https://ddrs-dev.dariah.eu/ also keeping in mind that the service is in a demonstration phase.
The DDRS infrastructure model (see Figure 8 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 to 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 to the questions are possible in an easy way via the administration section. This is necessary as the database used 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 specific research dataset in a structured way (this can be implemented in a similar way as the questionnaire). The questionnaire is also designed in a flexible way to allow further adjustments to 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. The current implementation is GDPR compliant as the user data gets submitted only to the selected repository contacts. The submitted user specific data is after sending not available to the DDRS.
- Currently[25] 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 impact 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 specific 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 component mails the completed data description form to the relevant repositories.
- A usage statistics component, currently Matomo. At this point it is not clear what kind of data could be collected by this service in the future. If the DDRS has a considerable user uptake in the future the usage statistics could become a valuable asset to be used for further added value services and to demonstrate the value of the service.

[25] 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.

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 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 related sub properties[26].
The DDRS now includes an additional database containing information on the points of contact 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 relies on re3data's external persistent identifiers in order to keep the information always bound to the same repository; contact information can only be connected to a single repository within re3data.
The DDRS supplementary database also includes a selected set of research data repositories of generic, national or European provenance. This ensures that a user will always 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.
However, considering these limitations the decision was still taken to use the re3data database. To our understanding re3data has the potential to grow in data quantity and usage and, for this end, it is a better choice than setting up an own exclusive database for the DDRS. Our assessment of the future development of re3data also implies 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 it would need access to their ElasticSearch servers or with the APIs that are being provided.
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).

[26] This aspect is described in detail in: Concept for a Data Deposit Recommendation Service: D 7.2 Design and Sustainability Plan for an Open Humanities Data Platform, https://halshs.archives-ouvertes.fr/halshs-01531337 chapter 5.3 user stories.

```
▸ 5:                                  {…}
▾ 6:
    _index:                           "frontend"
    _type:                            "repository"
    _id:                              "1117"
    _score:                           2.7554078
  ▾ _source:
      repositoryName:                 "DARIAH-DE Repository"
      repositoryNameLanguage:         "eng"
      repositoryUrl:                  "https://de.dariah.eu/repository"
    ▾ description:                    "DARIAH-DE supports digitally-enabled research and teaching in the arts
                                      research data as well as materials for research and teaching in the Dig
                                      Digital Research Infrastructure for the Arts and Humanities\" (http://d
                                      (http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri)."
      descriptionLanguage:            "eng"
      size:                           "45 collections; 2.732.920 documents"
      sizeUpdated:                    "2017-12-21"
      startDate:                      "2014"
      endDate:                        null
      missionStatementUrl:            "https://de.dariah.eu/wir-ueber-uns"
      versioning:                     null
      citationGuidelineUrl:           null
      enhancedPublication:            "no"
      qualityManagement:              "unknown"
      remarks:                        null
      created:                        "2015-01-20T09:39:18+01:00"
      updated:                        "2017-12-21T16:37:16+01:00"
    ▾ repositoryLanguages:
      ▾ 0:
          text:                       "eng"
      ▾ 1:
          text:                       "deu"
```


Figure 9: 94 repositories are listed as result of a query to the re3data Elasticsearch server (as of January 2021). The screenshot shows a snippet with only one repository.

Figure 9 shows a snippet of the search result of re3data's Elasticsearch server for the following query (it is not possible to 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 merged 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 available 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 number of repositories returned is reduced and the user can decide after each filter 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, a request to re3data is sent and the result is displayed. The main difference of the consolidated against the simultaneous approach is that the user doesn't see a "filter history". The user only receives the results, and, in some case, this may only be one or no repository. In terms of usability the simultaneous approach is therefore the better choice.
- **DDRS-ranked results**: multiple API retrievals of re3data are stored in the session and ranked for the user as a list. This concept is able to combine aspects of the two other concepts, but it is technically more elaborate and possibly not useful in all cases.

In practice a hybrid solution has been implemented. It is a combination of simultaneous retrieval and 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[27] as PIDs:

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

and ends up with 24 results[28]. 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

and ends up with 761 results. After applying the filter for both PID systems at once:

[27] An Archival Resource Key is a Uniform Resource Identifier (URI) that acts as a PID for any types of information objects. See also: https://en.wikipedia.org/wiki/Archival_Resource_Key

[28] All search requests described in this article have been retrieved in November 2020 and may have changed in the meantime, particularly in terms of the number of results.

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

only 13 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, the DDRS would be forced to launch multiple simple queries in order to retrieve meaningful repositories for users. This is a technical reason for liaising with the re3data's team in order to find a solution for this issue. Re3data kindly provided the team with a full ElasticSearch server on their private network which allows the DDRS to make complex queries more easily as seen below.

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

This provides 772 repositories (761 using DOI, 24 using ARK but including 13 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. In other words, the user checks five filters and the results compliant with 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[29], 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 does not include this option due to the limited number of humanities-specific research data repositories. This may change in the future.

7. Recommendations for future developments and sustainability

The sustainability of the service is crucial. The chosen concept, which builds upon an existing and well-established service - re3data - requires relatively little future maintenance compared to a completely new development. The platform has been 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[30].
An example for a future update requirement will be updating re3data's metadata schema. At the time of conceptualising the DDRS (2017), re3data uses the 2.2 version of their schema[31], but they already present version 3.0 on their website which will very likely imply changes in re3data's Elasticsearch configuration. 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 is also true if the DDRS wants to include other providers similar to re3data in the retrieval process.

[29] https://orcid.org/
[30] Although this relies on the resources available for the DDRS within the DARIAH framework.
[31] The used schema can be seen here: http://www.re3data.org/api/beta/repository/r3d100011839

An example of a likely future service extension may be related to the recommending functionality. With growing usage of the DDRS it may be useful to aggregate the usage statistics and analyse them in a way to enrich the recommendation results. Additional service extensions could also cover one or more aspects of the research data life cycle. Our chosen concept emphasises the use case of long-term preservation: the depositing of data for humanities researchers and the curation on the side of the archives. However, the DDRS may also be used in identifying suitable repositories for use in writing a project Data Management Plan (DMP)[32]. 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. Basically, two ways in which this could be implemented but they are not really related to the current implementation of the DDRS are feasible. Firstly, it should be possible for the depositor to simply post links to the deposited dataset on social media platforms, blogs, and project websites. 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[33]. 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)[34] and the Research Data Journal for the Humanities and Social Sciences[35]. The creation of a DARIAH data journal could be facilitated by the DARIAH Virtual Competence Centres (e.g. VCC3: Scholarly Content Management), for example through the organisation of a DARIAH Working Group to set up and maintain such a data journal. A recent activity in this regard - also an offspring from the Humanities at Scale project - is the Open Methods platform[36].

Other facets of the data life cycle that could be covered, is to find data and process or analyse 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 for an overview of service options[37]. However, 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 resources will be available to maintain the platform in the future, and extend it with new functionalities, will depend upon the uptake 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

[32] Some tools to help writing a DMP: https://dmponline.dcc.ac.uk or https://dmponline.be

[33] https://vocabs.dariah.eu/backbone_thesaurus/en/

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

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

[36] https://openmethods.dariah.eu/

[37] A few projects are currently working on these. The SSHOC (Social Sciences and Humanities Open Cloud - https://sshopencloud.eu/) project and its SSH Open Marketplace (coordinated by DARIAH) is in development and aims at providing tools and services alongside solutions, training materials, etc. in an effort of contextualisation of the data. The TRIPLE project (https://www.gotriple.eu/) will be the discovery platform for the OPERAS Research Infrastructure (http://operas-eu.org/) and will allow the users to find data, research profiles and projects within the SSH landscape. Both projects support SSH research in Europe.

legal status, clear decision-making structures and cost structures (Neuroth, Rapp 2016). At this point, the follow-up of the Humanities at Scale project or the DDRS as such is uncertain. In this regard it is also relevant to mention the DESIR[38] (DARIAH ERIC Sustainability Refined) project. This Horizon 2020-funded project, which ran from the beginning of 2017 until the end of 2019, developed means to enhance the usage and awareness of DARIAH and its services within the humanities research community and thereby contributed to the sustainability of the DARIAH digital research 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). Issues of updating notwithstanding, this service could be localised and hosted at a number of institutions. High bandwidth is not to be expected as this is just a simple (http) web service. At the current stage of the Humanities at Scale 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 needed service. But this may not be the right scale for a smaller infrastructure component like the DDRS. This 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 resources to run, it could also be hosted and maintained by one or more institutions as an (country-specific) in-kind contribution[39] to DARIAH-EU.

8. Conclusions

The Data Deposit Recommendation Service (DDRS) is a lightweight web service helping humanities researchers to identify suitable repositories to deposit research data. It has been developed as functional demonstrator within the DARIAH-affiliated humanities at Scale project. The complete documentation and code is available at GitHub: https://dariah-eric.github.io/ddrs/ for further development or re-use by third parties.
Beside the use case of identifying suitable deposit locations for humanities research data, the DDRS also aims at raising awareness for research data and research data repositories in the humanities.
The service has been designed knowingly relying on re3data, a widely known and established registry for research data repositories. This approach shows the potential of re-use and cooperation and is also reflected in the DDRS' flexibility to include other future functionalities beyond the repository recommendation.

9. Bibliography

Buddenbohm, S., De Jong, M., Larrousse, N., Minel, J.-L., Moranville, Y., Priddy, M. (2017): D7.2 Concept for a Data Deposit Recommendation Service: D 7.2 Design and Sustainability Plan for an Open Humanities Data Platform. https://halshs.archives-ouvertes.fr/halshs-01531337

[38] https://www.dariah.eu/activities/projects-and-affiliations/desir/

[39] More details on contributions to DARIAH https://www.dariah.eu/tools-services/contributions/

DARIAH General Assembly (2017): DARIAH Strategic Action Plan (STRAPL). News item: https://www.dariah.eu/2017/11/10/moving-forward-strategically-general-assembly-approves-dariah-action-plan/

European Commission Press Release (2016): European Cloud Initiative to give Europe a global lead in the data-driven economy. URL: http://europa.eu/rapid/press-release_IP-16-1408_en.htm

Neuroth, H., Rapp, A. (2016): Nachhaltigkeit von digitalen Forschungsinfrastrukturen. In: Bibliothek Forschung und Praxis 2016; 40(2). DOI https://doi.org/10.1515/bfp-2016-0022

Vierkant, P., Spier, S., Ruecknagel, J., Pampel, H., Fritze, F., Gundlach, J., Fichtmüller, D., Kindling, M., Kirchhoff, A., Göbelbecker, H.-J., Klump, J., Kloska, G., Reuter, E., Semrau, A., Schnepf, E., Skarupianski, M., Bertelmann, R., Schirmbacher, P., Scholze, F., Kramer, C., Witt, M., Fuchs, C., Ulrich, R. (2014): Schema for the description of research data repositories: version 2.2, DOI: http://doi.org/10.2312/re3.006