D 4.1 - Gap Analysis of DARIAH Research Infrastructure

Stefan Buddenbohm, Raisa Barthauer

D4.1 Gap Analysis of DARIAH Research Infrastructure

DESIR

DARIAH ERIC Sustainability Refined

INFRADEV-03-2016-2017 - Individual support to ESFRI and other world-class research infrastructures
Grant Agreement no.: 731081

Date: 01-12-2017
Version: 0.5

DESIR has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731081

Grant Agreement no.:	731081
Programme:	Horizon 2020
Project acronym:	DESIR
Project full title:	DARIAH-ERIC Sustainability Refined
Partners:	DIGITAL RESEARCH INFRASTRUCTURE FOR THE ARTS AND HUMANITIES GEORG-AUGUST-UNIVERSITAET GOETTINGEN STIFTUNG OEFFENTLICHEN RECHTS UNIVERSITEIT GENT UNIWERSYTET WARSZAWSKI FACULDADE DE CIÊNCIAS SOCIAIS E HUMANAS DA UNIVERSIDADE NOVA DE LISBOA CENTAR ZA DIGITALNE HUMANISTICKE NAUKE GOTTFRIED WILHELM LEIBNIZ UNIVERSITAET HANNOVER INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET AUTOMATIQUE KING'S COLLEGE LONDON UNIVERSITY OF GLASGOW KNIHOVNA AV CR V. V. I. HELSINGIN YLIOPISTO SIB INSTITUT SUISSE DE BIOINFORMATIQUE UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA UNIVERSITY OF HAIFA UNIVERSITY OF NEUCHATEL
Topic:	INFRADEV-03-2016-2017
Project Start Date:	01-01-2017
Project Duration:	36 months
Title of the document:	D4.1 Gap Analysis of DARIAH Research Infrastructure
Work Package title:	WP4 Technology
Estimated delivery date:	December 2017
Lead Beneficiary:	UGOE-SUB
Author(s):	Stefan Buddenbohm (buddenbohm@sub.uni-goettingen.de)

	Raisa Barthauer (barthauer@sub.uni-goettingen.de)
Quality Assessor(s):	Claudio Leone (leone@sub.uni-goettingen.de) Yoann Moranville (yoann.moranville@dariah.eu)
Keywords:	DARIAH, research infrastructure, sustainability, technology, gap analysis, demonstrators, visualisation, scholarly content management

Revision history

Version	Date	Author	Beneficiary	Description
0.1	30.06.2017	Stefan Buddenbohm	UGOE-SUB	First draft prepared for the F2F on July 11th in Hanover
0.2	20.10.2017	Stefan Buddenbohm	UGOE-SUB	Second draft based on the results from the F2F; Partner descriptions, completion of inventory, diagrams
0.3	10.11.2017	Raisa Barthauer	UGOE-SUB	Diagrams
0.4	17.11.2017	Stefan Buddenbohm	UGOE-SUB	Discussion of report at WP F2F on November 21st in Berlin
0.5	1.12.2017	Stefan Buddenbohm	UGOE-SUB	Completion of report and final feedback round with DESIR consortium
0.6	14.12.2017	All partners	All partners	Global checking
1.0	15.12.2017			Submission

Executive Summary

DESIR (DARIAH ERIC Sustainability Refined) work package 4 Technology contributes to DARIAH's long-term sustainability by technologically enhancing DARIAH's research infrastructure and services. To this end four new and innovative technology areas – entity-based search, scholarly content management, text analytic services and visualisation – shall shape DARIAH's profile.

The present document forms a prerequisite for the aforementioned objective and conducts for that reason a gap analysis of the current DARIAH infrastructure, concentrating on four areas: text analytic services, scholarly content management, entity-based search and visualisation. The competencies and resources of the three technology partners in this work package correspond to these four areas. The conceptual models - as proposal for the identified gaps - are subject to later deliverables in the project, but in the course of undertaking the gap analysis already first impressions of possible conceptual models are apparent and will be presented here. The concept of the gap analysis in this context means not only to identify missing tools or services in the DARIAH infrastructure landscape but also to recognize potential for enhancement with already existing tools and services.

Nature of the deliverable		
✓	R	Document, report
	DEM	Demonstrator, pilot, prototype
	DEC	Websites, patent filings, 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

DESIR has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731081. 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.

Table of Contents

1 Introduction and scope of the report

"Presently, there exists a vast library of tools, services, software [...]. One of the principal aims of DARIAH is to promote and facilitate the re-use of such materials, in order to accelerate new research, enable a more responsive and open publishing environment for humanities data, allow collaborations to be concluded more efficiently, with the overall goal of supporting not just digital services and complex digital objects in arts and humanities research but to broaden and deepen the adoption of digital methods and digital workflows amongst humanities scholars."[1]

This overall mission statement of DARIAH - the Digital Research Infrastructure for the Arts and Humanities[2] - is also guiding the DESIR project.

Against this background the aim of work package 4 and the document at hand is twofold: first and foremost an inventory of the current DARIAH infrastructure according to four specific areas will be compiled. These areas include entity-based search, scholarly content management, text-analytic services and visualisation. This structured overview represents the current status quo not only in terms of general infrastructure but also in terms of individual tools, services or software, being provided or maintained by DARIAH.

The inventory should ideally cover in this context all national branch activities – e.g. DARIAH-DE or DARIAH-FR – and includes also DARIAH-EU resp. the DARIAH ERIC[3] as supranational administration umbrella for all DARIAH activities. Nevertheless our gap analysis[4] will only result in a fragmentary picture of all existing activities, also because it is not necessary in this respective context. On the one hand this is connected to the by now huge bandwidth of European DARIAH activities and the steady expansion in terms of participating countries, community growth or differentiating services and resources. On the other hand there is currently no overarching inventory or registry, although there are several inventories on a national level. The upcoming DARIAH market place could possibly be some kind of central information hub for services, tools and resources - at least for the in-kind contributions of the partner countries and will be marked in the subsequent chapter. With regard to the description of work this fragmentary picture shouldn't be a problem as the main outcome of this work package shall be conceptual models with potential of contributing to the DARIAH landscape.

Secondly the resulting DARIAH inventory will undergo a gap analysis with respect to technology specialisations on side of the DESIR technology partners. The gap analysis will

[1] Baillot, Mertens, Romary (2016), p. 153
[2] http://www.dariah.eu/
[3] European Research Infrastructure Consortium
[4] The term gap analysis in the context of this document is not to be confused with the gap analysis in an economic sense, see https://en.wikipedia.org/wiki/Gap_analysis . Gap in our sense means gap in terms of infrastructure, services or resources.

focus on four main areas which have been predetermined by the description of work for the DESIR project:

- entity-based search,
- scholarly content management,
- text-analytic services and
- visualisation.

The document at hand should not only identify gaps in the DARIAH inventory but rather explore possibilities to close the gaps. This step will be detailed subject of the subsequent deliverable D4.2 Conceptual Models. D4.2 has to present conceptual modules for the identified gaps with respect to the aforementioned four areas. The further process of work package 4 will subject the conceptual models to a code-sprint, validate them, and prepare a proof-of-concept.

2 Technology partners in WP4

The following table provides a short overview on the involved partners in this work package. It has to be seen in relation to the subsequent tables in this paper which form an additional information basis for the gap analysis.

Table 1: Technology partners in DESIR WP4

Institution	Short description, focus of work	Employed technologies	Roles	Example
UGOE-SUB Göttingen State and University Library, Research & Development Department	RDD: Develops new infrastructure services for research; largest R&D library in Germany; 2016: 15 FTE; 14 third-party funded projects; over 80% third-party funding; founded in 2002 Research data centres: collecting, enriching, curating and preserving data Virtual research environments: collaborative data creation; processing and analysis	Java (web services); Javascript (Jquery, Ractive); RDF (esp. TGForms); existdb; elasticsearch; Maven ; Jenkins; Puppet; Tomcat; public and private git repositories	developer (maintainer, provider)	TextGrid VRE and repository for scholarly editions; XML-TEI; Java-Backend; Eclipse-based Frontend for all OS; developed since 2005; https://fontane-nb.dariah.eu http://bdn-edition.de/ http://www.mayawoerterbuch.de
ICM Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw	Basic unit of the university; leading HPC and R&D centre in Poland; provides research infrastructure within ICT and HPC; interdisciplinary research; applied	C/C++; Fortran; Java(Web, Desktop); Scala; Python; JavaScript; R; GitHub; GitLab; SVN; Redmine; Jenkins; Maven; Apache/Tomcat; Hadoop ecosystem;	developer maintainer provider	VisNow - a generic visualization framework http://visnow.icm.edu.pl

	mathematics; computational sciences and data science	multiple DB engines		
INRIA Institut national de recherche en informatique et en automatique represented by the ALMAnaCH team (Automatic Language Modelling and Analysis & Computational Humanities)	INRIA is the French institute for computer science and applied mathematics; it promotes "scientific excellence for technology transfer and society"; divided in various research centres; no major policies in terms of software sustainability or applied technologies; only a centralised repository on GitLab is provided; development approach depends on specific team; open software; hosted externally	Java; Python; Javascript (jquery, angularjs); Elasticsearch; lucene; mysql; Machine learning toolkits; NLP libraries; Maven; travis/jenkins; git; Json; XML (mostly TEI); XSLT	developer maintainer	(N)ERD - a (Named) entity recognition and disambiguation service: http://nerd.huma-num.fr http://github.com/kermitt2/nerd GROBID - Generation Of BIbliographic Data http://github.com/kermitt2/grobid http://github.com/kermitt2/grobid-ner http://github.com/kermitt2/grobid-quantities
L3S at the University of Hanover	Research interests of L3S: web science (Twitter, Wikipedia, etc.); Big Data (Hadoop, Spark, etc.); Text Mining; NLP; Data Mining; Machine Learning; Crowdsourcing; applied technologies and methods depend on the respective project, there is no mandatory standard procedure	Java, Maven, Hadoop, Spark, (Scala), Python, Tomcat, MySQL, git, SVN, Jenkins, Weka, RapidMiner, Eclipse, IntelliJ, NLP libraries	developer maintainer provider	BibSonomy - a social bookmarking system for web links and literature references http://www.bibsonomy.org/

As extension of the preceding information compilation we prepared an overview on the partners' applied workflows and technologies in order to better understand the course of the work in work package 4. By this we are able to evaluate the DARIAH inventory right from the start with the available resources and competencies in mind as preparatory work for the collection of ideas for conceptual models or demonstrators.

2.1 Applied technologies among the technology partners

The following table lists the applied development, maintenance and provisioning technologies amongst the partners in WP4. This information can be useful to assess compatibilities or synergies in the development of the conceptual models.

Table 2: Applied technologies among the DESIR technology partners in WP4

Question	ICM	INRIA	L3S	UGOE-SUB (RDD)
What workflows or methods do you employ? (Agile/Scrum vs. Waterfall, DevOps vs. HandOff, Git-Flow, etc.)	Agile/Scrum-like/Hand Off/git-flow	N/A	git-flow	Git-flow, some devops
What programming languages do you use? Which framework?	Java, C/C++, Python, Scala	Java, Javascript, Python	Java, JavaScript, Python, Scala, PHP	Java, Javascript (JQuery, Ractive), HTML, PHP, Python, C, C++, C#
Where and how do you host your code?	Github, GitLab (private+public)	Github, GitLab	Bitbucket & GitHub (public code), GitLab (internal code)	GitHub (for public code), GitLab (for internal code), Openprojects (internal+public)
Are you applying automated testing (unit/ integration)? If yes, which tools?	Yes: Jenkins, Gitlab-CI	Yes (Travis/CircleCI) we also use codeCov and coveralls	Yes (Jenkins) - only for BibSonomy	Yes Jenkins, also Gitlab-CI and some Travis
Are you applying (formalised) code review?	Yes (some projects, e.g. VisNow)	No	No (Yes for BibSonomy)	No
In which way do you document your code?	Javadoc formalised in code review (VisNow).	Comments (however not always a shared approach)	Eclipse settings enforcing comments	Javadoc, sometimes just comments, in part with Sphinx
Where do you operate the systems?	Own HPC Centre	Inria research center server	Dedicated/computing centre server rooms in Hannover, Kassel and Würzburg	GWDG (local computing center)
How do you provide systems? Openstack/ VMWare/ Bare Metal	Bare Metal/ VMWare/KVM	Our team has some dedicated Bare Metal servers.	Bare Metal	VMWare for important Services, Openstack, Some test systems on KVM on Bare Metal
Do you use configuration management? (in ITIL sense or tools like puppet/ chef/ ansible)	Yes: SVN, Salt, Ansible, Puppet	The ALMAnaCH team has 5 servers. There are dedicated teams doing some basic maintenance. The rest is manual work. The servers are usually allocated for demos and heavy processing.	No (config in RCS)	Yes: Puppet, No ITIL
How do you document systems and their state?	Wiki, Nagios, Icinga2	In a PAD, servers are usually not shared among many people.	Wiki, Nagios	Wiki, Nagios
Do you use tools for time series analytics / metrics?	Yes: MRTG, Munin, Zabbix	No	No	Yes: Grafana for Telegraf Data
Web/Usage Analytics	Yes: Piwik, Other	No	Awstats, Piwik	Piwik

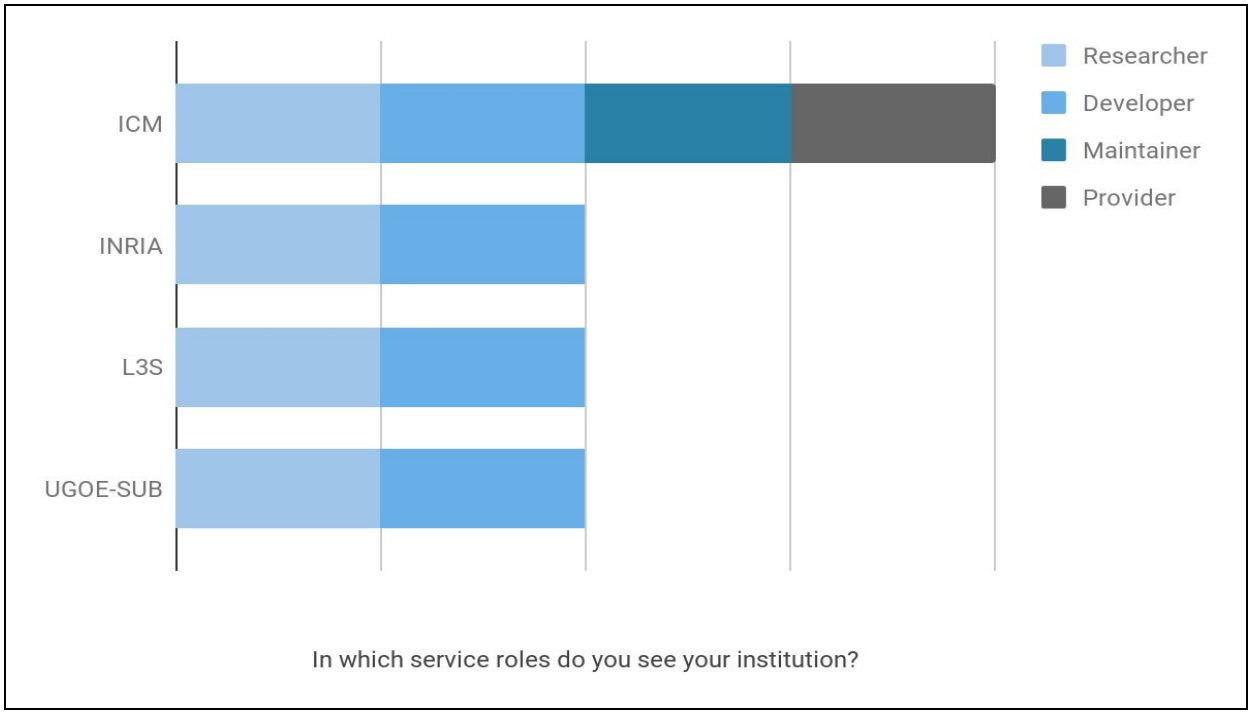
2.2 Service roles among the technology partners

The table below collects some meta-information on the roles of the technology partners. Aim of this paragraph is to be able to evaluate the yet to come conceptual models in a strategic perspective. Context: it is not uncommon that service are being developed by an institution and that the same institution subsequently takes further care of the service, be it as service provider, maintainer or in a developer role. As long as the institution already fulfils all of these roles, this may be unproblematic. In practice many institutions are unhappy with the maintainer and provider role but fulfil them given that there is no regulating policy available for this issue.

Table 3: Service roles among technology partners in DESIR WP4 (self assessment)

In which service roles do you see your institution?	ICM	INRIA	L3S	UGOE-SUB
Researcher	X[5]	X	X	X
Developer	X	X	X	X
Maintainer	X			
Provider	X			

Figure 1: Service roles of technology partners in DESIR



[5] X = applies.

The abovementioned diagram has to be read with some scepticism as with one exception – ICM – all partners envisage themselves not in the service provider role. Partly and motivated by the development and further support for certain services a partner might be sliding into a maintainer and/or provider role but usually don't take this role. Read with a certain degree of sincerity one can assume that all partner institutions see their focus on research and research-related development activities. Maintenance and provision comes inevitably here and there into effect but is clearly not seen as work focus and is always dependent on the available funding resources. This leads to certain consequences for the development of digital services and resources, mainly in terms of standardisation and documentation.

2.3 Current services or software among the technology partners

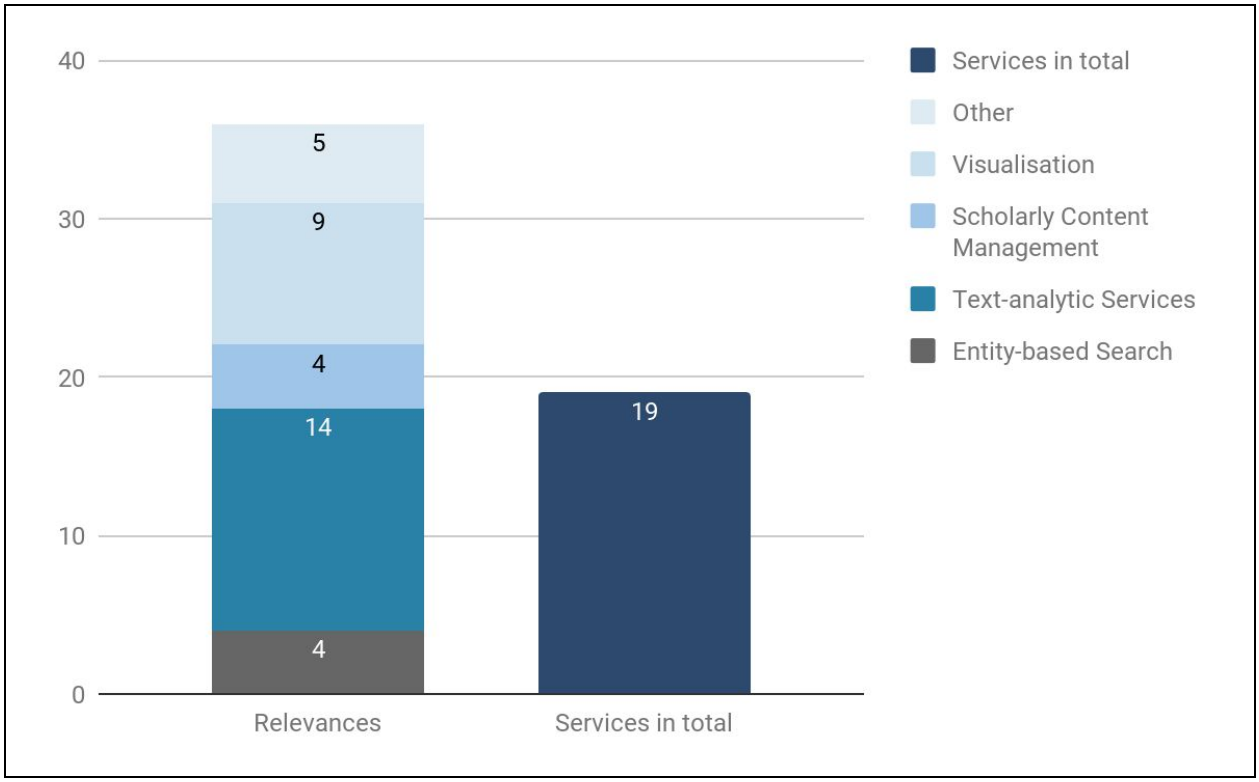
Table 4: Development and service competencies among the technology partners in DESIR WP4

Service	Partner	Tags	Entity-based search	Scholarly content management	Text-analytic services	Visualisation	Other
https://www.bibsonomy.org/	L3S	Social Bookmarking	X[6]	X			
http://git.l3s.uni-hannover.de/	L3S	Code hosting/ issue tracking					X
http://nerd.huma-num.fr	INRIA	Data extraction			X		
https://github.com/helgeho/ArchiveSpark	L3S	Data extraction	X		X		
http://scraper.bibsonomy.org/	L3S	Data extraction		X			
http://traces1.saclay.inria.fr/grobid	INRIA	Metadata conversion		X	X	X	
https://hal.archives-ouvertes.fr/DARIAH	INRIA	Publication repository		X		X	
http://traces1.saclay.inria.fr/inria/	INRIA	Repository analytics	X	X		X	X
https://textgridrep.org/en/	UGOE-SUB	Research data repository		X		X	X
https://de.dariah.eu/repository	UGOE-SUB	Research data repository		X		X	X
https://textgrid.de/	UGOE-SUB	Virtual research environment		X		X	

[6] X = applies

https://visnow.icm.edu.pl/	ICM	Data visualisation				X	
http://remotevis.icm.edu.pl/ (non-public)	ICM	Desktop cloud visualization				X	
http://cermine.ceon.pl/	ICM	Data extraction	X	X	X		
http://comac.ceon.pl	ICM	Bibliography analytics		X		X	
https://repod.pon.edu.pl	ICM	Research data repository		X			X
http://repozytorium.ceon.pl	ICM	Publication repository		X			
http://bibliotekanauki.ceon.pl	ICM	Journal platform		X			
http://otworzksiazke.pl	ICM	Book platform		X			
= 19 services			4	14	4	9	5

Figure 2: Services and software among the technology partners in DESIR WP4



The aforementioned diagram shows a quite even distribution of services and tools along the four categories suggested by DESIR's description of work. As a fifth category non applicable services and tools have been added as they complete the partners' inventory. The underlying n=services/tools falls short with 12 but one may recognise focal points in the development of services relevant for scholarly content management and visualisation of data. To relativise the picture even more, the distribution of services along the categories reflects the specific work package partners, hence the results are not surprising.

3 Partial catalogue of services and tools within DARIAH

3.1 Existing cataloguing activities within DARIAH

DESIR stands for DARIAH ERIC Sustainability Refined and is – similar to Humanities at Scale[7] (HaS) – an offspring project of DARIAH and therefore closely embedded in the DARIAH-ERIC and DARIAH-EU context. Out of context of the DARIAH-ERIC at least four main missions can be deducted[8], which guide the objectives of DESIR:

- Identify infrastructural needs of scholarly communities and provide help to fulfil those needs;
- Coordinate national contributions in establishing sustainable digital services for the arts and humanities;
- Contribute to the establishment of national infrastructure roadmaps;
- Participate in setting up the European agenda for infrastructures in Europe.

DESIR addresses the sustainability question for several kinds of activities, infrastructures or services originating from the DARIAH context and it is not the first project working on this question. Different from developing new resources or infrastructural components, DESIR is exploring opportunities to employ already existing resources (independent from DARIAH) as means to sustain certain infrastructure components and services. Within the context of HaS the DESIR work package 4 Technology and the actual deliverable 4.1 Gap Analysis are two activities that should be considered more thoroughly in this context:

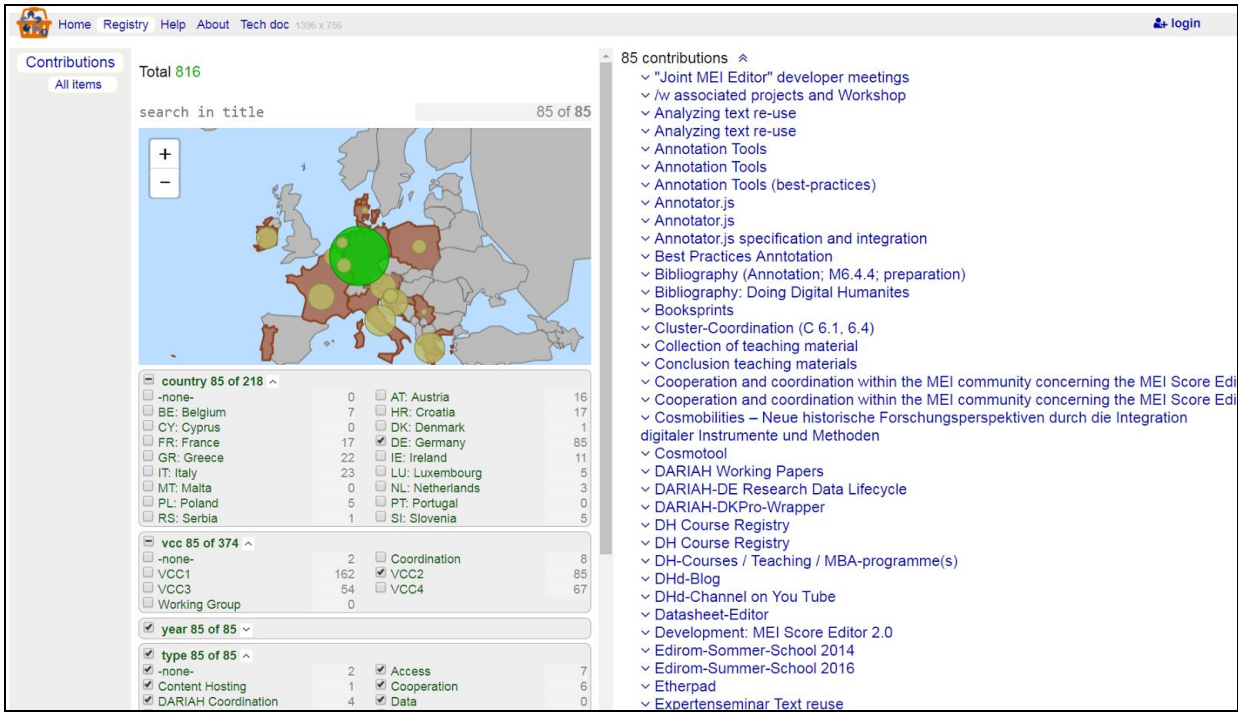
- DARIAH (In-kind) contribution tool[9] (see figure 3): the in-kind contribution tool has been an activity within the aforementioned HaS project's work package 5. Under the lead of DANS a ReactJS based interface has been developed which provides the user access to a database containing and allowing the assessment of the in-kind contributions of the national DARIAH partners to the DARIAH ERIC, basically an inventory. Not all of these in-kind contributions come in terms of infrastructure or

[7] Project website for Humanities at Scale: http://has.dariah.eu/
[8] Baillot, Mertens, Romary (2016), p. 152
[9] A provisional development instance of the In-kind contribution tool is available here: https://dariah-beta.dans.knaw.nl/data/contrib/filter , the technical documentation may be found here: https://dans-labs.github.io/dariah/ and the user manuals on Github here: https://github.com/Dans-labs/dariah/wiki

user oriented services but a considerable subset which is likely to be of interest within this context. There are plans to implement a DARIAH Hallmark for some of these resources. Possibly this concept and particularly the underlying criteria are also of interest in our context.

- DARIAH Marketplace: the launch of the Marketplace is planned for 2018. It will be "a central, easy-entry place for humanists seeking support for all kinds of digital aspects of their research. [...] It will contain a collection of software, tools, services, datasets, publication repositories and learning and training material and will establish visibility for them."[10] It is planned that the abovementioned DARIAH (in kind) contribution tool feeds into the DARIAH Marketplace. Not all of the listed resources are of interest in the context. Generally DARIAH differentiates resources between services and activities, under which such diverse activities such as events, software development or consulting may be subsumed.[11] Relevant for the gap analysis are in this regard mainly the services, more precisely (data) hosting services, (processing) services and support services.

Figure 3: The DARIAH In-Kind Contribution tool (as of November 2017)



[10] Frank Fischer, Jennifer Edmond (2017): Towards the DARIAH Marketplace - Position Paper. (only internally available) but also presented in a talk at the DARIAH Innovation Forum on November 2nd 2017 in Aarhus, Denmark: http://conferences.au.dk/dariah2017/preliminary-program/

[11] Humanities at Scale, Lisa de Leeuw, Dirk Roorda (2017), p. 14

Other relevant activities in this context include the DARIAH-IT Registry[12], the Joint Resource Registry[13] of PARTHENOS or the overhauled new website of DARIAH-EU, which also contains an overview of tools and services[14], and the DARIAH-DE list of tools and services[15]. This document considers these sources, particularly the DARIAH-DE list of tools and services and the DARIAH in-kind contribution tool. Nevertheless it has to be clear that the aforementioned activities follow a different scope than work package 5 of DESIR, which is particularly interested in identifying gaps to be closed by the current DESIR partners.
An interesting resource beyond the scope of DARIAH (as the above described initiatives except for PARTHENOS) can be found with TERESAH. TERESAH stands for Tools E-Registry for E-Social Science, Arts and Humanities and is the result of an DASISH[16] effort and currently under further development by the already mentioned Humanities at Scale project. A new metadata schema has been developed and has been implemented into an own instance of TERESAH[17]. It is a cross-community tools and knowledge registry aimed at researchers in the social sciences and humanities. It is of particular interest with its distributed approach as it crawls its data from different sources (e.g. community-driven websites) and aggregates them in a re-usable way, for instance in various RDF representations[18]. Although TERESAH presents an interesting example from a technological point of view in this context, its scope lays beyond the DARIAH resources which should be inventoried here.

3.2 Inventory of DARIAH activities in four areas

The following section provides an overview of the current infrastructure and service offerings of DARIAH according to the four areas: entity-based search, scholarly content management, text-analytic services and visualisation. As mentioned before, this deliverable is not tasked with the preparation of a comprehensive inventory but is said to focus on certain areas.
In the following step this inventory will be screened for gaps – or rather potentials – for contributions by the technological partners in DESIR. The table presents publicly offered services such as desktop applications, web applications, mobile applications and APIs, basically in the sense of service as the DARIAH (in kind) contribution tool understands them: (data) hosting service, (processing) service, support service and access to resources. The services may be aimed at basic infrastructure requirements, scholarly collaboration, conducting of research and user services. Where possible, the services are linked with a URL.

[12] http://it.dariah.eu/registro/en/
[13] http://www.parthenos-project.eu/Download/Deliverables/D5.2_Report_on_design_Joint_Resource_Registry.pdf
[14] https://www-dev.dariah.eu/what-we-provide/tools-list/
[15] https://de.dariah.eu/list-services which will perspectively be included in the DARIAH-EU Marketplace
[16] http://dasish.eu/
[17] http://teresah.dariah.eu/
[18] See here http://teresah.dariah.eu/about/rdf

The listed resources focus on services, tools and software instead of content. Therefore publication and research data repositories or digital editions are intentionally left out. This selection is not always possible, this is why to some degree databases have been incorporated in the inventory. Again: an exhaustive inventory of all available DH resources within DARIAH will be only within reach with the upcoming marketplace.

Also the reliability of the collected information has to be considered. Information on websites regarding the status of a service are often unclear. Is it still being developed, is it a productive service being used, is the service discontinued and just present for documentation purpose? To address this issue a column has been added in the table, displaying the maintenance status of the services. An x in this column indicates that from the information on the website we assume that the service is still active to some degree.

So far the following sources have been investigated for services[19]:

- DARIAH-DE: https://de.dariah.eu/list-services contains a comprehensive list of available services and tools that qualify for production use. As DARIAH-DE passed through two funding phases and is now in the midst of the third funding phase and paid considerable effort to the sustainability of the created infrastructure, the overview - nearly a service catalogue - on the DARIAH-DE website is a reliable portfolio of available and functional services. A comprehensive - although in German language - overview of the service portfolio can be found in Kálmán, Kong, Schwardmann (2016).[20]
- DARIAH-EU: the website http://dariah.eu/ gives currently only a glimpse overview of the resources that have been developed within the national branches of DARIAH. Even the relaunched website[21] of DARIAH-EU will only be able to offer a complete overview once the DARIAH Marketplace will be launched. The reasoning behind the DARIAH Marketplace and the according In-Kind-Contribution Tool is described a bit more in the previous chapter.
- DARIAH-IT: http://it.dariah.eu/sito/strumenti/ showcases a broad collection resources being developed in humanities-related research contexts in Italy. The relation to DARIAH is more or less loose which makes it difficult to identify services that are functional and accessible beyond a funded project period. Beyond this, it remains unclear along which criteria the resources have been selected or whether they are up to date. Therefore only a small assortment has been selected for the inventory below and still the listed resources have to be considered with caution.

[19] As of September 2017.
[20] Kálmán, Kong, Schwardmann (2016), p. 234ff.
[21] https://www-dev.dariah.eu/what-we-provide/tools-list/ as preliminary development instance.

- **DARIAH-GR: the website of Greek DARIAH branch http://www.dyas-net.gr/ contains a list of Greek humanities-related tools and services. As long as the resources seem to be available in English they have been considered for our inventory.**
- **DARIAH-BE: the website of DARIAH-BE http://be.dariah.eu/services-overview offers a comprehensive list of tools and services for digital scholarship in the arts and humanities in Belgium. Although it is not clear along which criteria the resources have been selected, the listed resources give the impression to be more or less functional and available.**
- **DARIAH-AT: http://www.digital-humanities.at/de/dha/projects the Austrian DH resources reveal a strong emphasis on digital editions or other content-focussed resources. Only a few of them are considered in the inventory below although they make the impression of being functional and accessible on a beyond-project base. Another feature of the Austrian resources is the CLARIN provenance. Services with a clear CLARIN provenance haven't been considered in the inventory.**
- **DARIAH-FR & Huma-Num: http://www.huma-num.fr/services-et-outils Huma-Num serves as gathering point for humanities-related French research services. Huma-num mission is to ensure the preservation of the scientific heritage of the laboratories and to reduce recurrent costs by pooling an infrastructure by co-managing and sharing tools, instruments and data management systems. Huma-num is also part of the DARIAH organisation providing their infrastructure for european projects.**
- **CLARIAH-NL: https://www.clariah.nl/en/ The Common Lab Research Infrastructure for the Arts and Humanities is not a national DARIAH branch as DARIAH-DE or DARIAH-BE but can be explained with the strong CLARIN uptake in the Netherlands. CLARIAH-NL offers a distributed infrastructure for the humanities and social sciences. Different from the national DARIAH websites above, CLARIAH doesn't offer list or registry of tools and services on its website.**
- **For Switzerland, the ASSH has open a German webpage for DARIAH announcements: http://www.sagw.ch/sagw/laufende-projekte/dariah.html**

The aforementioned list of resources is by no means complete but may be sufficient for an overview of the available services to deduct focal points or gaps.

Table 5: DARIAH inventory according to the four DESIR DoW categories

Service[22]	Tags	Entity-based search	Scholarly content management	Text-analytic services	Visualisation	Other category (not covered by DoW)	Maintained or applicable[23]
https://de.dariah.eu/gnd-service	Authority files	X[24]	X				X
https://de.dariah.eu/aai	AAI					X	X
https://www.zotero.org/groups/113737/doing_digital_humanities_-_a_dariah_bibliography	Bibliography		X		X		X
http://dhd-blog.org/	Blog		X				X
https://hypotheses.org/	Blog		X				X
https://www.youtube.com/channel/UCScSbG7XjiXbZVgilEpoPkw	Content, YouTube		X				X
https://www.oercommons.org/groups/dariah/229/	content		X				X
https://de.dariah.eu/en/cloudshare	Cloud/ File sharing					X	X
https://sourceforge.net/projects/meise/	Data analysis		X		X		(X)
https://de.dariah.eu/topics	Data analysis	X	X	X	X		(X)
https://de.dariah.eu/digivoy	Data analysis			X	X		X
https://de.dariah.eu/mysql-datenbankhosting	Database					X	(X)
https://pdrprod.bbaw.de/wiki/doku.php?id=de:start	Database, Research Data Repository		X	X	X		

[22] Service is defined according to D5.2 DARIAH (in kind) contribution tool as (Data) hosting service, (processing) service, support service and access to resources.
[23] Maintained or applicable is to understood as qualitative criteria in this regard. A service was recognized as maintained or applicable only if the hosting website appeared to be up to date.
[24] X = applies; (X) = could apply

Service[25]	Tags	Entity-based search	Scholarly content management	Text-analytic services	Visualisation	Other category (not covered by DoW)	Maintained or applicable[26]
https://de.dariah.eu/pnd-service	Database, Research Data Repository		X	X	X		X
https://de.dariah.eu/jenkins	Developer tools for CI workflows					X	
https://de.dariah.eu/etherpad	Etherpad					X	X
https://de.dariah.eu/geobrowser	Geovisualisation		X		X	X	X
https://de.dariah.eu/basisinfrastruktur-helpdesk	Helpdesk					X	(X)
https://de.dariah.eu/mailinglisten	List servers					X	X
https://colreg.de.dariah.eu/colreg/	Metadata, Research Data		X			X	
https://de.dariah.eu/cosmotool	Data analysis	X		X	X		
https://schereg.de.dariah.eu/schereg/registry/	Metadata conversion			X			
http://tadirah.dariah.eu/vocab/index.php	Metadata conversion, Thesaurus management	X		X			
https://de.dariah.eu/monitoring	Monitoring					X	X
https://de.dariah.eu/conedakor	Network visualisation, Research Data Repository	X	X				

[25] Service is defined according to D5.2 DARIAH (in kind) contribution tool as (Data) hosting service, (processing) service, support service and access to resources.
[26] Maintained or applicable is to understood as qualitative criteria in this regard. A service was recognized as maintained or applicable only if the hosting website appeared to be up to date.

Service[27]	Tags	Entity-based search	Scholarly content management	Text-analytic services	Visualisation	Other category (not covered by DoW)	Maintained or applicable[28]
https://de.dariah.eu/datasheet-editor	Metadata conversion, Geospatial visualisation				X	X	X
https://de.dariah.eu/dkpro-wrapper	NLP tools			X			X
https://www.gwdg.de/de/application-services/persistent-identifier-pid	PID service					X	(X)
https://de.dariah.eu/persistent-identifiers	PID service					X	X
https://projects.gwdg.de/	Project management					X	X
https://redmine.acdh.oeaw.ac.at/	Project management					X	(X)
http://calenda.org/	Public calendar					X	X
https://de.dariah.eu/repository	Repository, content		X				X
https://de.dariah.eu/publikator	Repository, content		X				X
http://books.openedition.org/	Repository, content		X				X
https://textgridrep.org/en/	Repository, content		X				X
http://www.revues.org/	Repository, content		X				X
https://hal.archives-ouvertes.fr/DARIAH	Repository, content		X				X
https://www.rechercheisidore.fr/apropos	Repository, content		X				

[27] Service is defined according to D5.2 DARIAH (in kind) contribution tool as (Data) hosting service, (processing) service, support service and access to resources.
[28] Maintained or applicable is to understood as qualitative criteria in this regard. A service was recognized as maintained or applicable only if the hosting website appeared to be up to date.

Service[29]	Tags	Entity-based search	Scholarly content management	Text-analytic services	Visualisation	Other category (not covered by DoW)	Maintained or applicable[30]
https://de.dariah.eu/generische-suche	Search		X				X
https://de.dariah.eu/survey-provisioning	Survey provisioning					X	(X)
https://registries.clarin-dariah.eu/courses/courses/	Teaching, content		X				X
https://de.dariah.eu/getty-service	Thesaurus management	X					(X)
https://de.dariah.eu/hosting-umgebung	Virtual machines/ web hosting					X	(X)
https://de.dariah.eu/virtuelle-maschinen	Virtual machines/ web hosting					X	X
https://textgrid.de/en/	VRE[31]		X	X			X
http://www.cendari.eu/	VRE		X	X			X
https://wiki.de.dariah.eu/display/DARIAH/Home	Wiki/ Collaborative platforms					X	X
http://www.mirabileweb.it/	Content, knowledge management system		X				X
http://www.lexicon.unisi.it/public/	Content, Data analysis			X			(X)
https://sourceforge.net/projects/evt-project/	Visualisation				X		X
https://www.idem.garr.it/	AAI					X	X

[29] Service is defined according to D5.2 DARIAH (in kind) contribution tool as (Data) hosting service, (processing) service, support service and access to resources.
[30] Maintained or applicable is to understood as qualitative criteria in this regard. A service was recognized as maintained or applicable only if the hosting website appeared to be up to date.
[31] Virtual Research Environment

Service[32]	Tags	Entity-based search	Scholarly content management	Text-analytic services	Visualisation	Other category (not covered by DoW)	Maintained or applicable[33]
http://elearning.enea.it/?lang=en	Teaching		X				
http://www.afs.enea.it/project/tigris/ASTEC.php	VRE				X		
http://www.evalita.it/2011/tasks/dependency_parsing	NLP tool			X			X
http://www.daphnet.org/index_de.html	VRE		X			X	X
http://www.languagelibrary.eu/owl/simple/	Linked data, data analysis	X					
http://trame.fefonlus.it/trame/index.html	VRE			X			
http://vcg.isti.cnr.it/3dhop/	Visualisation				X		X
http://vcg.isti.cnr.it/vcglib/	Visualisation				X		X
http://www.meshlab.net/	Visualisation				X		(X)
http://www.cophilab.eu:8080/CoPhiLabPortal/web/guest/home	Data analysis	X				X	
http://claviusontheweb.it/	VRE		X				X
http://www.italianlp.it/demo/linguistic-annotation-tool/	NLP			X			
http://www.corpusitaliano.it/preview/en/index.html	Content, repository		X				
http://www.italianlp.it/demo/read-it/	NLP			X			

[32] Service is defined according to D5.2 DARIAH (in kind) contribution tool as (Data) hosting service, (processing) service, support service and access to resources.
[33] Maintained or applicable is to understood as qualitative criteria in this regard. A service was recognized as maintained or applicable only if the hosting website appeared to be up to date.

Service[34]	Tags	Entity-based search	Scholarly content management	Text-analytic services	Visualisation	Other category (not covered by DoW)	Maintained or applicable[35]
http://www.italianlp.it/demo/t2k-text-to-knowledge/	NLP			X			
http://registries.dyas-net.gr/en	Content		X				X
http://be.dariah.eu/project/database-byzantine-book-epigrams-dbbe	Content, database		X				X
http://be.dariah.eu/project/database-southern-dutch-dialects-dsdd	Content, database		X				(X)
https://www.immibel.arch.be/virtualresearchenvironment	VRE					X	X
http://www.mamluk.ugent.be/Prosopography	Content, database		X				X
http://be.dariah.eu/project/ramses-online	Content, database, corpus		X				X
http://www.tic.ugent.be/	VRE			X		X	X
http://be.dariah.eu/project/dariah-vl-virtual-research-environment-service-infrastructure-vre-si	VRE					X	
https://transkribus.eu/Transkribus/	Data analysis			X		X	X
http://www.digital-humanities.at/de/dha/s-project/read-recognition-and-enrichment-archival-documents	VRE			X			
http://www.digital-humanities.at/de/dha/s-project/medcon	VRE, network analysis			X			
http://www.digital-humanities.at/de/dha/s-project/mapping-historical-networks-apis	VRE, content			X	X		
http://www.digital-humanities.at/de/dha/s-project/gams	Content, VRE		X				

[34] Service is defined according to D5.2 DARIAH (in kind) contribution tool as (Data) hosting service, (processing) service, support service and access to resources.
[35] Maintained or applicable is to understood as qualitative criteria in this regard. A service was recognized as maintained or applicable only if the hosting website appeared to be up to date.

Service[36]	Tags	Entity-based search	Scholarly content management	Text-analytic services	Visualisation	Other category (not covered by DoW)	Maintained or applicable[37]
http://www.digital-humanities.at/de/dha/s-project/dixit	Content, Teaching		X				
http://www.digital-humanities.at/de/dha/s-project/puzzle-4d	Content, Visualisation		X		X		
SUMS: 82	110	8	37	21	16	27	46

It is important to keep in mind, that the inventory is fragmentary. The reasons for have been explained above and are related to the situation in DARIAH that in 2017 - the time of writing - the marketplace is not yet available. There is no comprehensive overview on the DARIAH landscape. For this reason, the fragmentary inventory may be better understood as indication for specific competencies or centers of gravity. The large presence of scholarly content-related services may indicate such a center of gravity. The analysis of the figures follows below.

The table below is an aggregated facet of the inventory above. It groups all 110 services into three categories: infrastructure, research, and communication. The finding is – not surprisingly – that most DARIAH activities can be related more or less directly to research services.

Table 6: DARIAH inventory according to three aggregated categories

Service category	Quantity	Infrastructure	Research	Communication
AAI	2	2		
Authority files	1	1		
Bibliography	1		1	
Cloud/ File sharing	1	1		
Content (tutorial, teaching, blog)	27		27	
Data analysis	8		8	

[36] Service is defined according to D5.2 DARIAH (in kind) contribution tool as (Data) hosting service, (processing) service, support service and access to resources.

[37] Maintained or applicable is to understood as qualitative criteria in this regard. A service was recognized as maintained or applicable only if the hosting website appeared to be up to date.

Service category	Quantity	Infrastructure	Research	Communication
Database	7	7		
Developer tool	1	1		
Etherpad	1			1
Geospatial visualisation	2		2	
Helpdesk	1	1		
List server	1			1
Metadata conversion	4		4	
Monitoring	1	1		
NLP tool	5		5	
PID service	2	2		
Project management	2			2
Public calendar	1			1
Research data or publication repository	12		12	
Search	1	1		
Survey provisioning	1	1		
Thesaurus management	2		2	
Virtual machines/ web hosting	2	2		
Virtual research environment	13		13	
Visualisation	6		6	
Wiki	1			1
Corpus	1		1	
Knowledge Management System	1		1	

Service category	Quantity	Infrastructure	Research	Communication
Linked Data	1		1	
Network analysis	1		1	
SUM	110	20	84	6

Table 7: DARIAH inventory according to four DARIAH Virtual Competence Centers[38]

Service category	Quantity	*VCC1: e-Infrastructure*	*VCC2: Research and Education Liaison*	*VCC3: Scholarly Content Management*	*VCC4: Advocacy, Impact, Outreach*
AAI	2	2			
Authority files	1	1			
Bibliography	1			1	
Cloud/ File sharing	1	1			
Content (tutorial, teaching, blog)	27			27	
Data analysis	8	8			
Database	7	7			
Developer tool	1	1			
Etherpad	1	1			
Geospatial visualisation	2	2			
Helpdesk	1	1			
List server	1	1			
Metadata conversion	4	4			
Monitoring	1	1			
NLP tool	5	5			
PID service	2	2			
Project management	2	2			
Public calendar	1				1

[38] https://www.dariah.eu/activities/competence-centers-list/

Service category	Quantity	*VCC1: e-Infrastructure*	*VCC2: Research and Education Liaison*	*VCC3: Scholarly Content Management*	*VCC4: Advocacy, Impact, Outreach*
Research data or publication repository[39]	12			12	12
Search	1	1			
Survey provisioning	1	1			
Thesaurus management	2			2	
Virtual machines/ web hosting	2	2			
Virtual research environment	13	13			
Visualisation	6	6			
Wiki	1	1			
Corpus	1			1	
Knowledge Management System	1	1			
Linked Data	1	1			
Network analysis	1	1			
SUM	110	*66*	*0*	*43*	*13*

4 Interpretation

4.1 Methodology

Before an interpretation of the above standing findings is possible, the concept of the gap analysis has to be defined according to the present context.

First of all we don't conduct a gap analysis in the conventional sense, known from the economic sphere. In this sense a gap analysis is aimed at identifying sales or revenue gaps in a certain market or for certain products or services. The gap analysis takes into account the resources of the economic undertaking, usually a company, and relates them to the market segment in terms of customers, purchasing power or other factors.

A gap analysis in our context can be understood in the verbatim sense. The DARIAH landscape is mapped with focus on services, tools and resources and in the next step browsed for gaps. Correctly a gap can only be identified with consideration of the user requirements as they form the pattern which has to be filled or addressed with services and

[39] Repositories are listed in two VCCs.

resources. Clearly this hasn't been done in this document. It would have been impossible to gather the user requirements of the current DARIAH user base. This could be subject to a larger undertaking, which is not possible and not within the scope of this work package.
So one has to consider, that the aforementioned inventory of DARIAH's services and resources is fragmentary and represents a snapshot of the situation in autumn 2017. It has been compiled with regard to the requirements of the DESIR project, therefore focussing on four categories - entity-based search, text-analytic services, scholarly content management and visualisation. The inventory indicates centres of gravity of DARIAH in terms of services and resources. Areas with a lesser accumulation of services – for instance entity-based search – can be identified but one cannot inevitably deduct a gap from this. The gaps will be later defined in another, more indirect manner: with look at the DARIAH inventory the technology partners in WP4 draft ideas for possible innovative services and resources. As long as the idea for a service meets a potential demand in the DARIAH community and is coincidentally innovative and functional it addresses a gap in the sense of this work package. The analysis of the current DARIAH service portfolio leads to the following conclusions, which may be later used as basis for discussion of the conceptual models.

4.2 Interpretation of findings

DARIAH's services and resources concentrate on research related services. To a lesser degree services may be associated to infrastructure and communication (compare figure 4 below). These three categories come as an alternative facet to the four categories from DESIR's description of work. Although the gain of knowledge may be minor the result underpins the orientation of DARIAH as research infrastructure.
Interestingly the communication related share of services is the smallest one, which may be based on the already sufficient availability of (free or commercial) communication tools. It is clear that research infrastructure initiatives can't compete with the established and widely-used tools as Skype, forums, messengers, e-mail.
A different facet becomes possible when arranging the services along the DARIAH VCCs. The DARIAH Virtual Competency Centers[40] function as an umbrella structure to coordinate and promote DARIAH activities along four headlines: VCC1 -e-Infrastructure; VCC2-Research and Education Liaison; VCC3 -Scholarly Content Management Advocacy and VCC4 -Impact and Outreach. These categories differ a little bit from the infrastructural or technological view of the DESIR DoW but nevertheless the focus of DARIAH as research infrastructure becomes clear with 43 services relevant for scholarly content management (VCC3) and 66 services relevant for e-Infrastructure (VCC1).

[40] https://www.dariah.eu/activities/competence-centers-list/

Figure 4: Services according to the categories research, infrastructure and communication

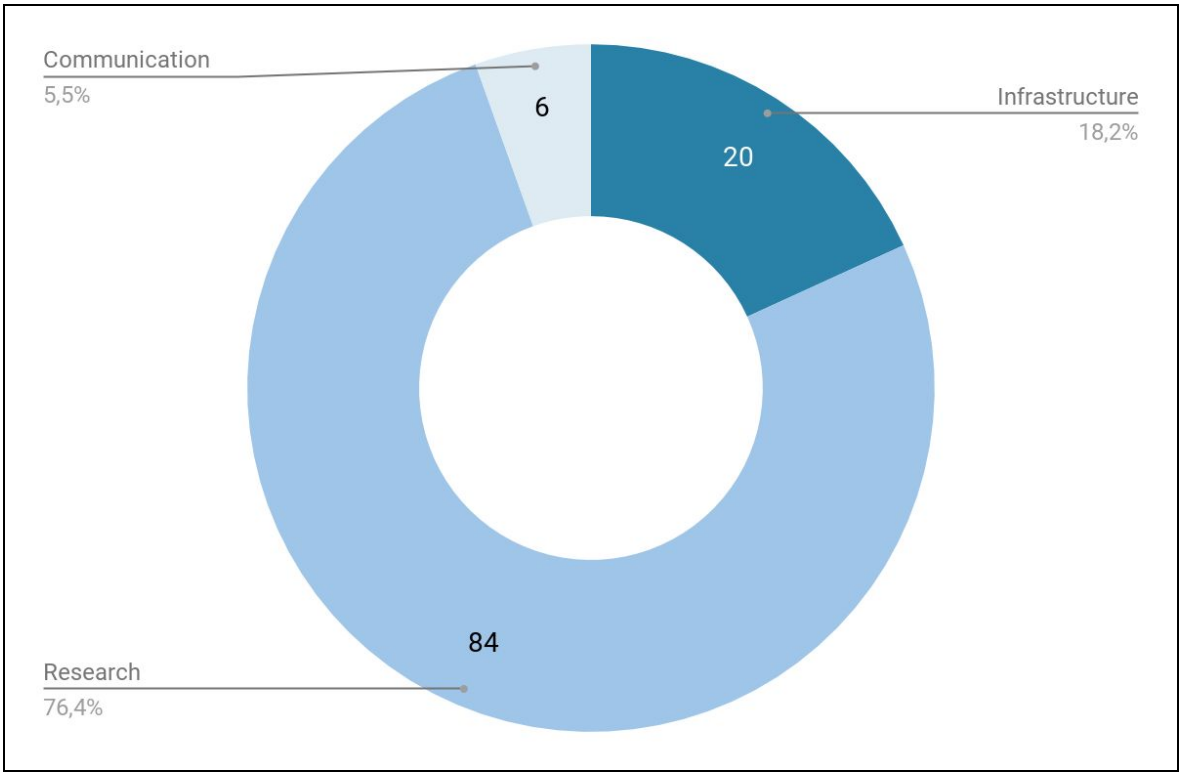
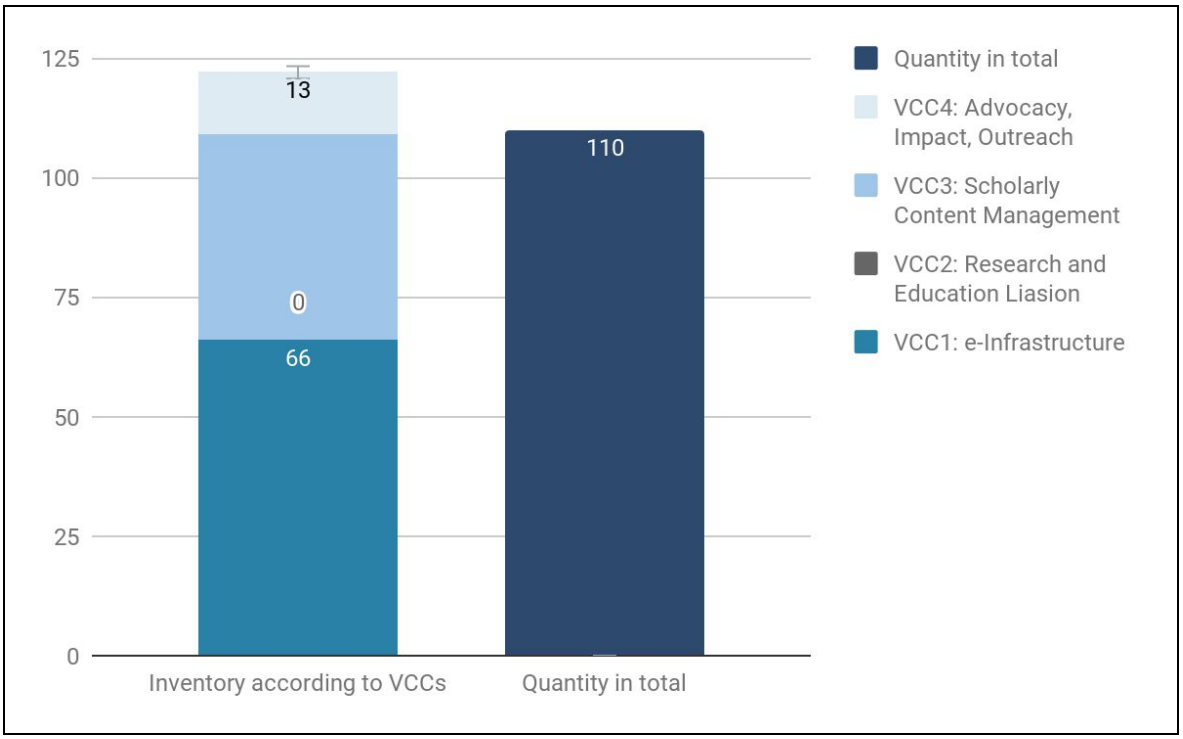
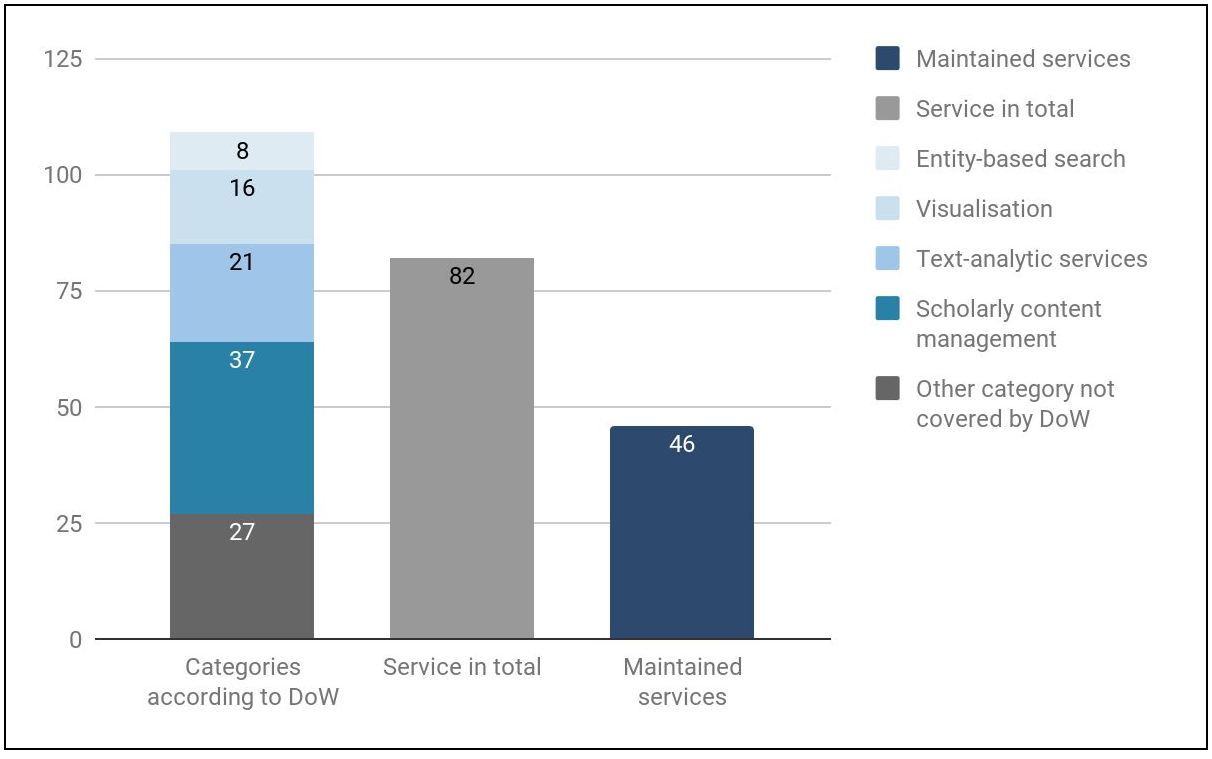


Figure 5: DARIAH inventory according to four DARIAH Virtual Competence Centers



A different picture emerges with look at the four DESIR DoW categories: most services are related to scholarly content management (37) and text-analytic services (21) and to a lesser degree to visualisation (16) and entity-based search (8). This may indicate that the DARIAH portfolio has the broadest portfolio in the area of scholarly content management and to a lesser degree in the field of entity-based search. It is likely that the conceptual models that will be developed in the course of the work package may reflect this distribution in one way or another. One can deduct from this that the number of only 8 entity-based search-services represent a gap. Again: a real gap could only be identified with regard to the user requirements. As long as the user requirements are met by the available services, this may be sufficient and there would be no need for developing further services. This can't be claimed with the available material. Another claim, which is more sustainable, could be to assume more possible connecting points with the available 37 services in scholarly content management or 21 services in text-analytic services. Therefore one could justify conceptual models in these areas as more connections seem possible.

Figure 6: Service[41] categories according to Description of Work

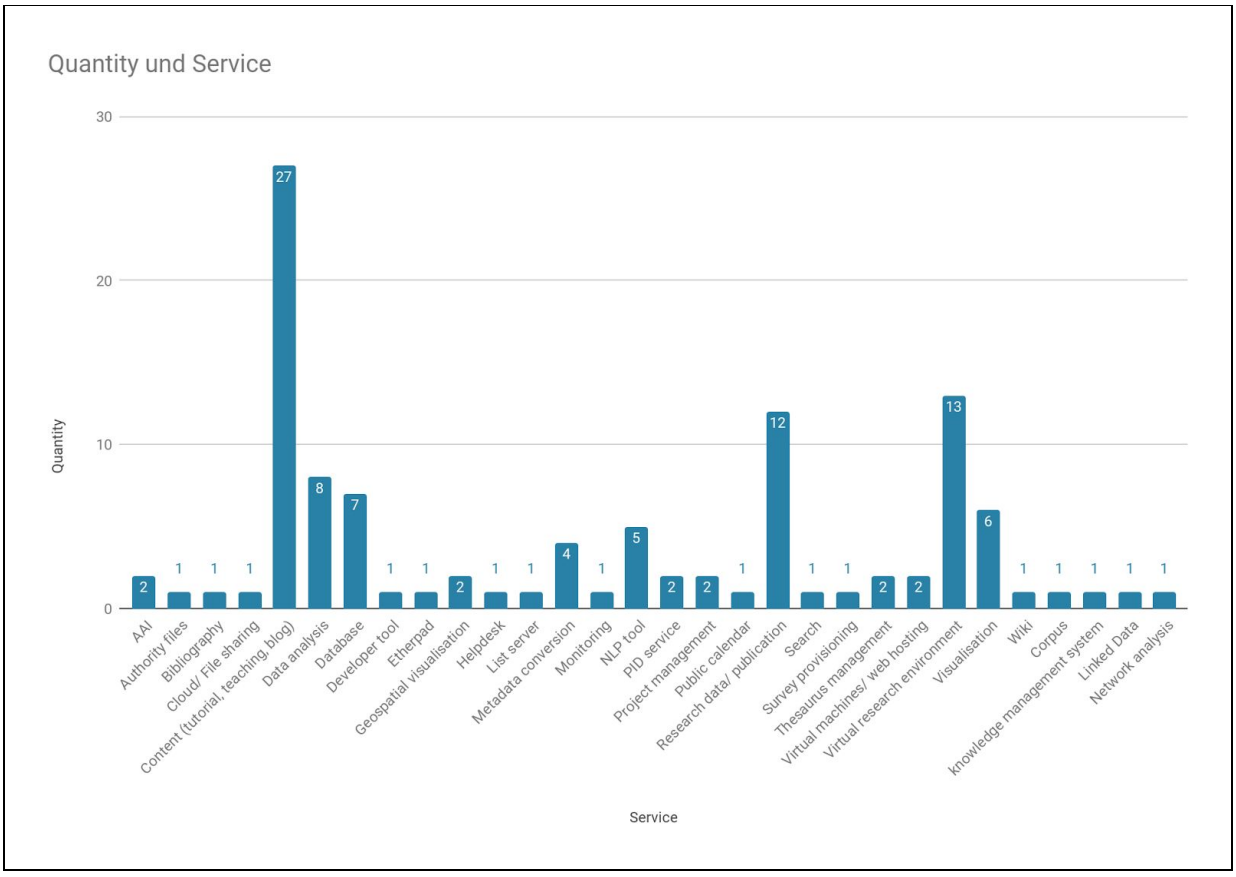


It may also be worthwhile to note that some services (27) could not be associated with any of the categories. This possibly bears the potential for conceptual models beyond the four DESIR categories, something which is not excluded by the project's description of work.

[41] Please note that the total of 82 services partly apply to more than one category, therefore a total of 110 is counted.

This grouping of services disguises the broad range of service forms (30) - illustrated in figure 7 - that have been found. DARIAH as research infrastructure offers a wide array of services, which makes it difficult to identify large gaps – as it will be explained later. Figure 7 clearly indicates a focus on content related services (27) such as blogs, repositories, tutorials or teaching material, but one has to consider that only a subset of them may relate to scholarly content as required in the description of work. The second largest group (13) is constituted by virtual research environments and research data repositories, whereby the term virtual research environment is quite broad. Aside from data analysis (8), visualisation tools (6), databases (7), NLP tools (5) or Metadata conversion (4) the other tag categories are represented a lot less.

Figure 7: Frequency of services in DARIAH



5 Conclusions

What implications for the work package and particularly the conceptual models can be drawn from the gap analysis?

As a reminder: the gap analysis is tasked with identifying gaps in the current DARIAH service landscape and shall deduct from the findings ideas for conceptual models that may

contribute to the DARIAH service portfolio and address specific demands from the user base. We have to admit that within the current situation of DARIAH - no central registry available until the DARIAH Marketplace will be launched in 2018 and a quite motley landscape of national DARIAH-related websites - and the resources at hand in this WP the results have to have a fragmentary character.

Also the broad range of services among the DARIAH partners - this is a finding of the gap analysis - makes it very difficult to identify major gaps, at least with the possible methodology for this report. A gap analysis should map the available services on the complete user requirements, something which is not possible within this DESIR work package.

But the inventory and its analysis allow to identify centres of gravity in DARIAH in relation to the other service categories. We have found that a bulk of the DARIAH services focus on scholarly content management and text-analytic services. But this may not be a surprising finding for a humanities-scoped research infrastructure. In both of these categories the services concentrate on content-related services like repositories, tutorials, blogs or other forms of content and on virtual research environments and data analysis tools.

The inventory also indicates areas where DARIAH is only represented with lesser services and resources, a finding that could be understood as a gap but doesn't have to be. It could also be interpreted the other way round: areas with considerably more services and resources offer more potential connection points for new ideas. This is true for entity-based search- and visualisation-related services. And it may also be true, that an area with only a few services available is already covered in a quite satisfying way for the users, so that there is no additional demand. But this can't be supported by our facts, actually.

A finding true for all services is that the development of services is to a large part research oriented and not intended to be offered on terms of a sustainable provision and as a service for the community. This is reflected by the column in table 5, which indicates that 46 out of 82 seem to be maintained or currently available. But this finding is on relatively unsure footing and would deserve a further analysis. The evaluation of services according to their maintenance and availability status is on the feature list for the DARIAH Marketplace as this is an important function for the users.

With look at the following steps in work package 4 Technology, the gap analysis suggests that for each category a few ideas should be drafted, especially paying attention to scholarly content-management and text-analytic services. The gap analysis indicates no real gap in an infrastructural sense, which is a good finding, but may not be helpful in pointing a direction for the conceptual models. DESIR takes into account the current DARIAH service landscape and intends to improve the available approaches. For this end, it will be important to relate the partners' competencies to DARIAH, which is already being underway with involving the

technology partners from the very beginning in this work package. This involvement is reflected in this document with chapter 2, describing the technology partners' resources and competencies.

6 Register of tables and figures

Tables

Figures

7 Sources

Anne Baillot, Mike Mertens and Laurent Romary (2016): Data Fluidity in DARIAH – Pushing the Agenda Forward. In: DARIAH-DE – Digitalität in den Geistes- und Kulturwissenschaften am Beispiel der digitalen Forschungsinfrastruktur DARIAH-DE. Bibliothek Forschung und Praxis (2016), 40(2): pp. 151-164. DOI http://dx.doi.org/10.1515/bfp-2016-0039

HUMANITIES AT SCALE, Dirk Roorda (2017): DARIAH Contribution Tool. Development instance provisionally available under: https://dariah-beta.dans.knaw.nl/data/contrib/filter, technical documentation under: https://dans-labs.github.io/dariah/ and user documentation under: https://github.com/Dans-labs/dariah/wiki/Help (Project Deliverable)

HUMANITIES AT SCALE, Lisa de Leeuw, Dirk Roorda (2017): D5.2 - DARIAH Space for inkind Contributions. (Confidential only for members of the consortium). (Project Deliverable)

Frank Fischer, Jennifer Edmond (2017): Towards the DARIAH Marketplace - Position Paper. http://conferences.au.dk/dariah2017/program/ (not a published version available at time of writing)

PARTHENOS, Nicola Aloia, Leonardo Candela, Franca Debole, Luca Frosini Matteo Lorenzini Pasquale Pagano (2017): D5.2 - Design of the Joint Resource Registry. URL: http://www.parthenos-project.eu/Download/Deliverables/D5.2_Report_on_design_Joint_Resource_Registry.pdf (Project Deliverable)

Tibor Kálmán, Xi Kong and Ulrich Schwardmann (2016): Die digitale Forschungsinfrastruktur DARIAH-DE: Angebotspalette für die Geistes- und Kulturwissenschaften. In: DARIAH-DE – Digitalität in den Geistes- und Kulturwissenschaften am Beispiel der digitalen Forschungsinfrastruktur DARIAH-DE. Bibliothek Forschung und Praxis (2016), 40(2): pp. 234-243. DOI https://doi.org/10.1515/bfp-2016-0041