INDIGO-DataCloud

INITIAL REQUIREMENTS FROM RESEARCH COMMUNITIES ANNEX 1.*PX*: SELECTED CASE STUDY FROM *DARIAH-DIGITAL RESEARCH INFRASTRUCTURE FOR THE ARTS AND HUMANITIES*

INPUT TO EU DELIVERABLE: D 2.1

Document identifier:	INDIGO-WP2-D2.1-ANNEX-1P0-V7
Date:	**27/05/2015**
Activity:	**WP2**
Lead Partner:	**EGI.eu**
Document Status:	**DRAFT**
Dissemination Level:	**CONFIDENTIAL (INTERNAL)**
Document Link:	

Abstract

This report summarizes the findings of T2.1 and T2.2 **for partner Px** along the first three months of the project. It is an integrated document including a general description of the research communities involved and the selected Case Studies proposed, in order to prepare deliverable D2.1, where the requirements captured will be prioritized and grouped by technical areas (Cloud, HPC, Grid, Data management) etc. The report includes an analysis of DMP (Data Management Plans) and data lifecycle documentation aiming to identify synergies and gaps among different communities.

I. COPYRIGHT NOTICE

Copyright © Members of the INDIGO-DataCloud Collaboration, 2015-2018.

II. DELIVERY SLIP

	Name	Partner/Activity	Date
From	<<**The author/editor in Px**>>	**Px**/WP2	
Reviewed by	**Moderators:** P.Solagna, F.Aguilar, J.Marco **Internal Reviewers:** <<To be completed by project office on submission to PMB>>		
Approved by	**PMB** <<To be completed by project office (no submission)>>		

III. DOCUMENT LOG

Issue	Date	Comment	Author/Partner
1	5-may-2015	First draft, v01	J.Marco, F.Aguilar CSIC
2	7-may-2015	Initial feedback on structure from all partners	F.Aguilar CSIC, A.Bonvin UUtrecht
3	18-may-2015	Draft discussed in f2f meeting in Lisbon	P.Solagna, EGI.eu F.Aguilar, CSIC
4-7	28-may-2015	Draft ready for initial community input, to be iterated with JRA, v07	P.Solagna, EGI.eu J.Marco, F.Aguilar, CSIC, I.Blanquer UPV
8	4-june-2015	Draft after input from community, v08	JRA?
9	7-june-2015	Draft revised also with JRA, v09	P.Solagna, EGI.eu F.Aguilar, CSIC
10	10-june-2015	Draft to be circulated for internal review, v10	P.Solagna, EGI.eu
11	20-june-2015	Comments included, version for release v11	P.Solagna, EGI.eu

TABLE OF CONTENTS

0 INTRODUCTION AND CONVENTIONS

PLEASE, READ CAREFULLY BEFORE COMPLETING THE ANNEX:

*This Annex is an example of compilation of the information needed to support adequately a **Case Study** of interest in a Research Community. Each partner in INDIGO WP2 is expected to provide such information along the first three months of the project (i.e. by June 2015), and it will be used to compile Deliverable D2.1 on Initial Requirements from Research Communities.*

There will be around 10 Annexes, for example Annex 1.P1 for partner 1 in WP2 (i.e. UPV), will cover Case Studies from EuroBioImaging research community.

The initial version will be discussed with INDIGO Architectural team to agree on a list of requirements.

Some relevant definitions:

***A Case Study** is an implementation of a research method involving an up-close, in-depth, and detailed examination of a subject of study (the case), as well as its related contextual conditions.*

We should focus on Case Studies that are representative both of the research challenge and complexity but also of the possibilities offered by INDIGO-DataCloud solutions on it!

*The Case Study will be based on a set of User Stories, i.e. how the researcher describes the steps to solve each part of the problem addressed. **User Stories** are the starting point of **Use Cases**, where they are transformed into a description using software engineering terms (like the actors, scenario, preconditions, etc). Use Cases are useful to capture the Requirements that will be handled by the INDIGO software developed in JRA workpackages, and tracked by the Backlog system from the OpenProject tool.*

The User Stories are built by interacting with the users, and a good way is to do it in three steps (CCC): Card, Conversation and Confirmation[1].

Use Cases can benefit from tools like "mock-up" systems where the user can describe virtually the set of actions that implement the User Story (i.e. by clicking or similar on a graphical tool).

Different parts of this document should be completed with the help/input of different people:

RESEARCH MANAGERS

-Section 1, SUMMARY, is to be reviewed/agreed with them as much as possible

RESEARCHERS

*-Section 2, INTRODUCTION is designed to be filled with direct input from (senior) researchers describing the interest of the application, and written in such a way that it can be included in related technical papers. It is likely that such introduction is already available for some communities (for example, for several research communities in WP2 like DARIAH, CTA,EMSO, Structural Biology, one may start from the **Compendium of e-Infrastructure requirements for the digital ERA[2] from EGI***

APPLICATION DEVELOPERS AND INTEGRATORS WITHIN THE RESEARCH COMMUNITIES

-Sections 3, 4, 5, 6: should be discussed from their technical point of view (including data management as much as possible).

MIDDLEWARE DEVELOPERS AND E-INFRASTRUCTURE MANAGERS

-Sections 7, 8: should be discussed with them

[1] For a nice intro, see: https://whyarerequirementssohard.wordpress.com/2013/10/08/when-to-use-user-stories-use-cases-and-ieee-830-part-1/ , and also https://whyarerequirementssohard.wordpress.com/2015/02/12/how-do-we-write-good-user-stories/ etc.

[2] *https://documents.egi.eu/public/ShowDocument?docid=2480*

The logical order to fill the sections is: 2,3,4,5,6,1,7,8. Sections 1 and 8 will go into deliverable D2.1.

Other conventions and instructions for this document:

As this document/template is to be reused, the convention to use it as a questionnaire is that:

1) -text in italics provides its structure and questions,

2) -input/content should be written using normal text, replacing **<input here>**

Also the following conventions are used to identify the purpose of some parts of the questionnaire:

Bold text in blue corresponds to indications/suggestions to complete the questionnaire

Bold text in dark red marks technical issues particularly relevant that should be carefully considered for further analysis of requirements

Text in red indicates pending issues or ad-hoc warnings to the reader

1 EXECUTIVE SUMMARY ON THE CASE STUDY

Summarize the research community applications/plans/priorities (max length 2 pages).
To be completed after section 2 and reviewed later. Supervision by a senior researcher is required.

1.1 Identification

- *Community Name*: **Digital Research Infrastructure for the Arts and Humanities (DARIAH)**
- *Institution/partner representing the community in INDIGO:* **Institute Ruder Boskovic (IRB)**
- *Main contact person*:
- *Contact email:*
- *Specific Title for the Case Study:* **Strengthening the Use of Scientific Distributed Computing in the Arts and Humanities**

1.2 Brief description of the Case Study and associated research challenge

Please include also a brief description of the community regarding this Case Study: partners collaborating, legal framework, related projects, etc.

Describe the research/scientific challenge that the community is addressing in the Case Study

Digital research methods have recently started to enter the mainstream of humanities, arts and social sciences research. Digital humanities have existed for years as a specialised field but the recent growth in the number of centres and research projects associated with digital methods in arts and humanities (A+H) and social sciences indicate that we are at a fundamental shift. The digital arts and humanities are at a critical point in the transition from a specialty area to a full-fledged community with a common set of methods, sources of evidence and infrastructure. All of these are necessary for achieving academic and data driven scientific recognition. Information and data-intensive, distributed, collaborative and multidisciplinary research is now the norm in many scientific areas, but they are still in an experimental phase in the arts and humanities research community. However, the art and humanities disciplines nowadays generate and analyse an increasing amount of data and show great potential for growth and evolvement of new technologies. Research process in the A+H become more and more data-intensive and therefore have to be supported by emerging research infrastructures. Also, a vast array of collaborations arises in the digital humanities across Europe in the form of spontaneously funded research networks and associations. What is lacking, however, is an infrastructure that would ensure that the state-of-the-art of these collaborations is preserved and integrated, and that common best practices and methodological and technological standards are followed.

1.3 Expectations in the framework of the INDIGO-DataCloud project

What do you think could be your main objectives to be achieved within the INDIGO project in relation to this Case Study?

The digital research infrastructures for the humanities can only be effectively established in the context of close interaction and connections between the areas of instruction, research, research data, and basis infrastructures, and they can only be successful when the entire spectrum of arts and humanities disciplines are involved. The A+H disciplines cover a wide range of very different research initiatives, with diverse requirements and different amounts and types of data used. However, the one thing that all these different research initiatives have in common is that they rely on easily accessible and reliable long-term data storage. The DARIAH community hopes that the collaboration with the INDIGO-DataCloud project will result in strengthening the utilization of Grid and Cloud resources for the research needs that arise in digital arts and humanities.

1.4 Expected results and derived impact

Describe the research results and impact associated to this Case Study.

Enhancement of digitally-enabled research across the digital art and humanities community, wherefore a sustainable, distributed research infrastructure is built and maintained. An essential component of the infrastructure is a long-term storage service serving a wide variety of A+H disciplines and accounting for their special requirements.

1.5 References useful to understand the Case Study

Include previous reports, articles, and also presentations describing the Case Study

[1] *Compendium of e-Infrastructure requirements for the digital ERA*, 2015

[2] *Report to the Technical Advisory Board (TAB) DARIAH-DE and CLARIN-D*, DARIAH-DE – Digital Research Infrastructure for the Arts and Humanities, March 28th, 2013

[3] Tonne, D., J. Rybicki, S.E. Funk, P. Gietz (2013) *Access to the DARIAH Bit Preservation Service for Humanities Research Data*. In: 21st Euromicro International Conference on Parallel, Distributed and Network-Based Processing pp:9-15. IEEE.

[4] Blanke, T., M. Bryant, M. Hedges, A. Aschenbrenner, M. Priddy (2011) *Preparing DARIAH*. In: IEEE 7th International Conference on E-Science pp:158-165. IEEE.

2 INTRODUCTION TO THE RESEARCH CASE STUDY

Summarize the Case Study from the point of view of the researchers (max length 3 pages + table). Input by the research team in the community addressing the Case Study is required.

2.1 Presentation of the Case Study

Describe the Case Study from the research point of view

In the DARIAH community, one of the biggest challenges for the near future is linked to the concept of big data. The arts and humanities have seen an exponential growth in digital research material, especially in the last decade, as a result of new born-digital material or large digitisation efforts in the EU and elsewhere. The biggest current field of research is to define the new digital methodologies to meet the requirements of humanities data that is particularly fuzzy and inconsistent, as it is not automatically produced, but is the result of human effort. Also, recently a lot of effort is being put to work towards more consistent cyber-infrastructure and away from ad hoc solutions with the aim of delivering more systematic investigations. To move beyond the state-of-the-art, DARIAH needs to achieve the integration of humanities research material on the grand scale. Therefore, the major challenge for DARIAH, also addressed within this Case Study, is to join up national/local knowledge in a sustainable, collaborative and lasting ecosystem. An essential component of the ecosystem is a long-term storage service serving a wide variety of disciplines and accounting for their special requirements.

2.2 Description of the research community including the different roles

Please include a description of the scientific and technical profiles, and detail their institutions

Describe the research community specifically involved in this Case Study

DARIAH is a social and technical infrastructure which is composed of people, expertise, information, knowledge, content, methods, tools and technologies for investigating, exploring and supporting work across the broad spectrum of the digital arts and humanities. Initially funded under the ESFRI programme, DARIAH comprises a number of national initiatives and it also aims to help other EU countries establish their own arts and humanities e-infrastructures, and to achieve new modes of collaboration between computing and humanities based on existing communities of practice.

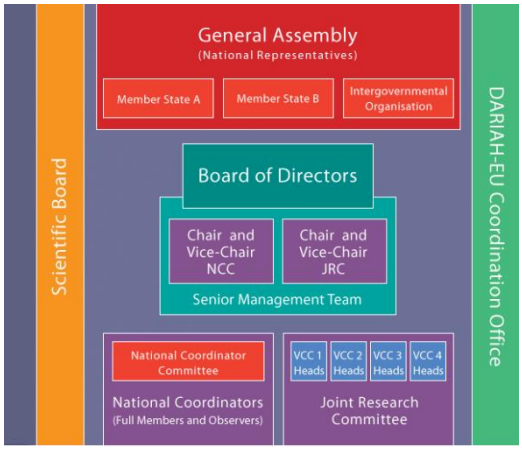


Figure 1: DARIAH organization

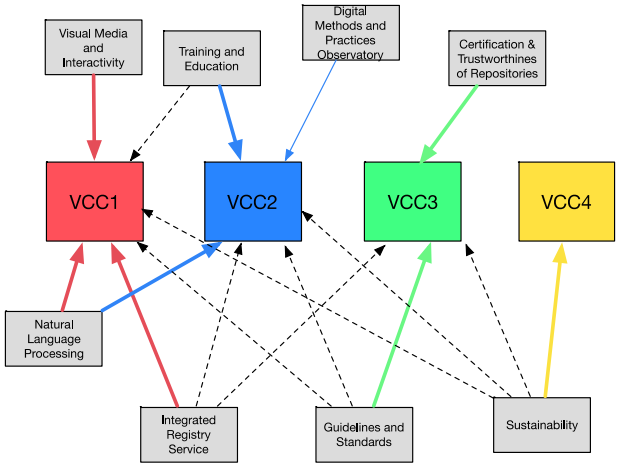


Figure 2: DARIAH's virtual competence centres

Internally, DARIAH is consists of more than 20 working groups organized around four virtual competency centres (VCCs) and a coordination office staffed and coordinated by its EU partners. The existing VCCs are: VCC1 e-infrastructure; VCC2 research and education; VCC3 content management; and VCC4 advocacy and outreach. While centred on a specific area of expertise, the VCCs are at the same time cross-disciplinary, multi-institutional and multi-national. DARIAH aims to provide its stakeholders a single point of contact in which a wide range of infrastructure and support services, which the VCCs can offer, are bundled into a smaller set of service packages targeting familiar and commonly requested activities. DARIAH partners are recognised contributors to national excellence in A+H research and successful collaborators in European research projects.

2.3 Current Status and Plan for this Case Study

Please indicate if the Case Study is already implemented or if it is at design phase.

Describe the status of the Case Study and its short/mid term evolution expected

DARIAH services are based on the existing national collaborations in digital arts and humanities. At the moment, DARIAH focuses on maximising the impact and collaboration across these kinds of national projects, but over time we see the unified DARIAH data and service market developing that will be enriched by tools and services from the digital arts and humanities community. DARIAH will guarantee this service market place and the services necessary to join up national services such as Persistent Identifiers (PIDs) and a federated search and authentication environment. DARIAH attempts to build services around communities, which can then be exchanged between communities in a virtual social marketplace that connects community workspaces with trusted DARIAH repositories of research data and services.

2.4 Identification of the KEY Scientific and Technological (S/T) requirements

Please try to identify what are the requirements that could make a difference on this Case Study (thanks to using INDIGO solutions in the future) and that are not solved by now.

Indicate which are the KEY S/T requirements from your point of view

In the future we need to work on a more decentralised approach that will in particular develop the social components of the e-Infrastructure. For DARIAH researchers, it would therefore be useful to have transnational access to virtual machines, data management services, persistent storage and instruments to investigate objects. This is next to the usual candidates of providing stable PIDs for resources and distributed authentication and authorisation. We already have these capacities in place in the various partner countries, but sharing these has proven to be challenging. Desirable for distinct research activities such as the analysis of manuscripts is the easy access to high performance computation infrastructure for the occasional burst in processing needs. Furthermore, a transparent data infrastructure that allows for the combination of many small scale but highly interrelated resources and is at the same time persistent across countries would be a great advantage. A polyglot persistent infrastructure that provides seamless access from localised web data storage all the way to long-term large digital archives would be one of our main future goals. Furthermore, as DARIAH is designed for the exchange of knowledge and services in a dedicated virtual social marketplace, we need open APIs to expose reusable services, as well as composition and aggregation facilities to work with these services

2.5 General description of e-Infrastructure use

Please indicate if the current solution is already using an e-Infrastructure (like GEANT, EGI, PRACE, EUDAT, a Cloud provider, etc.) and if so what middleware is used. If relevant, detail which centres support it and what level of resources are used (in terms of million-hours of CPU, Terabytes of storage, network bandwidth, etc.) from the point of view of the research community.

Detail e-Infrastructure resources being used or planned to be used.

DARIAH has just started it's collaboration with EGI within the EGI DARIAH Competence Center project. The EGI DARIAH CC is driven by the evidence of under-exploitation of Grid and Cloud-based e-Infrastructures in the research area of Social Sciences. The overall objective of the EGI DARIAH CC is to provide a wider and more efficient access to, and use of, research e-Infrastructures at European Grid Infrastructure (EGI) level, including transnational access, joint research and networking. EGI DARIAH CC will bring the network of the service providers, technology providers and application developer experts from EGI, to the experts in digital Arts and Humanities (A&H), bridging the gap between these groups. EGI DARIAH CC will support DARIAH-EU Virtual Competency Centres, especially VCC1 (e-Infrastructures) in porting services to make use of EGI resources. Empowering the services with national resources DARIAH-EU is able to extend the needed infrastructure to support European research programmes and projects in the Arts and Humanities

Also, within the Task “Enabling Users” of the GEANT3plus recently started the collaboration project “DARIAH integration with eduGAIN” with the aim to enhance the acceptance and deployment of eduGAIN within the humanities research infrastructures. Currently, the central components of the DARIAH AA infrastructure, as well as the services already supporting federated access, are operated by DARIAH-DE (German Branch of DARIAH) and subsequently exposed to eduGAIN via the German DFN-AAI identity federation.

There is currently a major review on-going within the DARIAH community on our future relationship with generic middleware work in the various national programmes. The review is carried out within the context of the German DARIAH -DE project and its second round of funding

with the aim to set up a generic e-Infrastructure Unit based on simple infrastructure needs. DARIAH-DE can in particular rely on the work done in the associated TextGrid consortium. This reflects the shift from an originally middleware-focussed (grid-based) e- Infrastructure (connecting to Globus middleware via GAT) towards a more computing centre-specific distributed storage solution. This approach will be replicated across the DARIAH communities. The set -up of grid nodes was too complicated for some of the noncomputing centre infrastructure partners and partly beyond their needs. The PKI based security has too high an overhead for the end -user (user certificates handling and reissue problems). With these lessons learned, DARIAH-DE favours simpler replication mechanisms (via iRODS) for distributed storage and user-friendlier Authentication and Authorisation mechanisms based on SAML. At the moment, DARIAH-DE is the leading national initiative regarding provision of hosting facilities, tools development and bit preservation. Resources offered within the DARIAH e-Infrastructure can be primarily used for either hosting services for the humanities or performing computations. In this regard, DARIAH-DE offers state-of-the-art solutions resulting from the long-year experiences of the computer and data centers participating in the project (Juelich Supercomputing Center, Rechenzentrum Garching, Gesellschaft für wissenschaftliche Datenverarbeitung mbh Göttingen, Karlsruhe Institute of Technology). It should be stressed at this point that the DARIAH Hosting Environment is an object of research itself, thus it is still in the pilot phase and is undergoing a number of changes. For storage purposes, DARIAH-DE developed the DARIAH Bit Preservation system which is providing a HTTP-based interface to storage resources, using a database to store basic metadata and relying on iRODS as its storage backend. [3]

2.6 Description of stakeholders and potential exploitation

Please summarize the potential stakeholders (public, private, international, etc.) and relate them with the exploitation possibilities. Provide also a realistic input to table on KPI.

Describe the exploitation plans related to this Case Study

The potential stakeholders that would benefit from establishing a sustainable DARIAH Repository are:

- Research groups across the arts and humanities
- Individual A+H researches
- Affiliated national/international projects in A+H
- A+H institutions
- Universities
- Cultural heritage sectors
- Publishers
- Funding agencies
- Goverment agencies

Please indicate (as realistic as possible) the expected impact for each topic in the following table:

Area	*Impact Description*	*KPI Values*

Access	*Increased access and usage of e-Infrastructures by scientific communities, simplifying the "embracing" of e-Science.*	• *Number of ESFRI or similar initiatives adopting advanced middleware solutions* ESFRIs: **<input here>** • *Number of production sites supporting the software* **<input here>**
Usability	*More direct access to state-of-the art resources, reduction of the learning curve. It should include analysis platforms like R-Studio, PROOF, and Octave/Matlab, Mathematica, or Web/Portal workflows like Galaxy.* *Use of virtualized GPU or interconnection (containers).* *Implementation of elastic scheduling on IaaS platforms.*	• *Number of production sites running INDIGO-based solutions to provide virtual access to GPUs or low latency interconnections* **<input here>** • *Number/List of production sites providing support for Cloud elastic scheduling* **<input here>** • *Number of popular applications used by the user communities directly integrated with the project products:* **<input here>** • *Number of research communities using the developed Science Gateway and Mobile Apps:* **<input here>** • *Research Communities external to INDIGO using the software products*: **<input here>**
Impact on Policy	*Policy impact depends on the successful generation and dissemination of relevant knowledge that can be used for policy formulation at the EU, or national level.*	• *Number of contributions to roadmaps, discussion papers:* **<input here>**
Visibility	*Visibility of the project among scientists, technology providers and resource managers at high level.*	• *Number of press releases issued*: **<input here>** • *Number of download of software from repository per year:* **<input here>** • *List of potential events/conferences/workshops:* **<input here>** • *Number of domain exhibitions attended* **<input here>** • *Number of communities and stakeholders contacted* **<input here>**
Knowledge Impact	*Knowledge impact creation: The impact on knowledge creation and dissemination of knowledge generated in the project depends on a high level of activity in dissemination to the proper groups.*	• *Number of journal publications:* **<input here>** • *Number of conference papers and presentations*: **<input here>**

Table 1 Key Performance Indicators (KPI) associated to different areas. Add in this table how your community would contribute to the KPIs. Note: this table will NOT be included in the deliverable.

3 TECHNICAL DESCRIPTION OF THE CASE STUDY

Describe the Case Study from the point of view of developers (4 pages max.)
Assemble it using preferably an AGILE scheme based on User Stories.

3.1 Case Study general description assembled from User Stories

Please describe here globally the Case Study. If possible use as input "generic" User Stories built according to the scheme: short-description (that fits in a "card") + longer description (after "conversation" with the research community). Provide links to presentations in different workshops describing the Case Study when available. Include schemes as necessary.

Describe the Case Study showing the different actors and the basic components (data, computing resources, network resources, workflow, etc.). Reference relevant documentation.

Considering the fact that the requirements of the DARIAH community regarding the INDIGO solutions are based on a universal need for accessible and reliable long-term data storage which would eventually lead to more sustainable and collaborative DARIAH ecosystem, the description of the Case Study in this section gives a more general overview of the complexity and variety of research areas within the DARIAH community, rather than a detailed technical description of one particular research case.

3.2 User categories and roles

Describe in more detail the different user categories in the Case Study and their roles, considering in particular potential issues (on authorization, identification, access, etc.)

User categories:

- A+H researches
  - from various disciplines, e.g. history, musicology, archaeology, art history, philosophy, literary studies, etc. (using different media types)
- Archivists and librarians
- IT experts

These user categories considered regarding their involvement in DARIAH:

- DARIAH members – have access to all services by using their DARIAH user login
- External researchers – can access services after requesting the DARIAH user login

3.3 General description of datasets/information used

List the main datasets and information services used (details will be provided in next section)

The data used in A+H research projects differs in size (a few kilobytes for a text file containing a letter or several gigabytes for a film record of an opera), quantity (a few image files of a rare and valuable manuscript up to several millions of image files of a whole library) and type as there is a variety of different formats for text, image, audio, and movies. The most important descriptive aspect and key

property of data handled within the DARIAH project, and the digital humanities in general, is heterogeneity. Humanities disciplines nowadays generate and analyze an increasing amount of data. At least parts of their research process therefore become more and more data-intensive and have to be supported by emerging research infrastructures.

3.4 Identification of the different Use Cases and related Services

Identify initial Use Cases based on User Stories, and describe related (central/distributed) Services

Some examples of different research projects in DARIAH which illustrate the diversity of use cases and heterogeneity of research data with their corresponding standards and metadata:

- A musicological project provides a complete overview of the work of one composer including scores, letters or recordings of an orchestra; need for viewing and editing MEI-encoded music documents in CMN (Common Music Notation).
    - MEI – Music Encoding Initiative: Standard for encoding music scores

- A scholar working in Jewish studies analyzes an old Jewish cemetery. He has to deal with inscriptions which have to be translated and access maps or chronicles from different decades based on an epigraphical database of Jewish cemeteries, funeral inscriptions and headstones using epigraphical metadata (information on the collection) and data (information on a single object).
    - EpiDoc – Epigraphic Documents: Standard for encoding epigraphic inscriptions in TEI XML

- A digitization project establishes a virtual library comprising of manuscripts that have been spread all over the world.
    - CIDOC-CRM – CIDOC Conceptual Reference Model: Ontology for cultural heritage data

- An archaeologist virtually reconstructs buildings from their remains. The data from results of the excavations will be used to create 3D models of the landscapes and buildings.
    - ADeX – Archaeological Data eXchange: Standard for the exchange of archaeological subject data

- A scholar analyses the historic development of narrative techniques based on a large collections of literary texts comprising about 2000 German novels, primarily from the 18th and 19th century.
    - TEI – Text Encoding Initiative: Standard for encoding textual data

All these different research initiatives have one thing in common: they rely on accessible, reliable long-term data storage.

3.5 Description of the Case Study in terms of Workflows

Summarize the different Workflows within the Case Study, and in particular Dataflows. Include the interaction between Services.

There are many the different aspects of dealing with data in the diverse spectrum of research projects within DARIAH such as data digitisation, digital annotation, data search and access, data analysis, data storage, etc. Therefore, different use cases have different workflows regarding their specific focus and aim of research.

From the perspective of an A+H researcher, access to research data is one of the most important aspects in digital research but often very difficult on a larger scale due to the diversity of digital data sources and the heterogeneity of the information they contain. The DARIAH federation infrastructure aims to address these problems by building a comprehensive framework of registries and generic services shown in Figure 3.

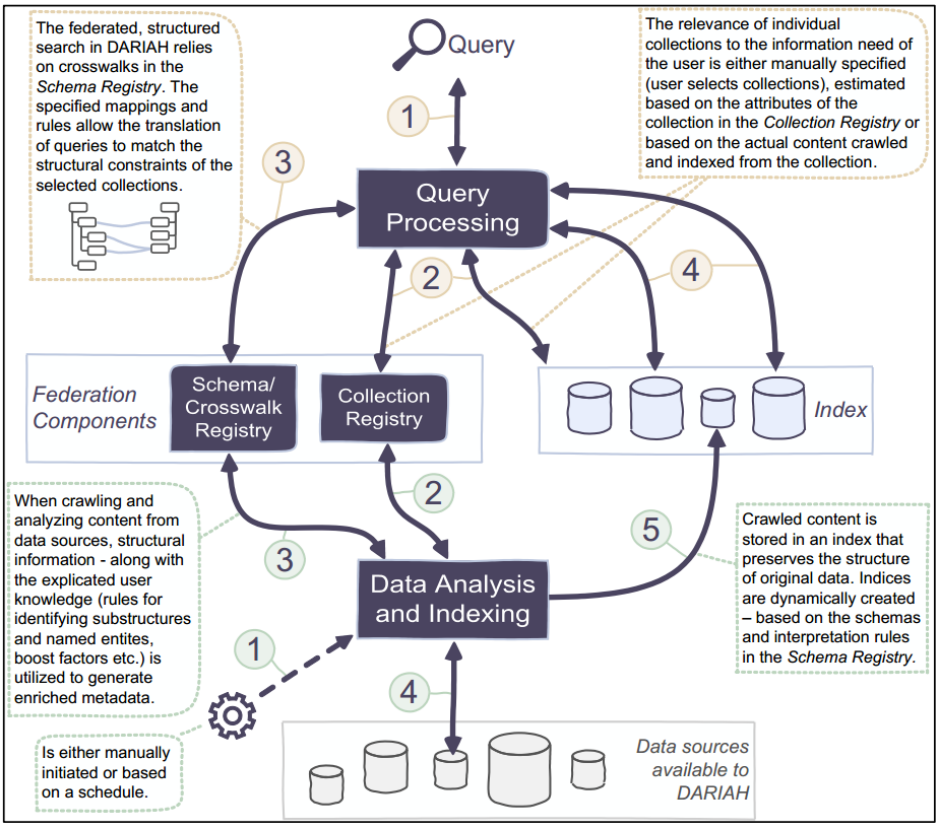


Figure 3: The DARIAH federation infrastructure

3.6 Deployment scenario and relevance of Network/Storage/HTC/HPC

Indicate the current deployment framework (cluster, Grid, Cloud, Supercomputer, public or private) and the relevance for the different Use Cases of the access to those resources.

<input here>

The following figures describe the current DARIAH storage architecture and potential integrations:

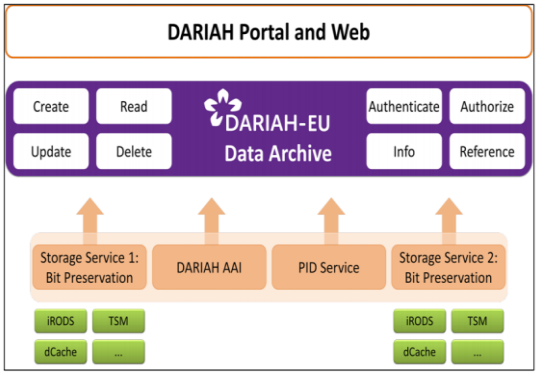


Figure4: The DARIAH Archive Service

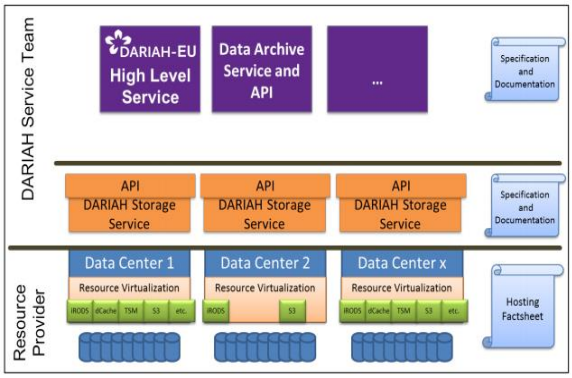


Figure5: Potential Integration

The DARIAH Bit Preservation, as a part of an archiving system for the arts and humanities, allows for a high performance, sustainable, and distributed storage of research data as the basis of virtual research environments. A RESTful API for the DARIAH Bit Preservation was developed which includes an administrative extension, and which is secured by an Authentication and Authorization Infrastructure (AAI) based on SAML. The implementation of the API offers distributed access by usage of the HTTP protocol and is able to handle a high number of files. Data transfer rates of up to 45 MB/s were achieved for uploading large files in the local network.

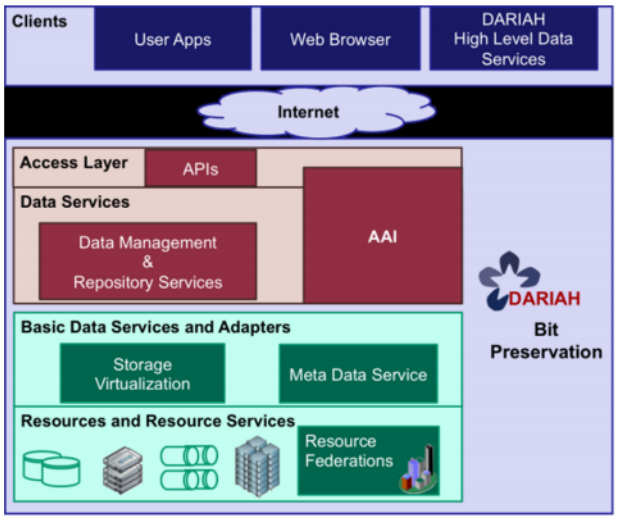


Figure 6: Architectural overview of the DARIAH Bit Preservation service

At the moment several projects coming from different disciplines and different funding use the DARIAH-DE storage service.

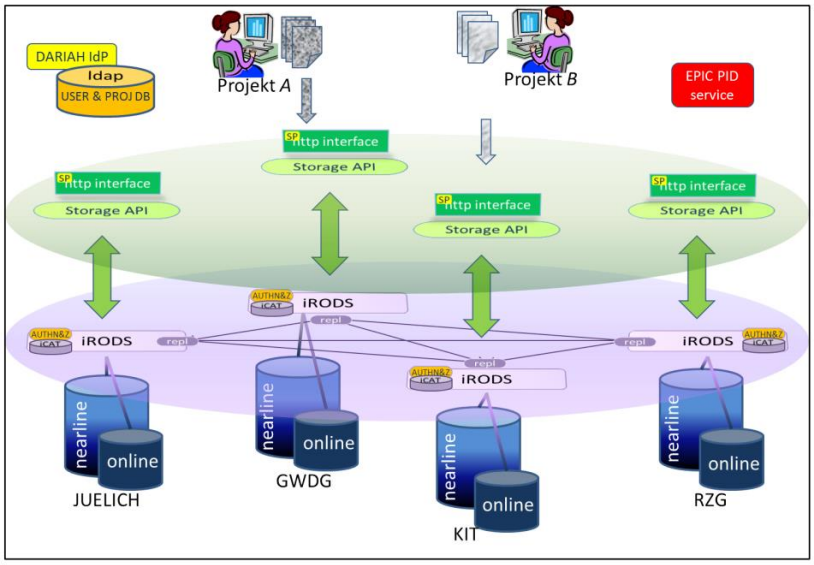
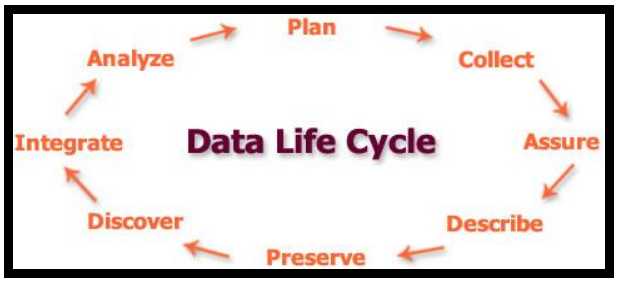


Figure7: DARIAH-DE Storage Federation Architecture

4 DATA LIFE CYCLE

INDIGO-DataCloud is a DATA oriented project. So the details provided in this complex section are KEY to the project. Please try to be as complete as possible with the relevant information.

Using the DataONE scheme, shown below, the different stages in the data life cycle are considered under the perspective of preparation of a DMP (Data Management Plan) following the recommendations of the UK DCC and H2020 guidelines.



BEFORE FILLING NEXT SECTIONS, CONSIDER CONSULTING:
https://www.dataone.org/all-best-practices-download-pdf and https://dmponline.dcc.ac.uk/

4.1 Data Management Plan (DMP) for this Case Study

According to EU H2020 indications[3], following UK DCC tool indications

4.1.1 Identification of the DMP

Plan identification: *<Code, ID>* **<input here>**

Associated grants: <Funded Projects, other grants> **<input here>**

Principal Researcher: **<input here>**

DMP Manager: **<input here>**

Description: **<input here>**

[3] *In Horizon 2020 a limited pilot action on open access to research data will be implemented. Projects participating in the Open Research Data Pilot will be required to develop a Data Management Plan (DMP), in which they will specify what data will be open. Other projects are invited to submit a Data Management Plan if relevant for their planned research. The DMP is not a fixed document; it evolves and gains more precision and substance during the lifespan of the project. The first version of the DMP is expected to be delivered within the first 6 months of the project. More elaborated versions of the DMP can be delivered at later stages of the project. The DMP would need to be updated at least by the mid-term and final review to fine-tune it to the data generated and the uses identified by the consortium since not all data or potential uses are clear from the start. The templates provided for each phase are based on the annexes provided in the Guidelines on Data Management in Horizon 2020 (v.1.0, 11 December 2013).*

4.1.2 DMP at initial stage (to be prepared before data collection)

The DMP should address the points below on a dataset by dataset basis and should reflect the current status of reflection within the consortium about the data that will be produced.

For each data set provide:

Description of the data that will be generated or collected; indicate its origin (in case it is collected), nature and scale and to whom it could be useful, and whether it underpins a scientific publication. Information on the existence (or not) of similar data and the possibilities for integration and reuse.

Data set reference and name **<input here>**

Data set description **<input here>**

Standards and metadata **<input here>**

Reference to existing suitable standards of the discipline. If these do not exist, an outline on how and what metadata will be created (see also below).

Connection to Instrumentation,

Sensors, Metadata, Calibration, etc (pending definitive form, see next sections)

<input here>

Vocabularies and Ontologies

Are they relevant? Internal vocabularies related to the specific fields. RDA groups.

(pending definitive form, see next sections)

<input here>

Data Capture Methods

Outline how the data will be collected / generated and which community data standards (if any) will be used at this stage. Indicate how the data will be organised during the project, mentioning for example naming conventions, version control and folder structures. Consistent, well-ordered research data will be easier for the research team to find, understand and reuse.

- *How will the data be created?* **<input here>**
- *What standards or methodologies will you use?* **<input here>**
- *How will you structure and name your folders and files?* **<input here>**
- *How will you ensure that different versions of a dataset are easily identifiable?* **<input here>**

Metadata

Metadata should be created to describe the data and aid discovery. Consider how you will capture this information and where it will be recorded e.g. in a database with links to each item, in a 'readme' text file, in file headers etc. Researchers are strongly encouraged to use community standards to describe and structure data, where these are in place. The UK Data Curation Center offers a catalogue of disciplinary metadata standards.

- *How will you capture / create the metadata?* **<input here>**

• *Can any of this information be created automatically?* **<input here>**
• *What metadata standards will you use and why?* **<input here>**

Data sharing

Description of how data will be shared, including access procedures, embargo periods (if any), outlines of technical mechanisms for dissemination and necessary software and other tools for enabling re-use, and definition of whether access will be widely open or restricted to specific groups. Identification of the repository where data will be stored, if already existing and identified, indicating in particular the type of repository (institutional, standard repository for the discipline, etc.). In case the dataset cannot be shared, the reasons for this should be mentioned (e.g. ethical, rules of personal data, intellectual property, commercial, privacy-related, security-related).

<input here>

Method for Data Sharing

Consider where, how, and to whom the data should be made available. Will you share data via a data repository, handle data requests directly or use another mechanism? The methods used to share data will be dependent on a number of factors such as the type, size, complexity and sensitivity of data. Mention earlier examples to show a track record of effective data sharing.

• *How will you make the data available to others?* **<input here>**
• *With whom will you share the data, and under what conditions?* **<input here>**

Restrictions on Sharing

Outline any expected difficulties in data sharing, along with causes and possible measures to overcome these. Restrictions to data sharing may be due to participant confidentiality, consent agreements or IPR. Strategies to limit restrictions may include: anonymising or aggregating data; gaining participant consent for data sharing; gaining copyright permissions; and agreeing a limited embargo period.

• *Are any restrictions on data sharing required? e.g. limits on who can use the data, when and for what purpose.* **<input here>**
• *What restrictions are needed and why?* **<input here>**
• *What action will you take to overcome or minimise restrictions?* **<input here>**

Data Repository

Most research funders recommend the use of established data repositories, community databases and related initiatives to aid data preservation, sharing and reuse. An international list of data repositories is available via Databib or Re3data.

• *Where (i.e. in which repository) will the data be deposited?* **<input here>**

Archiving and preservation (including storage and backup)

Questions to consider before answering:

•What is the long-term preservation plan for the dataset? e.g. deposit in a data repository

•Will additional resources be needed to prepare data for deposit or meet charges from data repositories?

Researchers should consider how datasets that have long-term value will be preserved and curated beyond the lifetime of the grant. Also outline the plans for preparing and documenting data for sharing and archiving. If you do not propose to use an established repository, the data management plan should demonstrate that resources and systems will be in place to enable the data to be curated effectively beyond the lifetime of the grant.

• What additional resources are needed to deliver your plan?

• Is additional specialist expertise (or training for existing staff) required?

• Do you have sufficient storage and equipment or do you need to cost in more?

• Will charges be applied by data repositories?

• Have you costed in time and effort to prepare the data for sharing / preservation?

Carefully consider any resources needed to deliver the plan. Where dedicated resources are needed, these should be outlined and justified. Outline any relevant technical expertise, support and training that is likely to be required and how it will be acquired. Provide details and justification for any hardware or software which will be purchased or additional storage and backup costs that may be charged by IT services. Funding should be included to cover any charges applied by data repositories, for example to handle data of exceptional size or complexity. Also remember to cost in time and effort to prepare data for deposit and ensure it is adequately documented to enable reuse. If you are not depositing in a data repository, ensure you have appropriate resources and systems in place to share and preserve the data.

Describe the procedures that will be put in place for long-term preservation of the data.

<input here>

Indicate how long the data should be preserved, what is its approximated end volume, what the associated costs are and how these are planned to be covered. **<input here>**

4.1.3 DMP at final stage (to be ready when data is available)

SCIENTIFIC RESEARCH DATA SHOULD BE EASILY ***DISCOVERABLE***

Questions to consider:

• How will potential users find out about your data?

• Will you provide metadata online to aid discovery and reuse?

Guidance: Indicate how potential new users can find out about your data and identify whether they could be suitable for their research purposes. For example, you may provide basic discovery metadata online (i.e. the title, author, subjects, keywords and publisher).

Are the data and associated software produced and/or used in the project discoverable (and readily located), identifiable by means of a standard identification mechanism (e.g. ***Digital Object Identifier****)?* **<input here>**

SCIENTIFIC RESEARCH DATA SHOULD BE **ACCESIBLE**

Questions to consider:

• Who owns the data?

• How will the data be licensed for reuse?

• If you are using third-party data, how do the permissions you have been granted affect licensing?

• Will data sharing be postponed / restricted e.g. to seek patents?

State who will own the copyright and IPR of any new data that you will generate. For multi-partner projects, IPR ownership may be worth covering in a consortium agreement. If purchasing or

reusing existing data sources, consider how the permissions granted to you affect licensing decisions. Outline any restrictions needed on data sharing e.g. to protect proprietary or patentable data. See the DCC guide: How to license research data.

Are the data and associated software produced and/or used in the project accessible and in what modalities, scope, licenses? (e.g. licencing framework for research and education, embargo periods, commercial exploitation, etc) **<input here>**

*SCIENTIFIC RESEARCH DATA SHOULD BE **ASSESSABLE** AND INTELLIGIBLE*

• What metadata, documentation or other supporting material should accompany the data for it to be interpreted correctly?

• What information needs to be retained to enable the data to be read and interpreted in the future?

Describe the types of documentation that will accompany the data to provide secondary users with any necessary details to prevent misuse, misinterpretation or confusion. This may include information on the methodology used to collect the data, analytical and procedural information, definitions of variables, units of measurement, any assumptions made, the format and file type of the data.

Are the data and associated software produced and/or used in the project assessable for and intelligible to third parties in contexts such as scientific scrutiny and peer review?, e.g. are the minimal datasets handled together with scientific papers for the purpose of peer review, are data is provided in a way that judgments can be made about their reliability and the competence of those who created them **<input here>**

***USABLE** BEYOND THE ORIGINAL PURPOSE FOR WHICH IT WAS COLLECTED*

• What is the long-term preservation plan for the dataset? e.g. deposit in a data repository

• Will additional resources be needed to prepare data for deposit or meet charges from data repositories?

Researchers should consider how datasets that have long-term value will be preserved and curated beyond the lifetime of the grant. Also outline the plans for preparing and documenting data for sharing and archiving. If you do not propose to use an established repository, the data management plan should demonstrate that resources and systems will be in place to enable the data to be curated effectively beyond the lifetime of the grant.

Guidance on Metadata:

• How will you capture / create the metadata?

• Can any of this information be created automatically?

• What metadata standards will you use and why?

Metadata should be created to describe the data and aid discovery. Consider how you will capture this information and where it will be recorded e.g. in a database with links to each item, in a 'readme' text file, in file headers etc.

Researchers are strongly encouraged to use community standards to describe and structure data, where these are in place. The DCC offers a catalogue of disciplinary metadata standards.

Are the data and associated software produced and/or used in the project useable by third parties even long time after the collection of the data? e.g. is the data safely stored in certified repositories for long term preservation and curation; is it stored together with the minimum

software, metadata and documentation to make it useful; is the data useful for the wider public needs and usable for the likely purposes of non-specialists? **<input here>**

INTEROPERABLE** TO SPECIFIC QUALITY **STANDARDS

• What format will your data be in?

• Why have you chosen to use particular formats?

• Do the chosen formats and software enable sharing and long-term validity of data?

Outline and justify your choice of format e.g. SPSS, Open Document Format, tab-delimited format, MS Excel. Decisions may be based on staff expertise, a preference for open formats, the standards accepted by data centres or widespread usage within a given community. Using standardised and interchangeable or open lossless data formats ensures the long-term usability of data?

See the UKDS Guidance on recommended formats

Are the data and associated software produced and/or used in the project interoperable allowing data exchange between researchers, institutions, organisations, countries, etc?, e.g. adhering to standards for data annotation, data exchange, compliant with available software applications, and allowing re-combinations with different datasets from different origins

<input here>

4.2 Data Levels, Data Acquisition, Data Curation, Data Ingestion

4.2.1 General description of data levels

Indicate if the DATASETS are organized into different levels (LEVEL-0, 1, 2, 3,4) and if so what are the relevant definitions and how DOI are provided. **<input here>**

4.2.2 Collection/Acquisition

Gathering RAW data

Specify how do you gather/collect your data (e.g. sensors, observations, satellites, etc.)? **<input here>**

How do you pre-process, transfer and store your RAW data? **<input here>**

From RAW Data to Calibrated Data

Describe the processes applied for Data Calibration, Validation, Filtering, etc. **<input here>**

4.2.3 Access to external data

Describe the identification and access to External Data **<input here>**

Indicate if there is a procedure for validation of External Data **<input here>**

4.2.4 Data curation

Specify any automatic check applied, like completing series, detecting outlier **<input here>**

Describe manual quality checks **<input here>**

Are there quality flags applied to the data? **<input here>**

4.2.5 Data ingestion / integration

Describe transformations applied to data taking into account ontologies/metadata. Indicate also if there is any "harmonization procedure" (to share/integrate data) and how linking internal and external data is made if relevant. **<input here>**

4.2.6 Further data processing

Describe, if relevant, the different additional processing steps (and the associated software and resources) applied to the (collected/curated) datasets to provide a "final" dataset collection that can be used in the analysis **<input here>**

4.3 Analysis

4.3.1 Basic analysis and standard analysis suites

Describe usual examples of basic analysis in the Case Study **<input here>**

Specify if software packages/tools like MATLAB, R-Studio, iPython,etc. are used **<input here>**

4.3.2 Data analytics and Big Data

Describe relevant examples of advanced analysis in the Case Study (like for example application of neural networks, series analysis, etc.) **<input here>**

Specify the resources and additional software required **<input here>**

Identify analysis challenges that can be classified as "Big Data" **<input here>**

List Big Data driven workflows **<input here>**

4.3.3 Data visualization and interactive analysis

Indicate the need for data and analysis results visualization **<input here>**

Indicate how visualization is made and if interactivity/steering is needed **<input here>**

Specify the User Interfaces (web, desktop, mobile, etc.) **<input here>**

4.4 Data Publication

Describe the information flow from the analysis to the publication **<input here>**

Indicate the requirements from publishers/editors to access data, and how it is made available (open data?) **<input here>**

5 SIMULATION/MODELLING

Describe the Simulation/Modelling requirements in this Case Study. Please identify also any other intensive CPU mainly activity as required.

5.1 General description of simulation/modelling needs

Describe the different models used (including references) **<input here>**

*Indicate the type and quantity of simulations needed in the Case Study, and how they are incorporated in the general workflow of the solution***<input here>**

5.2 Technical description of simulation/modelling software

For each simulation package:

Identify the simulation software **<input here>**

Provide a link to its documentation, and describe its maturity and support level **<input here>**

Indicate the requirements of the simulation software (hardware: RAM, processor/cores, extended instruction set, additional software and libraries, etc.) **<input here>**

Tag the simulation software as HTC or HPC **<input here>**

*List the input files required for execution and how to access them***<input here>**

Describe the output files and how they will be stored **<input here>**

Reference an existing installation and performance indicators **<input here>**

Specify if the simulation software is parallelized (or could be adapted) **<input here>**

Specify if the simulation software can exploit GPUs **<input here>**

Specify how the simulation software exploits multicore systems **<input here>**

Specify if parametric runs are required **<input here>**

Estimate the use required of the resources (million-hours, # cores in parallel, job duration, etc) **<input here>**

5.3 Simulation Workflows

Describe if there are workflows combining several (HTC/HPC) simulations or simulations and data processing **<input here>**

6 DETAILED USE CASES FOR RELEVANT USER STORIES

This section tries to put the focus on the preparation of detailed Use Cases starting from User Stories most relevant to the Case Study considered.

6.1 Identification of relevant User Stories

Examples of relevant User Stories linked to roles like for example Final User, Data Curator, etc.

List User Stories based on data collection, curation, processing, analysis, simulation, etc, that are considered most relevant for the Case Study being analyzed **<input here>**

For each relevant User Story:

Draft a basic card **<input here>**

Provide details from conversation with the researchers' teams **<input here>**

Draft as a Use Case **<input here>**

Analyze tools to support the definition of the Use Case (like mockups). Integrate in the analysis the requirements on user interfaces (like the use of mobile resources, under different flavours, access through web interfaces, etc.) **<input here>**

Describe the way to extract requirements and define acceptance criteria **<input here>**

Include if possible an example of support for Big Data driven workflows for e-Science, with requirements for scientific workflows management, under a "Workflow as a Service" model, where the proper workflow engines will be selected according to user needs and requirements.

In such case please describe the scenario for Big Data analysis, and assure that the Use Case considers which levels of workflow engines are needed (e.g., "coarse gran", which targeting distributed (loosely coupled) experiments, through workflow orchestration across heterogeneous set of services; "fine grain", which targeting high performance (tightly coupled) data analysis through workflows orchestration on big data analytics frameworks)

7 INFRASTRUCTURE TECHNICAL REQUIREMENTS

Describe the Case Study from the point of view of the required e-infrastructure support.
INDIGO Data-Cloud will support the use of heterogeneous resources.

7.1 Current e-Infrastructures Resources

Start from the current use of e-infrastructures.

7.1.1 Networking

Describe the current connectivity **<input here>**
Describe the key requirements (availability, bandwidth, latency, privacy, etc) **<input here>**
Specify any current issue (like last mile, or access from commercial, etc) **<input here>**

7.1.2 Computing: Clusters, Grid, Cloud, Supercomputing resources

Describe the current use of each of these type of resources: size and usage **<input here>**
Indicate if there is any mode of "orchestration" between them **<input here>**

7.1.3 Storage

Describe the current resources used **<input here>**
Discuss the key requirements (I/O performance, capacity, availability, reliability, any other QoS indicator) **<input here>**

7.2 Short-Midterm Plans regarding e-Infrastructure use

Plans for next year (2016) and in 5 years (2020).

7.2.1 Networking

Describe the proposed connectivity **<input here>**
Describe new/old key requirements (availability, bandwidth, latency, QoS, private networking, etc) **<input here>**
Specify any potential solution/technique (for example SDN) **<input here>**

7.2.2 Computing: Clusters, Grid, Cloud, Supercomputing resources

Describe the evolution expected: which infrastructures, total "size" and usage **<input here>**
Detail potential "orchestration" solutions **<input here>**

7.2.3 Storage

Describe the resources required **<input here>**
Discuss the key requirements (I/O performance, capacity, availability, reliability, any other QoS indicator) **<input here>**

7.2.4 SPECIFIC QUESTIONS REGARDING USE OF EGI.eu (FROM EGI DOC 2478)

Sample questions to capture details of a support case

These questions can help case supporters interview the case submitter and the NGIs to refine the technical details of the case and ultimately to move towards a suitable technical setup. These questions aim at understanding the user's need, the technical and other requirements/constrains of the case, and the impact that a solution would bring to the scientific community. These questions provide only guidance – Ticket owners can use other questions or even other methods to identify details of their support case(s).

- *What does the user/community want to achieve? (What's the user story?)*
- *For who does the case request resources for? (CPU/storage capacity, SW tools, consultant time, etc.) For a group? For a project? For a collaboration? Etc.*
- *What is the size of the group that would benefit from these resources, and where these people are? (which country, institute)*
- *Approximately how much compute and storage capacity and for how long time is needed? (may be irrelevant if the activity is for example assessment of an EGI technology)*
- *Does the user need access to an existing allocation (→ join existing VO), or does he/she needs a new allocation? (→ create a new VO)*
- *What is the scientific discipline?*
- *Which institute does the contact work for (or those he/she represents)?*
- *Does the case include preferences on specific tools and technologies to use?*
    - *For example: grid access to HTC clusters with gLite; Cloud access to OpenStack sites; Access to clusters via standard interdafaces; Access to image analysis tools via Web portal*
- *Does the user have preferences on specific resource providers? (e.g. in certain countries, regions or sites)*
- *Does the user (or those he/she represents) have access to a Certification Authority? (to obtain an EGI certificate)*
- *Does the user (or those he/she represent) have the resources, time and skills to manage an EGI VO?*
- *Which NGIs are interested in supporting this case? (Question to the NGIs)*

7.3 On Monitoring (and Accounting)

Please outline any requirements for monitoring of the platforms and the applications.

If you have specific tools already in use, please outline them.

Please also specify monitoring, metrics at different levels: system, performance, availability, network QoS, website, security, etc.

<input here>

7.4 On AAI

(From EGI, revise and check with WP4/5/6)

Describe the current AAI status of your community/research infrastructure

• Does your community/research infrastructure already use AAI solutions? **<input here>**

• Can you describe the solutions you have adopted highlighting as applicable: Technology adopted (e.g. X509, SAML Shibboleth,...), Identity Providers (IdP) federations integrated (e.g. eduGAIN) or approximate number of individual IdPs integrated, Solution for homeless users (users without an insitutional IdP), Solutions to handle user attributes **<input here>**

Describe the potential needs and expectations from an AAI integration in the ***services and platforms provided by INDIGO***

- *Type of IdP to be integrated (e.g. institutional IdP part of national federations and eduGAIN or non federated, social media credentials, dedicated research community catch-all IdP, ...)* **<input here>**
- *Preferred authentication technology, and requirements for support of multiple technology and credential translation services (e.g. SAML -> X509 translation)* **<input here>**
- *Community level authorization/attribute based authorization to support different authorization levels for the users* **<input here>**
- *Web access and/or non-web access* **<input here>**
- *Need for delegation (e.g. execute complex workflows on behalf of the user)* **<input here>**
- *Support for different level of assurance credentials, and need to use the information about users with lower level of assurance credentials to limit their capability* **<input here>**
- *Requirements for high level of assurance credentials (e.g. to access confidential/sensitive data)* **<input here>**

7.5 On HPC

Describe any specific issue related to the use of supercomputers.

<input here>

7.6 Initial short/summary list for "test" applications (task 2.3)

Software used	***Software/applications/services required, configuration, dependencies (Describe the software/applications/services name, version, configuration, and dependencies needed to run the application, indicating origin and requirements.)*** **<input here>**
Operating system requirements	**<input here>**
Run libraries requirements	***Run API/libraries requirements (e.g., Java, C++, Python, etc.)*** **<input here>**
CPU requirements (multithread,MPI, "wholenode")	**<input here>**
Memory requirements	**<input here>**
Network requirements	**<input here>**
Disk space requirements (permanent, temporal)	***Include the requirements for data transferring (upload and download of data objects: files, directories, metadata, VM/container images, etc.)*** **<input here>**
External data access requirements	**<input here>**
Typical processing time	**<input here>**
Other requirements	***Requirements for data synchronization*** ***Requirements for data publication*** ***Requirements for depositing data to archives and referring them*** ***Requirements for mobile application components for data storage and access*** ***Requirements for data encryption and integrity control-related functionality*** **<input here>**
Other comments	**<input here>**

Relevant references or URLs	**<input here>**

8 CONNECTION WITH INDIGO SOLUTIONS

<To be filled by INDIGO JRA >

8.1 IaaS / WP4

8.2 PaaS / WP5

8.3 SaaS / WP6

8.4 Other connections

9 FORMAL LIST OF REQUIREMENTS

<this will be further edited within WP2>

10 REFERENCES

R 1	
R 2	
R 3	
R 4	
R 5	