INDIGO-DataCloud

INITIAL REQUIREMENTS FROM RESEARCH COMMUNITIES ANNEX 1.*PX*: SELECTED CASE STUDY FROM *YOUR RESEARCH*

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

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

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.

---

[1] For a nice intro, see: https://whyarerequirementssohard.wordpress.com/2013/10/08/when-to-use-user-stories-

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

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*: **<input here>**
- *Institution/partner representing the community in INDIGO:* **<input here>**
- *Main contact person*: **<input here>**
- *Contact email:* **<input here>**
- *Specific Title for the Case Study:* **<input here>**

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

<input here>

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?

<input here>

1.4 Expected results and derived impact

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

<input here>

1.5 References useful to understand the Case Study

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

<input here>

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

Each Galaxy instance will be created by single users/group of users (e.g. people working in the same lab) in order to have a fully customizable version of the workflow manager. Each instance will provide to its creator full administrative powers, allowing the installation of new tools, control over resource management (disk quotas, computing time, etc..), access to stored data, etc... On the other hand each instance will be completely insulated from other instances running on the same infrastructure and the data stored in each instance (in order to be processed by the workflow manager) will be inaccessible/unreadable not only from other users of the same infrastructure lacking an account for that specific instance, but also by the infrastructure administrators. This layout provides the basis for a work environment where sensible data stored in each of the Galaxy instances will be accessible/readable only by who has a valid user account for the specific instance.

Each Galaxy instance will have to deal with highly heterogeneous data like genomic, transcriptomic, metagenomic, epigenomic and other -omic sequencing data obtained with Next Generation Sequencing techniques (.fastq or .fq file extensions), various types of genomic annotations coming in many formats (.bed, .gtf, .bedgraph. .wiggle, etc... file extensions), sequence alignments (.sam, .bam file extensions) and many more. Most files are in plain text format while some others are in binary format.

Each Galaxy instance will be a complete workflow manager for the bioinformatic processing of biological data. Each instance will have access to one or more Galaxy repositories in order to obtain and install new tools and workflows and update existing ones. In this way each Galaxy instance will be tailored to the specific needs of each user/group of users bypassing the barrier to installing new tools imposed by classic public Galaxy instances.

Users will be associated to projects, which will have a separate storage area and access to processing resources. Those processing resources may be seamlessly provided elsewhere (research infrastructures or even public clouds), depending on the workload and requirements of the study. Users will access through a web-based, simple interface.

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

Three levels of reseachers compose the community:

1 - Basic users: mostly use Galaxy to run simple analysis pipelines and usually have no means to set up a personalized instance due to lack of resources and/or competences. They would greatly benefit from the possibility of deploying their own customizable Galaxy instance, since they will obtain full

control over who access their data and the option to decide which tools should be available in their instance.

2 - Advanced users: use Galaxy to run complex workflows in a multi-user setup with the possibility to share data and results among other users of the same Galaxy instance. They often want to try different tools to perform the same analysis step in order to compare outputs and decide which approach better suits their specific data. Even advanced users often lack the resources/competences needed to set up and maintain a Galaxy instance and can greatly benefit from having one.

3 - Bioinformaticians: like advanced users but they usually have the competences (but not always the resources) to set up and maintain a Galaxy instance. They want to test custom pipelines and even tools they developed within the Galaxy environment. Freeing them from the burdensome activity of administering a Galaxy installation, like they have to do in many research environments, will make their time available for more valuable activities like bioinformatic tools and algorithms development and implementation, and the execution of complex data analysis.

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

A wide community of experts around the world is actively developing the Galaxy workflow manager and it has a very large users base among life science researchers. Recently this workflow manager has seen an interest even by other research communities and are reported active instances of Galaxy dedicated to chemistry and other disciplines. This case study aims to make the installation and administration of a customized Galaxy instance in a cloud environment available to every researcher without the need of any specific competence and currently it is in its design phase. ELIXIR-Italy (coordinated by CNR) will lead the pilot action and the National Institute for Nuclear Physics (INFN) of Bari will develop and deploy prototype on its cloud infrastructure, in collaboration with other technological partners of the ELIXIR-Italy, while CNR will provide coordination, feedback and perform usability tests. The prototype will exploit the hardware resources provided by INFN Bari in terms of cloud, grid and HPC. At the moment the site is able to provide more than 600 CPU/Cores, 3TB of RAM and more than 100TB of storage. We foresee to complete the project in about 18 months, divided as follows:

Month 1 -- Month 3

Requirements analysis and scouting of the already available technologies. (CURRENT PHASE)

Month 4 -- Month 8

Development of a first prototype of the core system.

Month 9 -- Month 12

User interface development.

Month 13 -- Month 15

Deployment of an alpha version of the software. Usability tests from real users and bug fixing.

Month 16 -- Month 18

Deployment of the beta version of the software. Stress tests and bug fixing.

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

The application case of the community has several requirements at different stages:

- Privacy and security. Despite the multi-tenancy of the computing infrastructure, data must not be accessible by different projects. Data, even anonymised, and produced results must be protected from the access of unauthorised users including the administrators of the cloud platform.

- Persistent Storage. Data must be kept even if the computing nodes are powered down. Frequent data transfers should be avoided as they may involve many GBs (up to TBs) of data.

- Software configuration. Each Galaxy instance will have specific requirements (in terms of resources and applications) that have to be fulfilled individually. Each subproject must fill-in a check- list form with the software and computing requirements and the back-end infrastructure should be compliant to them. Efficient execution. Instances may require resources or have external resources available, and the applications should work seamlessly. Automatic elasticity is taken for granted.

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.

The infrastructure will be supported by INFN. INFN is willing to contribute with resources to the use case. Currently, it is not using any external infrastructure.

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

Many life scientists, bioinformaticians and computational biologists across Europe and beyond are familiar with the Galaxy workflow manager environment. Giving them the option to have their own Galaxy setup will first of all empower them with the possibility to add any tool they like to the Galaxy interface without the need to ask to public instances administrators that may not want to modify their Galaxy installation for the sake of a single user. Then, public computational resources and infrastructures offering the Galaxy instance on-demand service could be better exploited by life

scientists without the expertise to use command line tools. Another big advantage of this solution is to give the option to investigators dealing with sensitive human data to use a friendly environment like Galaxy without the need to install and manage a local instance. In fact, these medical researchers often face legal constraints that restrain them from moving their data to external servers, thus denying the possibility to take advantage of public Galaxy instances to analyse their data, and giving them self-contained Galaxy instances completely insulated from others and even from the cloud environment administrators can solve this problem. Envisioning a not so distant future where NGS data will be extensively used even in the healthcare sector routinely, this approach could be also exploited to give to healthcare operators standardized tools and workflows to perform a number of -omics assays on patients data.

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* At least one within INDIGO and potentially another within ELIXIR-ITA.
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*: Many
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

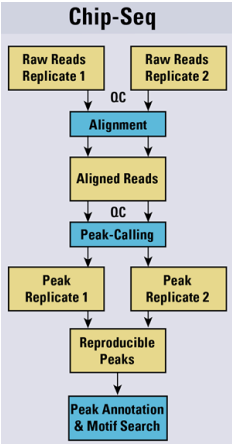
Describe the Case Study from the point of view of developers (4 pages max.)
Assemble it using preferably an AGILE scheme based on User Stories.

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.

Galaxy instances are mainly targeted to small to medium workgroups of life scientists. Typically they have to process in a reasonable amount of time sequencing data from NGS experiments. The workflows vary enormously between each type of experiment (e.g. ChIP-Seq, RNA-Seq, ChIA-Pet, RIP-Seq, Genomic (re)sequencing, Exome analysis, etc...) but they usually share the very first steps that consist in checking the quality of the data obtained by the sequencing facility and to map them on a reference genome, that is usually one of the most CPU demanding task of the analysis. From this point on the workflows diverge greatly depending on the type of the experiment and advanced analysis may diverge even among the same experiment type (e.g. ChIP-Seq analysis workflows vary greatly according to the type of the investigated protein). Each Galaxy instance will be customizable in line with the main activities of the working group it belongs to, thus providing the set of tools needed by the workgroup itself. Installation of novel/updated tools will be straightforward thanks to the Galaxy repositories, including the installation of custom software produced by the workgroup itself. Sharing of data and workflows among the workgroup components will be straightforward thanks to the Galaxy built-in sharing features. Reference data (e.g. genomes, annotation files, pre-computed genome indexes, etc...) will be available in two ways. The most used ones will be included in each Galaxy instance while for the most specific or exotic it will be possible for the user to upload them to the Galaxy instance as needed.



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.)

- Galaxy instance administrator: who have set up a Galaxy instance in a cloud environment using the Galaxy instantiator. An administrator has the possibility to customize his Galaxy instance with the bioinformatic tools he prefers, grant access to the instance to other users by creating accounts, set usage limits (e.g. disk quotas, number of concurrent jobs per users, etc...) and parameters. Administrators will likely be principal investigators, project task leaders and bioinformatic tools developers.
- Galaxy instance users: all the other researchers who have access to a Galaxy instance. They will use the workflow manager to perform their bioinformatic analyses using single tools or

workflows, with the possibility to share their data and workflows with other users of the same instance. They interact with the Galaxy administrator for any need that may arise, e.g. the installation of a new bioinformatic tool in the Galaxy instance. But, differently from what happens with standard public Galaxy instances, the Galaxy administrator will likely be in the same room / building and will be more receptive to users request since they are likely working for or with him on the same project.

3.3 General description of datasets/information used

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

Various biological databases run by EBI, NCBI and many others. Typical datasets will comprehend reference genomes, reference annotations, genome, transcript and protein sequences, epigenetic data and many others. Most common reference data, like genomes for model organisms, will be available for each instance in order to avoid unnecessary transfer of large data.

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

<input here>

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.

Once an instance is running users will have to upload the data to be analysed, while reference data will generally be already available within the Galaxy instance or in alternative obtainable from public biological databases. Based on the type of data and analysis required, users will have to select the appropriate tools from the Galaxy interface and set the parameters for the analysis. Users will be able to create without effort complex workflows by concatenating tools in order to create automated pipelines for the most frequent tasks. Outputs will naturally be of different kinds based on the analysis type. All of them will be downloadable directly from within the Galaxy instance and most of them will be viewable within the Galaxy instance itself both directly or via various graphical representations, or alternatively using external resources like genomic browsers.

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.

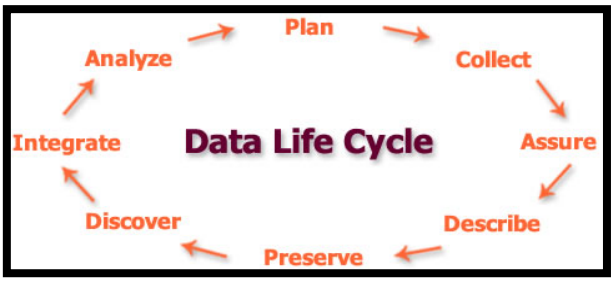
Most Galaxy public resources nowadays are running on single servers or cluster configurations based on their scope. While the Galaxy workflow manager itself does not need many computational resources, some of the bioinformatic tools that can be run with it, and in particular tools for NGS data analysis, are resource hungry and greatly benefits from heavily parallel computational environments and great amounts of storage and RAM. A cloud-based solution like the one proposed would allow tailoring the resources needed to each instance in order to achieve maximum efficiency. In fact, while as said many bioinformatic tools for NGS data analysis are resource hungry, they tend to be run in short bursts (e.g. mapping reads to a reference genome) and then be followed by a number of less

intensive tasks. Thus, a cloud-based solution represents an efficient way to make many Galaxy instances cohabit the same platform and sharing resources that would be otherwise used in a far less efficient way with a single instance/single physical server approach.

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?* Mostly NGS technologies.

• *What standards or methodologies will you use?* See above

• *How will you structure and name your folders and files?* Not applicable

• *How will you ensure that different versions of a dataset are easily identifiable?* Each user will use its own policy.

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?* Each Galaxy instance has a built-in metadata handler.

• *Can any of this information be created automatically?* Yes, and they will be.
• *What metadata standards will you use and why?* NA

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).

Each administrator will adopt its own policy based on the project(s) he is working on.

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?* Data can be easily shared among different users of the same instance.
• *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	