INDIGO-DataCloud

INITIAL REQUIREMENTS FROM RESEARCH COMMUNITIES ANNEX 1.*P7*: SELECTED CASE STUDY FROM *EGI FEDCLOUD*

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	Peter Solagna	**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*: **EGI Federated cloud use cases**
- *Institution/partner representing the community in INDIGO:* **EGI.eu**
- *Main contact person*: **Peter Solagna, Yin Chen**
- *Contact email:* **peter.solagna@egi.eu, yin.chen@egi.eu**
- *Specific Title for the Case Study:* **NA**

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.

EGI Federated cloud is a federation of institutional private cloud infrastructure. The federation supports a number of use cases and communities with diverse requirements and research fields.

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?

By deploying the specific technical use cases described in the next sections of the document. Fedcloud aims to improve the federation capabilities of cloud services in a distributed infrastructure.

The main areas of improvements are:

Improvements in the IaaS layer. Improve the features of the cloud storage, in particular the collaboration capabilities. Make easier and more reliable for the usersto share data and accessing concurrently storage areas and dataset.

Improvements in the PaaS layer: Advanced orchestration capabilities for deploying multi-services platforms.

1.4 Expected results and derived impact

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

The goal of the EGI Fedcloud requirements is to improve the user experience in using federated and distributed cloud services, overcoming the current technology limitations, to enable efficient deployment of the different use cases and maximise the use of the EGI resources.

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

The federated cloud use case is not limited to a single case study but includes a wide set of applications and use cases. This section considers some of the main use cases emerged from the Fedcloud outreach activities, focusing on the the following use cases, which have been deployed in oone or more sites of the Federated cloud:

- Chipster
    - Chipster is a user-friendly analysis software for high-throughput data. It contains over 300 analysis tools for next generation sequencing (NGS), microarray, proteomics and sequence data.
- READemption
    - READemption is a pipeline for the computational evaluation of RNA-Seq data. It was originally developed to process dRNA-Seq reads (as introduced by Sharma et al., Nature, 2010 (Pubmed)) originating from bacterial samples. Meanwhile is has been extended to process data generated in different experimental setups and from all domains of life.
- JAMS
    - JAMS is an java-based, open-source software platform that has been especially designed to address the demands of a process-based hydrological model development and various aspects of model application. JAMS is a framework to build up complex models out of simple components. Several hydrological models were implemented within JAMS (e.g. J2000, J2000g). Usually those models are applied to simulate hydrological dynamics in catchments with a size of 1km² to 100,000 km² in a temporal time step of hours to months.
- HAPPI
    - SCIence Data Infrastructure for Preservation with focus on Earth Science (SCIDIP-ES) brings together the state of the art in preservation technologies, represented by Earth Science repositories, and researchers for digital data preservation techniques. SCIDIP-ES HAPPI supports the archive manager and curator to capture and manage part of the Preservation Descriptive Information (PDI).

- INERTIA
    - INERTIA project addresses the "structural inertia" of existing Distribution Grids by introducing more active elements combined with the necessary control and distributed coordination mechanisms. To this end INERTIA will adopt the Internet of Things/Services principles to the Distribution Grid Control Operations.
- DRIHM
    - The DRIHM project is a European initiative running from 1st September 2011 to 28th February 2015 aiming at providing an open, fully integrated workflow platform for predicting, managing and mitigating the risks related to extreme weather phenomena.
    - Forecasting severe storms and floods is a key topic in HMR. Storms do not respect country boundaries so a pan-European approach to data access and modeling is necessary. These data challenge our current scientific understanding and call for focused and joint Hydro-Meteorological and ICT research to:
        - understand, explain and predict the physical processes producing such extreme storms;
        - understand the possible intensification of such events in the Mediterranean region and their physical origin;
        - explore the potential available ICT infrastructures to provide deeper understanding of those events through fine resolution modeling over large areal extents.
- BILS
    - BILS (Bioinformatics Infrastructure for Life Sciences) is a distributed national research infrastructure supported by the Swedish Research Council (Vetenskapsrådet) providing bioinformatics support to life science researchers in Sweden. BILS is also the Swedish node in the European infrastructure for biological information ELIXIR.
    - BILS would is evaluating the EGI Federated Cloud to get computational resources. BILS is willing to rewrite its current services to scale up compute in cloud and so exploiting the Cloud Elasticity.
    - The BILS portals is front-end to biological tools for not IT skilled users. Currently, all user compute jobs run in worker nodes in small clusters. BILS is interested to run the compute jobs on the EGI Federated Cloud and so scale-up.

EGI is also in contact with several other communities/research infrastructures not yet using Fedcloud, but in contact with EGI to discuss requirments to possibly support their use cases. One example is the Human Brain Project.

- Human Brain Project
  - The aim of the Human Brain Project (HBP) is to accelerate our understanding of the human brain by integrating global neuroscience knowledge and data into supercomputer-based models and simulations. This will be achieved, in part, by engaging the European and global research communities using six collaborative ICT platforms: Neuroinformatics, Brain Simulation, High Performance Computing, Medical Informatics, High Performance Computing, Neuromorphic Computing and Neurorobotics. For HBP a key capability is to deliver multi-level brain atlases that enable the analysis and integration of many different types of data into common semantic and spatial coordinate frameworks. HBP is looking for different type of repositories that allow interactive access to selected sub-set of data wanted and ultimately to be able to do analysis where that data sets are. The purpose is to leave the data in place, without moving it outside of the repositories.
  - Use Case 1: Remote interactive multiresolution visualization of large volumetric datasets. Make data available and accessible without moving large amounts of data. Typical dataset sizes can reach in the terabyte range, while a researcher may want to only view or access a small subset of the entire dataset.
  - Use Case 2: Feature extraction and analysis of large volumetric datasets. In this use case, a user would provide via a web service input parameters to a data application which would trace any recognized neuron structures using a selected algorithm. The output file would be returned via the webservice.

EGI-Engage project collaborates eight EU high-impact research infrastructures/communities by joint development of customised services based on core EGI capabilities, the so-called competence Centres (CCs). Communities collaborated through EGI-Engage CC program include, BBMRI-ERIC, ELIXIR, MoBrain, DARIAH, LifeWatch, EISCAT-3D and EPOS. Since ELCIXIR, MoBrain and LifeWatch have already good representations in INDIGO, we only include new use cases that emerge in EGI-Engage CCs having requirements for using Cloud as follows:

- BBMRI-ERIC CC
  - Thousands of biobanks in Europe have been collecting data, samples and images of millions of individuals in different stages of their lives, during disease and after recovery. Biobanking is currently evolving from local repositories to a pan-European RI the BBMRI-ERIC. The BBMRI CC facilitates the implementation of big data storage in combination with data analysis and data federation by integrating technologies from community projects, EGI and other e-Infrastructures.
  - The CC will capture requirements and provide technology demonstrators to:

- Increase biobank interoperability and data discovery in BBMRI-ERIC community by providing a secure and standard way to share biobank high-throughput data,
- Provide biobanking community with a federated infrastructure for big data storage and intensive data analysis,
- Facilitate the efficient use of bio-resources by supporting visibility and sharing, while also respecting the protection level required by owners of the data and samples,
- Facilitate the efficient use of economic resources in BBMRI-ERIC by providing a common informatics infrastructure for storage and processing of big data.

- DARIAH CC
  - It Aims to widen the usage of the e-Infrastructures for Arts and Humanities research. The CC will develop and provide a workflow-based science gateway based on the generic-purpose WS-PGRADE and gLibrary technologies, adapted and tailored to the needs of users coming from the field of Arts and Humanities. The gateway will provide access and compute services for data residing in distributed grid and cloud storages. The gateway will be validated and enriched with the 'Multi-Source Distributed Real-Time Search and Information Retrieval' application (SIR). The CC will engage with Arts and Humanities communities to attract more applications and users to the gateway.
- EPOS CC
  - It aims to drive the future design of the use of grid and cloud for the integrated solid Earth Sciences research as part of the European Plate Observing System (EPOS). The CC will (1) identify and validate authentication and authorisation services, (2) will test cloud resources and usage models, (3) provide knowledge transfer services between e-Infrastructure and EPOS communities.
- Disaster Mitigation
  - The objective of the this CC is to make available customised IT services to support the climate and disaster mitigation researchers to gain a deeper understanding of the most serious natural disasters that affect Asia (e.g. earthquakes, tsunamis, typhoons) and to mitigate multi-hazards via data-intensive, e-Science techniques and collaborations. The task strongly builds on experts from the Asia-Pacific region who will create virtual research environments with embedded services and simulations that enable the sharing of disaster-related data, tools, applications and knowledge among field- workers, scientists, and e-Infrastructure experts, shortening the time they can respond to natural disasters.

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

- Chipster
  - Life scientists face three major requirements when analyzing next generation sequencing data: Installation of a large number of software and reference data, which need to be kept up to date. Unix and programming skills. Computing platform with sufficient CPU and memory
- READemption
- JAMS
- HAPPI
- INERTIA
- DRIHM
  - Running various hydrological models in the EGI Federated Cloud. Almost all these models need a Windows environment to be executed.
- BILS

Non yet in fedcloud

- Human Brain Project
-

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

- Chipster
  - The Chipster virtual machine provides a comprehensive collection of up-to-date analysis tools and Ensembl-based reference data in a ready-to-use format. The data and the tools can be used either on command line, or via an intuitive GUI which also provides powerful visualizations, workflow functionality, and analysis metadata tracking.
- READemption
  - The functions which are accessible via a command-line interface cover read processing and aligning, coverage calculation, gene expression quantification,

differential gene expression analysis as well as visualization. In order to set up and perform analyses quickly READemption follows the principal of convention over configuration: Once the input files are copied/linked into defined folders no further parameters have to be given. Still, READemption's behavior can be adapted to specific needs of the user by parameters.

- JAMS
  - Typically, the computing time of a single model run is in the range from minutes to hours (on a single workstation). However, those models have many parameters, which must be estimated indirectly during a calibration process. For model calibration an evolutionary optimization algorithms was adopted, which evaulate the model several thousand times. Therefore this calibration process takes days to weeks on a single workstation. To speed up the processing JAMS has been thread-based parallelized. This thread-based parallelization is used for example to evaluate several parameter combinations in parallel during optimization. The model calibration run frequently, several times a week.
- HAPPI
  - HAPPI is a web service which has run successfully on the EGI federated cloud, with limited requirements: VMs with 2 cores, 4 GB of RAM plus block storage access
- INERTIA
  - The INERTIA Use case comprises the following applications:
  - Use Case 1: Selection of the optimal portfolio on an emergency grid operation - Congestion issues are related with the power that flows among the lines or the transformers of a power system. The power flows among the different network's equipment must be maintained within acceptable operational limits (equipment's thermal limits) in order to prevent equipment failures. In emergency situation a selection of optimal local hub can be used to keep the grid stable.
  - Use Case 2: Selection of the optimal portfolio on a non emergency grid operation - This use case describes the situation where the total Consumption of an Aggregated portfolio should be optimized based on the reception of data from market operations. This can be the prediction of tariff information, the operation of Reserve Markets.
  - Use Case 3: Monitoring of energy data (Aggregator Hub installation) - A rich monitoring tool is the prerequisite to determine state and performance of the total hub portfolio. It has to collect measured data on different time-scales from DERs, Building Energy Management Systems (Local Hubs) and also manage the regulation of data provision with a large variety of graphical data representations. Different monitoring approaches are delivered for the available data depending on the time horizon.

    - Use Case 4: Setting User Preferences - Final Occupants will have the capability of explicitly setting their preferences through a personalized Ambient UI. Moreover, they will have full control over DERs within their personalised working area, using traditional controls (light switches, HVAC panels etc.). If they feel that their preferences are violated, any automated control will be postponed, informing the system of the corrective action performed by the user and the potential discomfort caused.
    -

- DRIHM
    - The appliances are available in the EGI AppDB, they include a contextualisation tool installed in the image.
    - Context. tested in OpenStack site: update on OCCI-OS needed
- BILS
    - The application has been tested in the EGI federated cloud, and there is interest to continue using cloud resources for the computational task of their platform.

Non yet in fedcloud

- Human Brain Project
    - Currently the UBP applications are not using distributed e-infrastructures, but there are already datasets available in Europe and beyond, to be federated.

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

- Chipster
    - Complex deployment through contextualisation
    - Shared block storage exported as NFS up to 1 TB
- READemption
    - High capacity VMs
    - Block storage capabilities
- JAMS
    - Contextualization, heavy computation
- HAPPI
- INERTIA

- Backend services hosted in federated cloud services. Client services are mobile applications.
- VM: 2 cores, 4 GB of RAM
- Tomcat App server, Block storage

- DRIHM
  - Tenant configured to start only 1 instance of the VM image, this requirement is due to limited amount of licenses. The distributed cloud should be able to manage licenses per user community.

- BILS

Non yet in fedcloud

- Human Brain Project
  - A multi-terabyte storage capacity. Each image will typically range from 1-10TB.
  - A compute node with fast IO bandwidth to storage device (to be specified shortly)
  - The ability to deploy a Python-based service (BBIC, see appendix I) and supporting libraries (HDF5, etc).
  - A standardized authentication/authorization/identity mechanism.
  - Interactive access to dataset
  - Federation of sample neuroscience-based volumetric datasets provided by HBP, OpenConnectome, Allen Institute and others.
  - A multiprocessor compute node with high speed access to the storage device.
  - Transfer rate of 1GB/sec per repository

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.

- Chipster
- READemption

    - The READemption community already have a computational infrastructure which covers most of the capacity requirements of the community but it cannot cope peaks of demand. To fulfill the requests in such situations the community used Amazon to manage outburst computing. The usual workflow: create some instances, upload the data, run the application, get the result and kill the VM, usually lasts 1 or 2 days. As an alternative to commercial clouds the community as used the EGI federated cloud platform.
- JAMS
- HAPPI
- INERTIA
    - The use case run in the CESNET Fedcloud site in the first deployment, test, phase.
- DRIHM
- BILS
    - Production quality Object Storage
    - Hundreds of GB to be shared between different VMs
    - Tools to exploit cloud elasticity
    - JAVA OCCI API

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

All the use cases described in this annex are linked, if not already deployed, with the EGI federated cloud. Being EGI a mature European-wide e-infrastructure, the availability of services and capabilities in the federated cloud opens access to a big use base across several countries.

- Chipster
    - While Chipster has become very popular, many users and institutes in Europe are still struggling to set up their own server as they lack a suitable computing platform. Ability to launch Chipster easily in the cloud is critical to solve this problem.
- READemption
- JAMS
- HAPPI
- INERTIA

- DRIHM
- BILS

From the EGI Competence centres:

- BBMRI-ERIC CC
    - BBMRI CC will exploit how to enable research consortia to implement integrated data analysis across BBMRI biobanks, central public repositories and their own 'study' data. Therefore they will deploy tools to support heterogeneous data processing workflows in the EGI Cloud. This includes Virtual Machines (VMs) with tools for data staging (e.g. B2STAGE, B2SHARE from EUDAT), for connectivity to relevant data providers (ranging from de-centralized data with access restrictions such as individual biobank data to data from central repositories such as ENA/EGA) and for platform facilities executing analysis workflows (e.g. BiobankCloud).

- EPOS
    - The direct communication channel between the related working groups and the EPOS CC is of major importance. This is established through EPOS CC partners involved in EPOS, but the EPOS CC will also organize a proactive dissemination strategy to provide application adapted information about the capabilities, which may be available by the EGI infrastructure. Another main focus will be laid on the exploration of dynamic Cloud services from applications used in the ESFRI and providing them on discipline specific marketplaces.

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.*	• *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:*

	Use of virtualized GPU or interconnection (containers). Implementation of elastic scheduling on IaaS platforms.	**<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.

<input here>

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.)

<input here>

3.3 General description of datasets/information used

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

<input here>

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.

<input here>

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>

4 DATA LIFE CYCLE

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

9.1 Storage

- Reliable and efficient mechanism to share data
- Block storage:
  - Accessible only from within a VM (in the same site)
  - Data sharing not possible
  - High performance: low latency
  - Easy usage and integration
- Object storage:
  - Accessible via REST API
  - Allow data sharing
  - Requires a client to be integrated within the application
- Satisfy both high performance and shared access

9.2 PaaS

Abstract IaaS complexity to end users:

- Setup complex deployment
- Exploit cloud elasticity →horizontal scalability
- Simply deployment of web services
- Automatic scalability
- Monitoring

Support a set of PaaS/Orchestrators

- Preferably standard based
- Avoid technology lock-in

PaaS for specific communities integrated with Indigo services?

- Arvados from Curoverse for BioInformatic

9.3 Data management

Data aware brokering

- Move computation close to the data
  - Dataset registry
  - Location, replicas and access permission rules (e.g. VOs enabled)
- Tools to start VMs close the data
  - Infrastructure topology hidden to end users
  - VM images with ad-hoc sw pre-loaded
- Related activity in EGI - AppDB
  - Dataset registry under development
  - Life Science data replication VT

- Brokering in the roadmap

Data replication

- Move data close to computation

Datasets stored in community repository (ES..,EIDA) or other e-infrastructure (e.g. EUDAT)

- Data Caching service

Create replicas of subset of datasets in the EGI infrastructure

Interfaces towards external repositories

Make the process transparent to the end users

9.4 Network

- Build rich networking topologies
  - Added value for end users
- Configure advanced network policies in the cloud
  - QoS
  - LB-aaS, VPN-aaS, firewall-aaS, etc.
  - monitoring
- Extend current standard interfaces to support network configuration

10 REFERENCES

R 1	
R 2	
R 3	
R 4	
R 5	