INDIGO-DataCloud

INITIAL REQUIREMENTS FROM RESEARCH COMMUNITIES

EU DELIVERABLE: D 2.1

Document identifier:	INDIGO-WP2-D2.1-V21
Date:	**12/07/2015**
Activity:	**WP2**
Lead Partner:	**EGI.eu**
Document Status:	**RELEASED**
Dissemination Level:	**PUBLIC**
Document Link:	

Abstract

This report summarizes the findings of the WP2 tasks on “Research Communities Requirements” and “Defining support to Research Data” along the first three months of the project, providing a key input to the architectural design of INDIGO-DataCloud solutions. This integrated document includes a detailed description of the Research Communities involved and of the Case Studies proposed, including Data Management and Computational Intensive issues. Based on the inputs provided by the WP2 partners, an initial list of the requirements for each application has been collected, that has been used in turn to assemble a list of common requirements that has been provided as input to the discussion on the technical Architecture of INDIGO-DataCloud solutions.

I. COPYRIGHT NOTICE

II. DELIVERY SLIP

	Name	Partner/Activity	Date
From	Peter Solagna	WP2/NA2/EGI.eu	10 July 2015
Reviewed by	**Moderator: Jesús Marco** **Reviewers:** Fernando Aguilar, Ignacio Blanquer, Sandro Fiore, Massimiliano Rossi	CSIC, UPV, CMCC, INGV	14 July 2015
Approved by	**PMB**		16 July 2015

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 from partners on structure included, v02	F.Aguilar CSIC
3	18-may-2015	Draft to be discussed with each community, v03	P.Solagna, EGI.eu F.Aguilar, CSIC M.Rossi, EMSO
4	22-may-2015	Draft with initial community input, to be iterated with JRA, v04	P.Solagna, EGI.eu F.Aguilar, CSIC M.Rossi, EMSO
5	26-june-2015	Draft revised with updated inputs from user communities, v05	P.Solagna, EGI.eu F.Aguilar, CSIC Y.Chen, EGI.eu
6	30-june-2015	Draft to be circulated for internal review, v06	P.Solagna, EGI.eu
7	3-july-2015	Tables of Requirements included v07	I.Blanquer, EGI.eu
8	6-July-2015	Methodology section included v08	Yin Chen, EGI.eu
9	10-july-2015	Sections completed, version ready for internal review v19	P.Solagna, EGI.eu F.Aguilarf, CSIC
10	12-july-2015	Revised version v20	J.Marco, CSIC
11	14-july-2015	Reviewed version v21, released to PMB	P.Solagna EGI,eu, J.Marco, CSIC, S.Fiore CMCC, Massimiliano Rossi, INGV

TABLE OF CONTENTS

1 EXECUTIVE SUMMARY

This report summarizes the findings of the WP2 tasks on "Research Communities Requirements" and "Defining support to Research Data" along the first three months of the project, providing a key input to the architectural design of INDIGO-DataCloud solutions.

A list of Case Studies has been proposed and analyzed in detail by the different Research Communities on the basis of the interest on INDIGO-DataCloud solutions:

Research Community	Case Study/Application	
LifeWatch	**Monitoring and Modelling Algae Bloom in a Water Reservoir**	
	TRUFA (Transcriptomes User-Friendly Analysis)	
EuroBioImaging	**Medical Imaging Biobanks**	
INSTRUCT	**Molecular dynamics simulations**	
LBT CTA	**Astronomical Data Archives** **Archive System for the Cherenkov Telescope Array**	
WeNMR	**HADDOCK portal**	
ENES	**Climate models inter comparison data analysis**	
Galleries, Libraries, Archives, Museums	**eCulture science Gateway**	
ELIXIR	**Galaxy as a Cloud service**	
EMSO	**MOIST-multidisciplinary oceanic information system**	
DARIAH	**Big Data in Arts and Humanities**	
EGI Virtual Teams Competence Centres	**Chipster**	**BILS**
	READemption	**Human Brain Project**
	JAMS	**BBMRI-ERIC CC**
	HAPPI	**DARIAH CC**
	INERTIA	**EPOS CC**
	DRIHM	**Disaster Mitigation**
	CANFAR	**LoFAR**

A complete list of requirements per Case Study has been compiled internally from the detailed Annexes describing them that have been provided by the different Research Communities.

This long list with more than 100 requirements, classified by type (Computational/Storage/PaaS) and rank (mandatory/convenient/optional) has been used in turn to extract a short list of common requirements, classified into three categories:

Category A: Computational Requirements

Category B: Requirements linked to Storage

Category C: Requirements on Infrastructure

This list of requirements is the main outcome of this deliverable, and it is provided as input to the work of the JRA teams within INDIGO-DataCloud project. It was used as starting point in the f2f meeting beginning of July in Valencia oriented to define the INDIGO Architecture.

The list is enumerated below, and details are provided in the corresponding section of this deliverable.

#REQ	Description
CO#1	Deployment of Interface SaaS
CO#2	Deployment of Customized computing back-ends as batch queues
CO#3	Deployment of user-specific software
CO#4	Automatic elasticity of computing batch queues
CO#5	Terminal access to the resources.
CO#6	Privileged access
CO#7	Execution of workflows
CO#8	Provenance information
CO#9	Cloud bursting
CO#10	Data-aware scheduling
CO#11	Provisioning of efficient Big Data Analysis solutions exploiting server-side and declarative approaches
CO#12	Execution across multiple centres.
CO#13	On-line processing of data
CO#14	Special hw configuration - MPI, multicore, GPGPU
SO#1	Shared storage accessible like a POSIX filesystem
SO#2	Persistent data storage
SO#3	Long-term availability of results
SO#4	Local user storage
SO#5	Availability of reference data
SO#6	Interoperability with application domain specific software and services (e.g. IS-ENES/ESGF)
SO#7	Metadata management / Database as a Service
SO#8	Share data capabilities
SO#9	Data replication
SO#10	Distributed storage
SO#11	Dropbox-like storage
PL#1	Global-level AAI
PL#2	On-line access to data
PL#3	Network configuration
PL#4	Monitoring and operation

When considering the Use Cases, it was considered important to remark the need to include different roles (final user, developer, managers) in their analysis and to try to find general solutions to support common requirements. Two such complete generic solutions (User Community Computing Portal Service and User Community Computing Portal) are also introduced in this deliverable to trigger the discussion with the JRA teams.

An initial set of additional questions have been already posed back by JRA teams, that will be in turn considered and answered by the research communities, providing an initial refinement of the requirements. This work will be done along the line agreed of supporting an Agile development method, so that the current versions of the Annexes of the different Case Studies will be updated and improved along the next months.

2 INTRODUCTION

The INDIGO-DataCloud project was designed with a clear objective in mind:

"The proposal is oriented to support the use of different e-infrastructures by ***a wide-range of scientific communities, and aims to address a wide range of challenging requirements posed by leading-edge research activities conducted by those communities.*** *Indeed, the force driving this proposal is the interest of these communities and the organizations supporting them, from many different fields in science, from biomedicine to astrophysics, from cultural heritage to climate, participating in very relevant initiatives at European level, such as INSTRUCT, ELIXIR, EMSO, DARIAH, LIFEWATCH, etc.*

The tasks within WP2/NA2 are oriented to assure that this objective is achieved. In particular tasks *T2.1, Research Communities Requirements* and *T2.2, Defining support to Research Data*, should provide a key input to the architectural design of INDIGO-DataCloud solutions.

Among the subtasks indicated in the proposal within T2.1, this deliverable addresses:

-Analyse the use cases proposed by the communities participating to the consortium. Capture the requirements for efficiently running the applications and workflows on Cloud, Grid or HPC infrastructures.

-Capture requirements generated by user communities not part of the project (such as the EGI Federated Cloud users), which are relevant for the outputs of the project

and presents the outcome of a third subtask:
-Produce an integrated document with the requirements captured, prioritized and grouped by technical areas

Task 2.1 will continue along next months, in particular also including the subtask related to the interaction with other parallel projects that are starting recently[1],

-Liaise with the INFRADEV-4 projects to enable synergies between the projects, and interoperability between the INDIGO outputs and the VRE to be deployed by the E-INFRA-9 projects.

Notice also that this deliverable D2.1 will be updated in Month 9 by deliverable D2.4, on *"Confirmation of support to initial requirements from JRA design and extended list of requirements"*

Regarding Task 2.2, it shall "*undertake a survey on the research communities to collect and analyze the individual Data Management Plans (DMP) and data-life-cycle documentation with the aim to ensure that the full data cycle and components will be supported in INDIGO, and with the aim to provide adequate specifications for the compliance with INDIGO.*"

This deliverable presents also the outcome of this first (but quite detailed) survey, considering in particular the following subtask:

-Development of individual search activities to acquire and analyze the available DMP of the research communities/infrastructures with special attention to distributed/heterogeneous data services and catalogues, and to available open data.

[1] *Notice that a Concertation meeting is scheduled on the week starting 23rd November*

It is not an unexpected surprise the wide variety observed in the results of the survey, and certainly additional iteration is needed before addressing the two other subtasks:

-Acquisition of procedure details/parameters (i.e., DMP, Collection, Authenticity & Provenance, Data Preservation) to elaborate the specifications for data ingestion and use in INDIGO

-Definition of the specifications of INDIGO ingestion integrity test

that will provide the input to deliverable D2.7 "*Specifications of data ingestion and use in INDIGO"*.

The structure of this report is the following one:

-First of all a brief summary about the Research Communities participating in INDIGO is presented in section 3, to introduce them in the context of the project objectives.

-The following section 4 presents the methodology used to gather the requirements of the different Research Communities, and in particular the design of the template used.

-Section 5 summarizes the Case Studies reported in the different Annexes.

-Section 6 indicates the specific requirements extracted within WP2/NA2 from each Case Study, which are reported complete in the Appendix (section 9) and **proposes a set of common requirements that are the main outcome of the deliverable, and are provided as input to JRA teams to be taken into account along the definition of the INDIGO-DataCloud Architecture**.

-Section 7 provides two examples of generic solutions based on the analysis of the previous Case Studies.

-And section 8 summarizes the next steps for the refinement and improvement of these requirements, including in particular the interaction with JRA team.

-Finally, the appendix (section 9) explicitly details the list of requirements gathered.

Additionally, the Annexes provide the complete input on the different Case Studies from each Research Community. They can be found at https://grid.ifca.es/wiki/INDIGO/WP2/D2.1 and in the WP2/NA2 Wiki space within INDIGO-DataCloud OpenProject.

3 RESEARCH COMMUNITIES

Eleven different Research Communities represented by corresponding partners are participating in WP2/NA2 in INDIGO. The "SIMPLIFIED IMPACT TABLE" prepared for the proposal is presented below indicating the different areas covered (Life Sciences, Physical Sciences and Astronomy, Social Sciences and Humanities, and Environmental Sciences), the Research Communities involved, examples of the different applications to be supported, and the potential impact of INDIGO solutions.

SIMPLIFIED IMPACT TABLE **SELECTED OBJECTIVES versus REQUESTS/ POTENTIAL IMPACT FOR COMMUNITIES** **O1: Development of the INDIGO Platform based on open software without restrictions on the e-Infrastructure**	**Life Sciences**	**Physical Sciences & Astronomy**	**Social Sciences & Humanities**	**Environmental Sciences**
Research Communities & Initiatives , including ESFRIs	**ELIXIR INSTRUCT/ WeNMR EuroBioImaging**	**CTA LBT WLCG**	**DARIAH DCH-RP**	**EMSO LIFEWATCH ENES**
Examples of Applications	**HADDOCK GROMACS AMBER GALAXY**	**MIDAS, IRAF, IDL, Geant4 ROOT/PROOF Geant4**	**Fedora Digital Libraries**	**Delft3D R-Studio TRUFA MATLAB**
Design and development of a Platform providing advanced users and community developers a powerful and modern environment for development work. This includes programming and scripting tools, and composition of custom applications and software deployment	RELEVANT	**CRITICAL**	RELEVANT	**CRITICAL**
Developing a framework to enable the transparent execution on remote e-infrastructures of existing popular applications like MATLAB / OCTAVE, ROOT, MATHEMATICA, or R-STUDIO.	RELEVANT	**CRITICAL**	MINOR	**CRITICAL**
Provide the services and tools needed to enable a secure composition of services from multiple providers in support of scientific applications.	**CRITICAL**	**CRITICAL**	RELEVANT	RELEVANT
Develop and implement a solution that is able to deploy in a transparent and powerful way both services and applications in a distributed and heterogeneous environment made by several different infrastructures (EGI Grid and Federated Cloud, IaaS Cloud, Helix Nebula, HPC clusters)	**CRITICAL**	RELEVANT	MINOR	RELEVANT
Develop the capability in the PaaS to provide unified data access despite geographical location of data, including APIs access, based on existing standards, or virtually mount like a POSIX device to worker node, cloud virtual machines, personal computer etc.	**CRITICAL**	RELEVANT	**CRITICAL**	RELEVANT

Table 1: Simplified Impact Table as presented in INDIGO proposal

The distribution of the INDIGO partners is shown on top of the ESFRI map in the following figure:

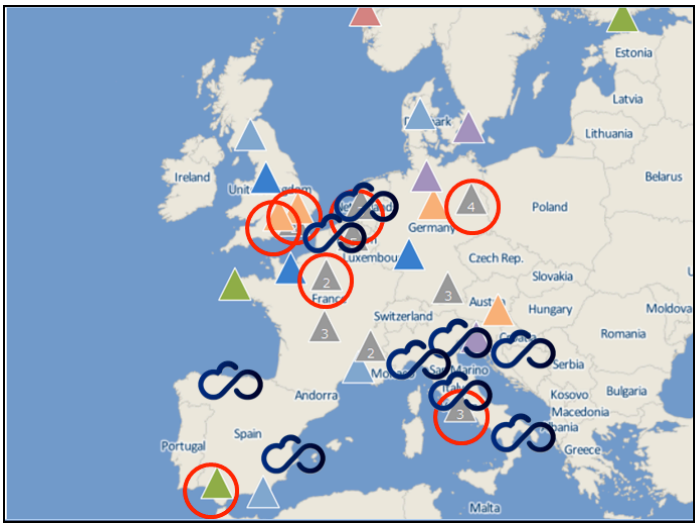


Figure 1: Distribution of WP2/NA2 Partners. Related ESFRIs are circled in red

The following table provides the relation of Research Communities ordered by the participating partner, and including the contacts that are directly contributing to INDIGO-DataCloud WP2/NA2

#	Partner	Research Community	Area/Type	Contact
P0	CSIC	**LifeWatch**	Environmental ESFRI	Jesús Marco Fernando Aguilar
P1	UPV	**EuroBioImaging**	Biological and Medical Sciences ESFRI	Ignacio Blanquer
P2	CIRMMP	**INSTRUCT**	Biological and Medical Sciences ESFRI	Antonio Rosato
P3	INAF	**LBT** **CTA**	Physical Sciences ESFRI	Cristina Knapic Riccardo Smareglia Stefano Gallozzi L. Angelo Antonelli
P4	U. Utrecht	**WeNMR**	Biological and Medical Sciences Virtual Research Community	Alexandre Bonvin
P5	CMCC	**ENES**	Climate and Earth System Modeling	Sandro Fiore Giovanni Aloisio
P6	ICCU	**G**alleries, **L**ibraries, **A**rchives and **M**useums	Social Sciences and Humanities Cultural Heritage	Sara Di Giorgio Antonio Davide Madonna
P7	EGI.eu	**EGI** **Virtual Teams** **Competence Centres**	All Areas	Peter Solagna Yin Chen
P8	CNR	**ELIXIR**	Biological and Medical Sciences ESFRI	Federico Zambelli
P9	INGV	**EMSO**	Environmental Sciences ESFRI	Massimiliano Rossi Laura Beranzoli Manuela Sbarra
P10	RBI	**DARIAH**	Social Sciences and Humanities ESFRI	Eva Cetinic Karolj Skala

Table 2: Research Communities participating in INDIGO-DataCloud WP2

The following information about the Research Communities is a summary of the complete information that has been provided directly by these partners, through the Annexes. It is compiled here as considered useful in particular for internal communication.

3.1 Biological and Medical Sciences

3.1.1 EuroBioImaging

EuroBioImaging addresses population imaging: the large-scale acquisition and analysis of medical images in large human cohorts. Population imaging research community is focusing on building an overarching infrastructure for integrated data management and large-scale analysis of medical imaging data linked to clinical records. Such infrastructure will have sites providing open access to their data storage capacity to test a uniform image archiving and analysis infrastructure which may cross-link population imaging studies in different research centres. Projects will focus on the requirements for common data acquisition, storage, exchange, and analysis.

Notwithstanding the relevance of automatic image recognition, the advance in clinical radiology is not yet significant enough. Three fundamental barriers arise. The first such barrier is the **lack of access to medical imaging data**. Access to Picture Archiving and Communication System-PACS data has been severely restricted due to ethical-legal statues. Consequently, while other disciplines train algorithms on databases of over million images, the overwhelming majority of contemporary medical CAD publications are created upon data sets of fewer than 100-200 samples. The fragmented nature of medical imaging data also restricts re-use, not only of actual data sets but also of results, thereby limiting scientific progress to the efforts of an isolated research group. In addition, the lack of access to standard validation sets and inter-study comparison tools yields inefficiency in resource allocation which in turn may represents a critical barrier to entry for small research teams.

A second significant barrier to research involves the **challenge of managing available data**. Despite of the standards, image archives are subject to various proprietary techniques, making it cumbersome to consolidate data from several sources. Moreover, the requirements for the huge size of radiology examinations cannot be met by many small to medium-sized research entities.

A third barrier is encountered once the medical informatics researcher has retrieved the data. Extensive **cross-disciplinary collaboration** with a radiologist is necessary to identify the content in radiographic data, construct experiments and validate results. Such collaboration is often not feasible in current research environments.

Therefore, the community needs data repositories connected to effective processing infrastructures that could run reference or customised software pipelines as well as visualizing the results. For this purpose, the BIM-CV (Medical Imaging Biobank of the Valencia Region, BIM-CV for its initials in Spanish) consortium is a positively evaluated candidate node to EuroBioImaging ESFRI that comprises The Polytechnic University Hospital La Fe (HUPLF-https://www.acim.lafe.san.gva.es/), the Valencia Regional Health Authorities, the Valencian Network of Biobanks (RVB-http://grupos.fisabio.san.gva.es/web/rvb) and the Polytechnic University of Valencia (www.upv.es). BIM-CV is in the process of signing a Joint Research Unit Agreement for regulating their participation in projects. The main mandate of BIM-CV is to create such population imaging node.

3.1.2 INSTRUCT

Structural biology deals with the characterization of the structural (atomic coordinates) and dynamic (fluctuation of atomic coordinates over time) properties of biological macromolecules and adducts thereof. The dynamic properties of these systems are crucial to many aspects of their biological function, such as recognition of molecular partners or diffusion of small molecules (substrates, products, inhibitors) to/from the active site of catalytic machineries. These properties are hard to characterize experimentally in a direct and comprehensive (i.e. for all atoms) manner at the atomic level. Consequently researchers in the field largely rely on computer simulations to tackle the dynamic

aspect of structural biology. Simulations can be validated by comparison to different types of experimental data.
There are several bottlenecks for the community to benefit of state-of-the-art simulations of molecular dynamics, like complex tools, advanced computational power not always available (GPGPU, MPI libraries) or large output datasets.
CIRMMP and U. Utrecht have been involved in addressing the above points, through the implementation of dedicated web portals to run short simulations on a grid computational infrastructure.

3.1.3 WeNMR

Protein interactions that are critical to all cellular processes establish an intricate and dynamic molecular network – the interactome – in which subtle miscommunications often result in disease. The large gap between the number of interactions and available experimental 3D structures calls for complementary computational methods to produce accurate predictions and guide experimentalists. This is the field of computational structural biology, which has seen in the last decade fascinating developments both in software and hardware. Computational structure prediction is nowadays routinely considered an integral part of research. The docking field, in particular, has thrived in the last decade since the beginning of the CAPRI (Critical Assessment of PRedicted Interactions) experiment, in which the participants are asked to predict the structure of an unknown biomolecular interaction. Computational modelling of complexes has grown into a well-accepted complementary method to classical experimental techniques.
P10, U. Utrecht, is the main developer of HADDOCK. The Bonvin group is developing and distributing the software (http://bonvinlab.org/software), and also operating the web portals (both local versions running on the group computing resources and the grid-enabled version offered via the **WeNMR VRC**.
End users of HADDOCK consist of a large scientific community worldwide with difference backgrounds and expertise, ranging from bachelor students to experienced researchers and even commercial companies. They mostly interact with the web portal front end and/or run a local version of the software.

3.1.4 ELIXIR

ELIXIR unites Europe's leading life science organizations in managing and safeguarding the massive amounts of data being generated every day by publicly funded research. It is a pan-European research infrastructure for biological information.

ELIXIR will provide the facilities necessary for life science researchers - from bench biologists to cheminformaticians - to make the most of our rapidly growing store of information about living systems, which is the foundation on which our understanding of life is built. One of the services provided by ELIXIR is computing and also tools to take advantage of that, like Galaxy, an open-source web-based tool to perform biomedical researching.

3.2 Social Sciences and Humanities

3.2.1 DARIAH

Digital research methods have recently started to enter the mainstream of humanities, arts and social sciences research. Digital humanities have existed for years as a specialized field but the recent growth in the number of centres and research projects associated with digital methods in arts and humanities (A+H) and social sciences indicate that we are at a fundamental shift. The digital arts and humanities are at a critical point in the transition from a specialty area to a full-fledged community with a

common set of methods, sources of evidence and infrastructure. All of these are necessary for achieving academic and data driven scientific recognition. Information and data-intensive, distributed, collaborative and multidisciplinary research is now the norm in many scientific areas, but they are still in an experimental phase in the arts and humanities research community. However, the art and humanities disciplines nowadays generate and analyze an increasing amount of data and show great potential for growth and evolvement of new technologies. Research process in the A+H become more and more data-intensive and therefore have to be supported by emerging research infrastructures. Also, a vast array of collaborations arises in the digital humanities across Europe in the form of spontaneously funded research networks and associations. What is lacking, however, is an infrastructure that would ensure that the state-of-the-art of these collaborations is preserved and integrated, and that common best practices and technological standards are followed.

3.2.2 Galleries, Libraries, Archives and Museums

The Central Institute for the Union Catalogue for the Italian Libraries and for Bibliographic Information (ICCU) promotes and coordinates cataloguing and documentation activities of the library heritage.
InternetCulturale (a project promoted by ICCU) is a portal that provides analysis of cultural issues through multimedia resources (itineraries, exhibitions, authors and works, 3D programmes) devoted to literature, science, art and music.
The digitized material is produced by its partners (composed by a wide library community) and it is related with metadata according to national and international standards.

The Institutions partners of InternetCulturale manage their own archive (usually books, but there are also manuscripts, music, periodical and so on) composed by digital resources and local metadata (MAG, METS, PREMIS). The metadata are ingested by InternetCulturale via OAI-PMH. So, the digital objects remain in the native archives, even if they are showed in InternetCulturale thanks a viewer. Moreover, the cultural institutions can take advantage of a service, the MAGTECA, a digital archive that provides a free management and preservation of digital collections ingested (with a web resolution) with related metadata (MAG).

Considering the current development of InternetCulturale (and MAGTeca service), there are some important challenges about, security, speed, migration to new formats, storage and preservation.
InternetCulturale provides the publication of librarian material into a unique access point (www.internetculturale.it/opencms/opencms/it/) structured in OPAC SBN (catalogue researching), Manusonline, Edit16, historical catalogues, the Digital Library (which also contains the MAGTeca) and the website.
The research provides faceted browsing, that allows to filter the results. The search engine works on the Index, accomplished through a data extraction process from the original databases and the results are showed thanks a shared profile based on the Dublin Core standards (and its extentions) and finally the data are indexed.

Moreover, the MAG descriptive metadata allows a more detailed response. Indeed, the presence of ontologies allows the semantic expansion of queries and the automatic or semi-automatic identification of related terms and proposing suggestions within the detailed template associated with the selected object.

3.3 Environmental and Earth Sciences

3.3.1 LifeWatch

LifeWatch is the european e-science infrastructure (ESFRI) for biodiversity and ecosystem research. It aims to provide advanced capabilities for research on the complex biodiversity system and also to provide answers to policy problems that impact directly to our lifestyle. LifeWatch will take advantage of external resources from other initiatives, infrastructures, projects, distributed centers and single research groups. The capabilities offered by the LifeWatch, as an e-Science infrastructure, allow users to tackle the big basic questions in biodiversity, as well to address the urgent societal challenges concerning biodiversity, ecosystems and other crosscutting issues.

3.3.2 ENES

The European Network for Earth System modelling, ENES, was launched in 2001. Upon the establishment of the network and in the following years, several institutions including university departments, research centres, meteorological services, computer centres, and industrial partners, agreed to work together and cooperate to discuss strategies to accelerate progress in climate and Earth system modelling and understanding.
ENES aims to:

1. **help** in the **development** and **evaluation** of **state-of-the-art climate and Earth system models**,
2. **facilitate** focused **model intercomparisons** in order to assess and improve these models,
3. **encourage exchanges** of software and model results,
4. **help** in the **development of high-performance computing facilities** dedicated to long high-resolution, multi-model ensemble integrations.

As reported above, a major challenge for this community is the development of comprehensive Earth system models capable of simulating natural climate variability and human-induced climate changes. Such models need to account for detailed processes occurring in the atmosphere, the ocean and on the continents including physical, chemical and biological processes on a variety of spatial and temporal scales and how interactions between these processes can be perturbed as a result of human activities.

The development and use of realistic climate models requires a sophisticated software infrastructure and access to the most powerful supercomputers and data handling systems (data processing, analysis, archiving, sharing, visualization, preservation, curation, and so on).

In such a context, large scale global experiments for climate model intercomparison (CMIP) have led to the development of the Earth System Grid Federation (ESGF) a federated **data infrastructure** involving a large set of data providers/modeling centres around the globe (the IS-ENES project provides the European infrastructure for ENES, as well as the European contribution to ESGF). ESGF provides support for search & discovery, browsing and access to climate simulation data and observational data products.

3.3.3 EMSO

EMSO is a large-scale European Research Infrastructure in the field of marine environmental sciences supported by EC through ESFRI- European Strategy Forum on Research Infrastructures and by 10 Member States. EMSO is based on a European-scale network of fixed-point seafloor and water column observatories with the basic scientific objective of long-term monitoring, mainly in real-time, of processes related to the interaction between the geosphere, biosphere, and hydrosphere, including natural hazards. EMSO is a geographically distributed infrastructure composed of several deep-sea monitoring systems deployed on specific sites in the European seas, from the Arctic to the Black Sea

through the Atlantic Ocean and the Mediterranean, thus forming a widely distributed pan-European infrastructure.
MOIST is a data management system presently in use within EMSO for some multi-parametric observatories focused on standards, open accessibility and web services. This web interface allows discovering, visualizing and downloading the metadata and data related to seafloor campaigns with GEOSTAR-type observatories from 1998 to present. The datasets discovery can be done through the navigation menu (in particular by sites, projects, instruments).

3.4 Physical Sciences

3.4.1 CTA

The Cherenkov Telescope Array (CTA) is a large array of Cherenkov telescopes of different sizes and deployed on an unprecedented scale. It will allow significant extension of our current knowledge in high-energy astrophysics.
The CTA observatory will impose a new model to the very-high energy (E>1Tev) gamma ray astronomy data management. The CTA community will create a huge array of ~100 telescopes in two different sites (one in the northern hemisphere one in the southern hemisphere). The CTA will offer worldwide unique opportunities to users with varied scientific interests and support a growing number of young scientists working in the evolving field of gamma-ray astronomy.
The CTA will, for the first time in this field, provide open access via targeted observation proposals and generate large amounts of public data, accessible using Virtual Observatory tools. The CTA aims to become a cornerstone in a networked multi-wavelength, multi-messenger exploration of the high-energy non-thermal universe. During the ongoing preparatory phase of the project, CTA Monte Carlo (MC) simulations campaigns are distributed on the grid via the EGI CTA Virtual Organisation.

3.4.2 LBT

The Large Binocular Telescope (LBT) is an optical telescope for astronomy located on Mount Graham in the Pinaleno Mountains of south eastern Arizona, and is a part of the Mount Graham International Observatory. The LBT community supports a joint project with a number of partners, including the Istituto Nazionale di Astrofisica, also partner of INDIGO. Large Binocular Telescope is equipped with several instruments able to investigate different subjects in different wavelengths and for very diverse scientific fields of investigation. Each instrument has its own peculiarities and capabilities and has to be controlled, configured and operated in particular way respect to the others. Starting from the diverse proposals, data produced by each instrument is stored in a permanent archive and first handling consists in data delivery to partner archival mirrors sites. Each partner institution manage both private and public data and after data storing, performs data reduction and calibration. After the first data reduction some other data manipulation and information extraction could be done locally by expert users via already existing tools or pipelines. Again data sharing and data delivery to the final users are required.

3.5 Other Research Communities

As already indicated, the objective is also to include requirements generated by other Research Communities that do not participate directly in the project but which could be relevant for the INDIGO-DataCloud project outputs. EGI.eu has established relevant contact with different Research Communities of different areas through the Virtual Teams and Competence Centers that are reported in the following sections.

4 METHODOLOGY USED

This section presents the methodology used to gather the requirements of the different Research Communities.

The process started with the development of a common template described below (see any of the Annexes *Pxx* for complete details), including different sections to compile the information that was considered to be relevant for INDIGO. The idea of a comprehensive template as well as the structure of this template was agreed in the f2f WP2/NA2 meeting in Lisbon in May, and the different communities completed a first version of their annexes along June. Along the biweekly teleconference meetings, the discussion on the different sections of the Annexes helped to prepare updated versions. These improved versions were then used to extract the specific and common requirements (see next section). The details are provided below.

4.1 Design of a template for Case Studies

In order to capture the community requirements in an efficient and systematic way, a common template was designed and has been used by the partners to provide this information.

The design of the template followed after an initial discussion at the kick-off meeting in Bologna about the well-known "impedance mismatch" between researchers and ICT teams, and the difficulties to translate the user needs into final requirements, even in an Agile framework like the one being used in INDIGO. The idea of **relying as much as possible on User Stories**, closer to the researcher language and needs, was preferred over the possibility of using directly a full framework, like the one implemented by OpenProject, to express the requirements and provide direct input to a backlog.

The design of the template uses a Case Study oriented approach. **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*. When collecting requirements, communities were informed to focus on Case Studies that are representative both of the research challenge and complexity but also of the possibilities offered by INDIGO-Data Cloud solutions on it.

The Case Study is 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 then tracked by the Backlog system from the OpenProject tool.

The template serves as a structured framework with guiding questions concerned by INDIGO development workpackages. The inquiries cover the following aspects:

- Executive summary on the case study, where a community is asked to give background information about research challenge, explaining the expectation for INDIGO technology and potential impacts the study case would deliver.
- Introduction to the research Case Study, where a community is asked to present the Case Study, community roles to be involved, key technological and e-Infrastructure requirements, and potential exploitations.
- Technical description of the Case Study. Questions in this section intend to drill down into technical details to understand the computational functionalities a community would need. In the

last section, the community was asked to describe the Case Study from the research point of view that they are familiar with. In this section, they are step-by-step guided to break the case study into user stories and think in more detail, in term of Use Cases, how those user stories could be achieved through computational functions, how those computational elements interact with each other via workflows, and what the deployment framework and scenarios could be. This last section was included to be able to take into account the ICT point of view of the Research Communities themselves, as for example in many ESFRI initiatives there is already an ICT team and the best way to assure that INDIGO solutions will be useful is to engage those teams.

- Data Life Cycle. This section is specially designed since INDIGO-DataCloud is a DATA oriented project. The community is asked to describe in details about their Data Management Plans, data & metadata model, and operations involved in data life cycle, e.g., data acquisition, data curation, data process, data analysis, data visualization, and data publication. Much of this section is based on the requirements posed to H2020 projects, and follows the recommendations of the DMP tools developed at the Digital Curation Center in UK and within the DataOne initiative in US.
- Intensive Computing. Initially labeled as Simulation/Modeling, since simulation/modeling is a computing intensive process, which is highly related to INDIGO technology, this section intends to inquire community requirement details about simulation packages, computing capacities needed, simulation workflows, and other aspects related to simulation/modeling. The section was extended and re-labeled to take into account other intensive computing processes, like for example post-processing and analytical tasks.
- 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. A community is asked to list use stories based on data collection, curation, processing, analysis, simulation, etc. that are considered most relevant for the Case Study being analyzed. In particular they are requested to 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, a community is asked to describe the scenario for Big Data analysis, and assure that the Use Case considers which levels of workflow engines are needed (e.g., “coarse grain”, targeting distributed (loosely coupled) experiments, through workflow orchestration across heterogeneous set of services; “fine grain”, targeting high performance (tightly coupled) data analysis through workflows orchestration on big data analytics frameworks).
- Infrastructure technical requirements. While previous sections focus on use cases, functional requirements and data model, this section focuses on e-Infrastructure resources and computing capacities, e.g., CPU, storage, and networking. It also includes the inquiry questions about AAI and (service) monitoring aspects.
- Connection with INDIGO solutions. This section is left open to include the ideas and feedback of technology providers, in particular the INDIGO JRA workpackages, to consider how the community requirements could be efficiently and effectively supported by INDIGO implementation.

As described above, by including all these aspects of the Case Study, like computational functionalities, data model, technology in use, and capacity to request, the design ensures that filling such template will gather sufficient information for the INDIGO implementations. The template has been reviewed and cleaned up internally by the WP2/NA2 team.
The template has also been reviewed and commented by INDIGO JRA leaders to assure that it

captures their expectations on the collection of requirements.

The template, as well as each of the Annexes, is considered to be a live document along the whole life of the project. In its current format (as a word file), versioning is used to track the changes and improvements.

4.2 Collection of Input from Research Communities

After obtaining the approvals within WP2/NA2 and from JRA leaders, the template was distributed to the Research Communities detailed in table 2, asking them to provide at least a relevant Case Study.
A brief summary of the context of these communities has been given in Section 3. All of them have provided the corresponding Annexes, and the list of Case Studies follows below:

#	Partner	Research Community	Case Study/Application	
P0	CSIC	LifeWatch	**Monitoring and Modelling Algae Bloom in a Water Reservoir**	
			TRUFA (Transcriptomes User-Friendly Analysis)	
P1	UPV	EuroBioImaging	**Medical Imaging Biobanks**	
P2	CIRMMP	INSTRUCT	**Molecular dynamics simulations**	
P3	INAF	LBT CTA	**Astronomical Data Archives** **Archive System for the Cherenkov Telescope Array**	
P4	U. Utrecht	WeNMR	**HADDOCK portal**	
P5	CMCC	ENES	**Climate models inter comparison data analysis**	
P6	ICCU	Galleries, Libraries, Archives, Museums	**eCulture science Gateway**	
P7	EGI.eu	EGI Virtual Teams Competence Centres	**Chipster**	**BILS**
			READemption	**Human Brain Project**
			JAMS	**BBMRI-ERIC CC**
			HAPPI	**DARIAH CC**
			INERTIA	**EPOS CC**
			DRIHM	**Disaster Mitigation**
			CANFAR	**LoFAR**
P8	CNR	ELIXIR	**Galaxy as a Cloud service**	
P9	INGV	EMSO	**MOIST-multidisciplinary oceanic information system**	
P10	RBI	DARIAH	**Big Data in Arts and Humanities**	

Table 3: List of Applications/Case Studies provided by the different Research Communities

This resulting collection of more than 20 Case Studies from representative communities from different scientific disciplines provides a relevant sample of Europe-wide requirements for INDIGO-

DataCloud. This information, including this deliverable, is public and can be shared with other technology providers that want to gain insights of communities' needs.

4.3 Identification the technology requirements and prioritization

The annexes provided by the user communities are very much use case focused. As described above, the user communities have been asked to report on their needs from a researcher perspective, starting from the topics in which they are more familiar, their research, their software tools, rather than requirements for generic cloud, computing or storage technologies.

The research use cases translate into the requirements for the INDIGO JRA work packages starting from the gaps identified in the use stories detailed in the annexes.

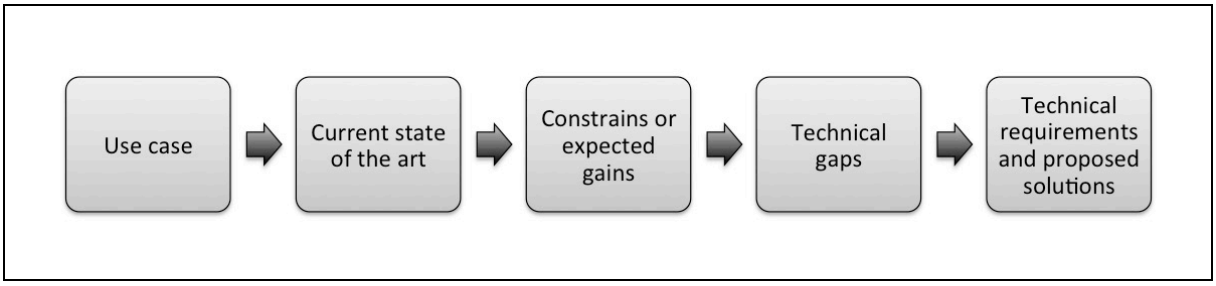


Figure 2: process of identification of the requirements from the user stories.

The process that led to the final set of requirements is as described in figure 2. The use case is the description of the workflow that the user needs to perform to do their research, the information include for example which type of data, where the dataset are hosted, which processing is performed on the input data and/or to generate the outputs and where the outputs have to be stored. The current state of the art is the way the workflow is currently implemented by the researchers, as reported in the annexes this includes a very diverse set of use of IT resources that goes from downloading locally the data to process in the personal computer and use its computing power, to using a computing cluster or even distributed e-infrastructures.

From the analysis of the current state of the art it has been possible to identify the constraints and the expected gains for the users, where they saw limitations in the work they are doing, and where chances of improvements are. The users either see limits in their current workflows or they need to increase their productivity, for example to analyse or produce bigger datasets, reduce waiting time, share their data or more in general collaborate with other researchers in a easier way. The specific constrains and limitations reported by the user communities have been grouped under a set of technical gaps, still linked to the user communities use cased but more focused on the actual technology used or potentially supporting the use case, one example could be "Data cannot be accessed efficiently in a concurrent fashion". These technical gaps or functional requirements are then translated in technical requirements that can be addressed by the INDIGO solutions.

The constraints, gaps, functional requirements, technical requirements and the proposed solutions have been summarized in tables to be easily presented to the rest of the INDIGO community. These tables are reported in the appendix (section 9).

Given the considerable number of requirements, WP2/NA2 produced a first attempt of prioritization, with an internal prioritization and an external prioritization.

Internal, means the prioritization done by the research communities themselves from their point of view. During the first analysis of the annexes the WP2/NA2 team – based on the description of the use cases - classified the requirements in three levels of priority ranking: Mandatory, Convenient and Optional. Where "Mandatory" means that without this development the use case cannot consistently improve ("must"), "Convenient" is a requirement that would further improve the workflow for the users, but it is less critical ("should"), while "Optional" means that the requirements is a nice-to-have but not critical ("can"). The user communities' contacts have then been asked to review the priorities and comment or modify them.

External means that the prioritization is done by looking at the complete picture of all the requirements produced by the communities. Every use case provided by the user communities identified several technical requirements, and the INDIGO solutions often fulfil several requirements, this already provided a first analysis of the importance of the developments planned in INDIGO, some solutions will support more use cases and therefore can be more relevant. This information will be used as an input in the prioritization process.

The contacts of the user communities have been asked to check and comment or edit the summary tables to confirm that the analysis reflects their actual status.

4.4 Interaction with the INDIGO JRA development work packages

The first target for the requirements gathered from WP2/NA2 is the INDIGO Project Management Board (PMB), where all the work packages are represented. The INDIGO technical architecture, the services and the components that will be offered to our users, will be defined within the PMB and the requirements are the first input for this design process, since the communities are the main stakeholders of INDIGO.

This communication is not, of course, a one-way channel. The development packages in particular, the JRAs, have to analyse the requirements as they have been reported by NA2, and identify the proposed solutions and the fit them in the architecture, but first the requirements must be clear and agreed. To reach the optimum level of understanding of the use case and the technical implications behind them, there will be likely need for some direct interaction between the developers groups in INDIGO and the communities, which in some cases will require several interactions, including dedicated calls.

To track the evolution of the requirements the individual technical requirements will be recorded in a requests tracker, or ticketing system, in order to easily record the questions/answers and the outputs of the calls. The whole consortium of INDIGO will be able to access these tickets, unless with the use there will be the need to limit access to it, and this will foster interaction and contributions to this requirements analysis process, and – most importantly – the dissemination of the requirements in the INDIGO community.

This tickets system will be provided as part of the OpenProject tool deployed by WP1.

This process of finalizing the requirements with the JRAs is starting in the moment of writing, and will continue during the development phase. Once the services and solutions will be available, the ticket will be associated to a particular release and closed, if no further feedback is required. It is more likely, though, that the users will feed feedback to the JRAs about the development producing requirements for improvements for a second release of the services.

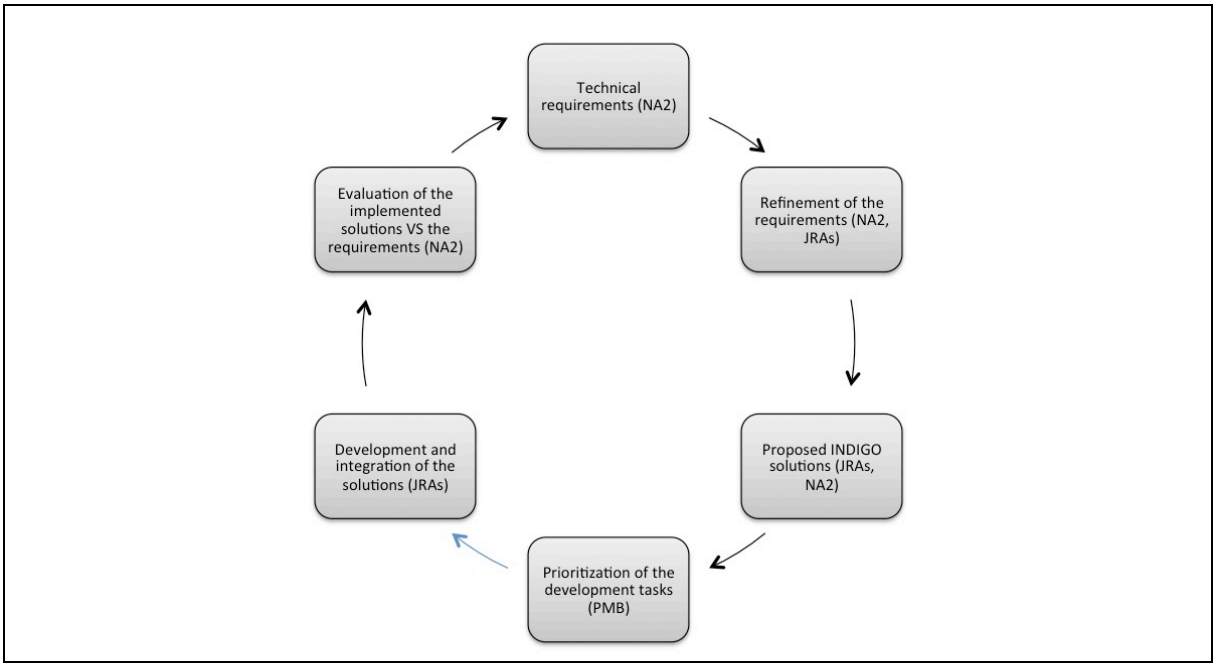


Figure 3: requirement definition and implementation cycle

As Figure 3 shows, we foresee a cyclic approach to requirements analysis. The beginning of the cycle is the first box where the technical requirements are defined within NA2, and then the requirements are refined and analysed with the JRAs work packages teams, to reach a set of proposed solutions for development. The development and integration tasks will be then prioritized by the INDIGO PMB, considering as inputs the results of the development/integration activities is then evaluated within WP2/NA2, tested versus the use cases and further feedback is provided in form of Technical requirements to improve the developments, and improve the software.

Using a tracker for this purpose will serve to (i) make sure that the right information is associated to the requirements, and (ii) track which requirements are progressing and which are not.

5 SUMMARY OF CASE STUDIES

This section provides a summary of the Case Studies proposed by the Research Communities. Additionally, the Annexes provide the complete input on the different Case Studies from each Research Community. They can be found at https://grid.ifca.es/wiki/INDIGO/WP2/D2.1 and in the WP2/NA2 Wiki space within INDIGO-DataCloud OpenProject.

An overview is provided in the following table:

# Partner Research Community	Case Study/Application
P0 CSIC **LifeWatch**	**Monitoring and Modelling Algae Bloom in a Water Reservoir** Support of hydrodynamic and water quality modelling including data input-output management and visualization.
	TRUFA (Transcriptomes User-Friendly Analysis) Deployment of a web-based workflow for RNA-seq performing including user management and computing of the different steps (with specific software), which requires specific hardware infrastructure in terms of memory or CPU. Data management is also part of the use case
P1 UPV **EuroBioImaging**	**Medical Imaging Biobanks** The virtual Biobank integrates medical images from different sources and formats. This case study includes all the steps needed to manage these images, like analysis, storage, processing (pre, post), etc. Privacy is a constraint to take in account for user management.
P2 CIRMMP **INSTRUCT**	**Molecular dynamics simulations** Support of Molecular Dynamics simulations of macromolecules that needs specific requirements in terms of computing (GPGPUs e.g.) using a pipeline of software that combines protocols that automate the step for setup and execution of these simulations.
P3 INAF **LBT** **CTA**	**Astronomical Data Archives** Data management and analysis using different tools such as data discovery, comparison, cross matching, data mining and also workflows. Basically the case study could be depicted as follow in some points with different requirements and users: data production, data reduction, data quality, data handling and workflows, data publication and data link to articles.
	Archive System for the Cherenkov Telescope Array (CTA) Data management, treatment and flow of data, big data archiving and processing, open data access.
P4 U. Utrecht **WeNMR**	**HADDOCK portal** Deployment and support of HADDOCK and the incorporation of a large variety of data from NMR and other biophysical methods.
P5 CMCC	**Climate models inter comparison data analysis**

ENES	Linked to the Coupled Model Inter comparison Project (CMIP).	
P6 ICCU **Galleries, Libraries, Archives and Museums**	**eCulture science Gateway** Digital repository collection support that allows users to upload, download digital documents and manage metadata, in collaboration with INFN Catania.	
P7 EGI.eu **Virtual Teams** **Competence Centres**	**Chipster**	**BILS**
	READemption	**Human Brain Project**
	JAMS	**BBMRI-ERIC CC**
	HAPPI	**DARIAH CC**
	INERTIA	**EPOS CC**
	DRIHM	**Disaster Mitigation**
	CANFAR	**LoFAR**
P8 CNR **ELIXIR**	**Galaxy as a Cloud service** Deployment of Galaxy instance that should support all the software/steps needed by the pipeline over, for example, a virtual cluster or cloud instances. Users also have to be managed in every instance.	
P9 INGV **EMSO**	**MOIST- multidisciplinary oceanic information system** Support for MOIST portal, which includes data storage and management (metadata, documents, etc.), web portal and applications (harvesting, data discovery, data visualization and analysis, etc.).	
P10 RBI **DARIAH**	**Big Data in Arts and Humanities** Strengthening the Use of Scientific Distributed Computing in the Arts and Humanities	

Table 4: Summary of Applications/Case Studies provided by the different Research Communities

Notice that the EGI Federated Cloud use case is not limited to a single case study, but includes a wide set of applications and use cases whose main need is the use of cloud computing.
The following sections provide more details about the different Case Studies.

5.1 LifeWatch

5.1.1 Algae Bloom

The researchers (typically biologists) want to use data collected in an instrumented platform in the water reservoir and also in some tributaries to:

- Monitor the evolution of the potential eutrophication of the water reservoir
- Use these data as input to a model (hydrological and biological), understand how well it compares with real data, and what are the main parameters that may affect the eutrophication evolution.
- Implement a predictive/alarm framework to enable warnings to the water management authorities about the water quality and the expected evolution in the time framework of weeks or months

The main objective of the use case is to install and execute automatically the whole monitoring and modeling platform on Cloud resources and HPC resources on demand, to make it available on different service providers for the users of the collaboration.

The adequately "tuning" of the model, and also the exploration of the predictions under different conditions, requires the iterative/coupled execution of the different simulation components. A SaaS/PaaS solution, including also the possibility of "finding the best solution through a parametric exploration", could be of great interest. The SME plans to use intensively the Cloud as the basic framework, and they would need to migrate their usual tools (visualization using Matlab free package, analysis using excel/python/R scripts) and run them on the quite large simulation output (order of hundredths of terabytes). The solution needs to be scalable to be implemented for as many water reservoirs as needed.

The roles and the responsibilities identified for this use case are:

- The researchers (typically linked to the limnology community) that study the evolution of the water quality, in particular eutrophication. In this case, the core team is integrated by biologists and environmental researchers working at an SME, Ecohydros.
- Water management authorities, that include also biologists, chemists and civil engineers. Main institution in this example is the Confederación Hidrográfica del Duero (CHD) in Spain.
- ICT groups, like the one at IFCA, supporting the implementation of the instrumentation and the simulation. Also other companies, like ITG, involved in the Life+ project.
- Other groups in limnology, an area that covers a wide spectrum of professionals in other sectors (like other biologists interested in the evolution of the cyanophicea, including genetic topics).
- Typical institutions: public research organizations (like CSIC, and also CEDEX in Spain), Universities (like the University of Cantabria and the Universities Autónoma and Complutense de Madrid in Spain)

Currently the monitoring/data collection e-infrastructure is not in the Cloud, but the plan is to provide such a setup by the end of this year. The resources will be provided by IFCA nodes integrated in the EGI.eu FedCloud. The simulation can run in FedCloud as well, but an optimized remote access has to be provided to researchers to be able to examine/visualize the large data outputs. Remote interactive post-processing using either Matlab, R or Excel, is required. Preferably all the components should run in the Cloud. The integration of the full data workflow is needed to enable the biologists in the SME to execute it fully in the Cloud: from the storage of the instrumentation data to its use as input for the DELFT-3D model, that requires HPC resources for execution, to the post-processing using standard software like Excel or R or Matlab to process large outputs.

Remote access is a requirement for post-procesing. Enabling HPC/HTC workflows in the Cloud is required. Parametric runs would largely benefit the validation. Currently the output visualization requires Matlab plug-in.

Implementation of the DMP would help a lot to make the solution scalable and sustainable.

The use case involves two main user stories:

User Story A): SME team wants to model the hydrodynamic behavior of the water reservoir, to reproduce the thermocline and predict the onset and completion times of the water column stratification, with special interest in its final phase (september/october).

User Story B): SME team wants to predict algae bloom based on model, and validate against previous year detailed analytical measurements

In particular there is need for a service for the access to the data from the INSTRUMENTED PLATFORM that is located in CdP water reservoir and replicated in a server at IFCA. This service allows accessing / transferring all the data required by the model.

The different model components are currently processed in a wide range of computing resources: from personal computers (water quality model, low resolution hydrodynamics), Altamira supercomputer (medium-high resolution hydrodynamics and water quality when needed) or cloud VMs (medium-high resolution hydrodynamics and water quality when needed).

Remote post processing of output: after processing, the (large/very large) output is stored where the processing was made and any replication requires high bandwidth connectivity. Analysis is made from personal computers using Excel or Matlab. Currently data analysis are stored in PCs and uploaded to a "SugarSync" folder for sharing with ECOHYDROS researchers when it is tagged as relevant model. This process should be transformed into a remote post processing service in the Cloud.

5.1.2 TRUFA

Since the introduction of the RNA-seq methodology around 2006, studies based on whole transcriptomes of both model and non-model species have been flourishing. RNA-seq data are widely used for discovering novel transcripts and splice variants, finding candidate genes, or comparing differential gene expression patterns. The applications of this technology in many fields are vast, including researches on, for example, splicing signatures of breast cancer, host–pathogen interactions, the evolution of the frog immunome, the plasticity of butterfly wing patterns, the study of conotoxin diversity in Conus tribblei and the optimization of trimming parameters for de novo assemblies.

Despite the tremendous decrease in sequencing costs, which allows virtually any laboratory to obtain RNA-seq data, transcriptome analyses are still challenging and remain the main bottleneck for the widespread use of this technology. User-friendly applications are scarce and the post-analysis of generated sequence data demands appropriate bioinformatics know-how and suitable computing infrastructures.

When a reference genome is available, which is normally the case for model system species, a reference-guided assembly is preferable to a de novo assembly. However, an increasing number of RNA-seq studies are performed on non-model organisms with no available reference genome for read mapping (particularly those studies focused on comparative transcriptomics above the species level), and thus require a de novo assembly approach. Moreover, when a reference genome is available, combining both de novo and reference-based approaches can lead to better assemblies. Analysis pipelines encompassing de novo assemblies are varied, and generally include steps such as cleaning and assembly of the reads, annotation of transcripts, and gene expression quantification. A variety of software programs have been developed to perform different steps of the transcriptome analysis (RNA-seq), but most of them are computationally intensive. The vast majority of these programs run solely with command lines. Processing the data to connect one step to the next in RNA-seq pipelines can be cumbersome in many instances, mainly due to the variety of output formats produced and the postprocessing needed to accept them further as input. Moreover, as soon as a large computing effort is required, interactive execution is usually not feasible and an interface with the underlying batch systems used in clusters or supercomputers is needed.

Application of next-generation sequencing (NGS) methods for RNA-seq has become increasingly accessible in recent years and are of great interest to many biological disciplines including, eg, evolutionary biology, ecology, biomedicine, and computational biology. Although virtually any research group can now obtain RNA-seq data, only a few have the bioinformatics knowledge and computation facilities required for transcriptome analysis.

TRUFA (TRanscriptome User-Friendly Analysis) is an open informatics platform offering a web-based interface that generates the outputs commonly used in de novo RNA-seq analysis and comparative transcriptomics. TRUFA provides a comprehensive service that allows performing dynamically raw read cleaning, transcript assembly, annotation, and expression quantification. Due to the computationally intensive nature of such analyses, TRUFA is highly parallelized and benefits from accessing high-performance computing resources. The complete TRUFA pipeline was validated using four previously published transcriptomic data sets. TRUFA's results for the example datasets showed globally similar results when comparing with the original studies, and performed particularly better when analyzing the green tea dataset. The platform permits analyzing RNA-seq data in a fast, robust, and user-friendly manner. Currently accounts on TRUFA are provided freely upon request at https://trufa.ifca.es. TRUFA has been developed by IFCA in collaboration with MNCN (Spanish Natural Science Museum, also in CSIC). Access to the web portal is available under subscription for the research community.

5.2 EuroBioImaging: the Virtual Biobank

This Virtual Biobank will integrate data with a high degree of heterogeneity, like medical images with all the interpretations of DICOM3 standards from the different manufacturers, parametric maps of imaging biomarkers, 3D reconstructions of anatomy, source codes of image processing algorithms, and associated clinical data and variables. The main data types managed by the platform will be plain DICOM files (one DICOM file per image) and also NIFTii and analyse formats (*.nii and *.hdr/*.img file extensions, respectively), that can handle entire volumes in a single file and are widely extended among medical image processing scientists community.

A pipeline-based architecture will be used for the development of quantitative imaging procedures. Image analysis methods will be classified in those used for quality control data (i.e., signal noise ratio plot), data pre-processing (i.e., segmentation, filtering, interpolation), data analysis (i.e., brain volumes quantification, T2 mapping, perfusion analysis, lung emphysema quantification) and data measurement (i.e., histogram based analysis, multivariate statistical analysis). Rules will be defined for the interconnection of the different types of processing modules for pipeline creation.

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 and may require interactive access to the applications.

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.
- Persistent Storage. Data must be kept even if the computing nodes are powered down. Frequent data transfers should be avoided as they may involve TBs of data.
- Software repository. Pre-configured software packages for the commonly used image processing, pipeline orchestration and visualization tools should be available.
- Software configuration. Each subproject 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. Projects may require resources or have external resources available, and the applications should work seamlessly. Automatic elasticity is taken for granted.

5.3 INSTRUCT: Molecular Dynamics

Molecular dynamics (MD) simulations of macromolecules have grown to become a standard tool for complementing experiments, providing a structural basis for rationalizing in vitro and in vivo observations, and for suggesting new experiments. Advances in algorithms and hardware have allowed ever larger systems to be simulated for ever longer times, providing, among other things, exciting views on function-related dynamics and assembly of full virus particles, protein folding events and mechanisms, and long time scale dynamics. A variety of software tools is available to apply MD methods. However, due to the complexity of the underlying methods their usage requires both a lot of experience and sufficient computing power. The latter can be satisfied by using cloud computing to gain access to appropriate computing resources, which can include also High Performance Computing (HPC) systems. The complexity of methods and resources instead must be tackled by providing a user-friendly and intuitive interface to the scientists.

Here we seek to implement a pipeline that combines protocols that automate the step for setup and execution of MD simulations, using state-of-the-art approach, appropriate data and metadata management, and state-of-the-art approaches to the analysis of MD trajectories, aiming also at simplifying the comparison with different kinds of experimental data for validation of the simulations. The pipeline shall be integrated by a transparent mechanism to provide the most appropriate computational resources for each specific simulation.

Users will be associated to projects, and guided in their work via simple graphical interfaces, e.g. embedded in the browser. The range of applications can extend from “simple” free dynamics to energy calculations, docking of ligands and macromolecular partners, simulation of substrate/product diffusion.

The key scientific and technical challenges of this case study are:

- Long-term storage, also in order to enable complex post-simulation analysis of trajectories;
- Capability to move analysis tools to the simulation data in order to avoid data transfer;
- Availability of pre-configured software packages both to run simulations and to analyse the trajectories produced;
- Flexibility to use different computational architectures as needed, via standardize, optimized protocols;
- Tools to monitor the execution of the calculations and to cope with power-down or failure.

5.4 LBT: Astronomical Data Archives

The Large Binocular Telescope (LBT) is equipped with several instruments able to investigate different subjects in different wavelengths and for very diverse scientific fields of investigation. Each instrument has its own peculiarities and capabilities and has to be controlled, configured and operated in particular way respect to the others. Starting from the diverse proposals, data produced by each instrument is stored in a permanent archive and first handling consists in data delivery to partner archival mirrors sites. Each partner institution manages both private and public data and after data storing, performs data reduction and calibration. After the first data reduction some other data manipulation and information extraction could be done locally by expert users via already existing

tools or pipelines. Again data sharing and data delivery to the final users are required. Merging the current application developed specifically for one or a set of instruments, it would be nice to have a more extended view of the Astronomer needs that could be basically summarized into the following.

Astronomers would expect in the future developments and in particular from the promising improvements of the Cloud technology an integrated infrastructure where the possibility to find all the e-infrastructure an existing Telescope like LBT offer like distinct and non-correlated support software. More in detail, some existing tools like the observation proposals, the data reduction and the data quality calculations, continuing to data handling and workflows, the subset used in papers to the data publication for educational purposes are completely independent and uncorrelated processes also from the user point of view.

It would be of great improvement on the efficiency of data discovery, comparison, cross matching, data mining and publication if an integrated infrastructure could host all those processes and transparently accessed by astronomers and users in general.

The data life cycle in Astronomy as in several other disciplines involves also the data acquisition preparation. The starting point is the Astronomer that, following the necessities of his/her personal field of investigation, requires more data to extract information from. Since the peculiarities of each astronomical facility, Astronomers has to investigate the more appropriate Telescope / Instrument filling the necessities and start to plan a well formed and motivated proposal for the Observing time request. Nonetheless, he has to verify no other Observations with the same target and same techniques had already been submitted. This checking process is usually done only in the target facility, but would be nice if some of the competitive facilities (i.e. facilities that offer the same capabilities) could share those kinds of information, in order to optimize the re-usage of the same kind of observations like suggested in the H2020 guidelines. After the proposal acceptance and the real observation, data reduction, and interactive data handling are the following processes to gain the scientific goals. After data calibration and reduction the scientific result is part of the paper and publication products and is paralleling part of catalogues over which data mining is performed. Sub sets of data are to be collected as unique sets and the Digital Object identifiers standard cataloguing technique is one of the min used to preserve the dataset content. Last (in terms of time but not importance) is the data publication for Educational purposes.

5.5 CTA: Archive system for the Cerenkov Telescope Array

Very High Energy (VHE) gamma-ray astronomy with CTA is evolving towards the model of a public observatory where guest observers will submit observation proposals and have access to the corresponding data, software for scientific analysis and support services. The technical implementation of The CTA Data Management Sub-Project (henceforth referred to as Data Management) therefore aims to fulfil the requirements of a public observatory guaranteeing reliable processing, ensuring quality of services for access, dissemination and transmission of data.

The CTA data and their scientific products need to be preserved in a dedicated archive guaranteed to provide open access to a wide and diverse scientific community. Data Management provides scientific data products through community-based standards (e.g. the Virtual Observatory) and relies on existing e-infrastructure for data transmission and dissemination.

Handling and archiving the large amount of data generated by the instruments and delivering scientific products according to astrophysical standards is one of the challenges in designing the CTA observatory.

The high data rate of CTA, together with the large computing power requirements for data processing and Monte-Carlo simulations, require dedicated computer resources. Furthermore the participation of scientists from within CTA Consortium and from the greater worldwide scientific community necessitates a sophisticated scientific analysis system capable of providing unified and efficient user access for all data levels, software and computing resources.

The main scope of the project is the design of the CTA Science Data Centre, which is in charge of the off-site handling of data reduction, Monte Carlo simulations, data archiving and data dissemination. The remote (e.g. intercontinental) transmission of data from CTA sites to the CTA archive is one of the key services that the Science Data Centre administers at both ends: off and on the CTA site. The development and provision of software and middle-ware services for dissemination including observation proposal handling is a task that Data Management guarantees to be interfaced with the Operation Centre.

The services and components that the CTA Data Management is in charge of at the CTA sites include: the execution of on-site scientific data reduction pipelines, the real-time analysis software, the on-site temporary archive system as well as the data quality monitoring.

The CTA Archive will handle all the data produced by the Observatory and thus it has to be properly designed in order to reach desired goals. It has to respond to three main issues, which represents projects requirements:

- the treatment and flow of data from remote telescopes;
- “big-data” archiving and processing;
- open access to all data.

The design is inspired by the lessons learned from current and past Atmospheric Cherenkov Telescopes, from existing astronomical observatories, and finally from the technical know-how of major computing and data centres that serve large international projects and world-wide communities.

5.6 WeNMR: HADDOCK

Protein interactions that are critical to all cellular processes establish an intricate and dynamic molecular network – the interactome – in which subtle miscommunications often result in disease. The large gap between the number of interactions and available experimental 3D structures calls for complementary computational methods to produce accurate predictions and guide experimentalists. This is the field of computational structural biology, which has seen in the last decade fascinating developments both in software and hardware. Computational structure prediction is nowadays routinely considered an integral part of research. The docking field, in particular, has thrived in the last decade since the beginning of the CAPRI (Critical Assessment of PRedicted Interactions) experiment, in which the participants are asked to predict the structure of an unknown biomolecular interaction. Computational modelling of complexes has grown into a well-accepted complementary method to classical experimental techniques.

The U.Utrecht partner has developed for over ten years now the integrative, information-driven docking approach, HADDOCK (http://www.bonvinlab.org/software/haddock2.2/haddock.html).

It supports the incorporation of a large variety of data from NMR and other biophysical methods HADDOCK has demonstrated a strong performance in the blind docking experiment CAPRI.

The HADDOCK software and its associated web portals are fully operational and in use by a large community. The web portal is currently operated at Utrecht University exclusively, on physical machines. The main aim of this use case is to virtualize the web portal and the required computation infrastructure underneath it, in order to be less dependent on local hardware and facilitate possible deployment at other (possibly within company) sites.

The key technological requirement is a virtualize computational infrastructure that provides both the frontend for the HADDOCK web portal, together with enough computing resources to run the calculations, e.g. a virtualized cluster, with master node controlling the computations and serving the web portals, associated compute nodes (for a minimum of 100 cores) with scheduling system for the jobs, and possibly federated user identification.

5.7 ENES: Climate Model Inter comparison Data Analysis

The case study on *climate models intercomparison data analysis* is directly connected to the Coupled Model Intercomparison Project (CMIP). CMIP studies output from coupled ocean-atmosphere general circulation models that also include interactive sea ice. These models allow the simulated climate to adjust to changes in climate forcing, such as increasing atmospheric carbon dioxide. CMIP began in 1995 by collecting output from model "control runs" in which climate forcing is held constant. Later versions of CMIP have collected output from an idealized scenario of global warming, with atmospheric CO2 increasing at the rate of 1% per year until it doubles at about Year 70. CMIP output is available for study by approved diagnostic sub-projects. The WCRP CMIP3 multi-model dataset archived at PCMDI, included realistic scenarios for both past and present climate forcing. The research based on this dataset has provided much of the new material underlying the IPCC 4th Assessment Report (AR4). The WCRP CMIP5 experiment has provided the bases for the IPCC AR5. The CMIP5 experiment design has been finalized with the following suites of experiments: (i) Decadal Hindcasts and Predictions simulations, (ii) "long-term" simulations, and (iii) "atmosphere-only" (prescribed SST) simulations for especially computationally-demanding models.

CMIP5 has promoted a standard set of model simulations in order to:

- evaluate how realistic the models are in simulating the recent past,
- provide projections of future climate change on two time scales, near term (out to about 2035) and long term (out to 2100 and beyond), and
- understand some of the factors responsible for differences in model projections, including quantifying some key feedbacks such as those involving clouds and the carbon cycle.

CMIP5 notably provides a multi-model context for 1) assessing the mechanisms responsible for model differences in poorly understood feedbacks associated with the carbon cycle and with clouds, 2) examining climate “predictability” and exploring the ability of models to predict climate on decadal time scales, and, more generally, 3) determining why similarly forced models produce a range of responses[2].

With specific regard to the CMIP* context, the Case Study will focus, in particular, on a specific set of data analysis. More specifically:

- Anomalies analysis
- Trend analysis
- Climate change signal analysis

Moreover, the output related to these three classes of data analysis will be considered as a basis for additional data analysis experiments, such as:

- Tracking analysis (e.g. tropical cyclones, oceanic water masses)

[2] CMIP5 - http://cmip-pcmdi.llnl.gov/cmip5/

- Transport analysis (e.g. Moc, oceanic transport, atmospheric transport, atmospheric rivers identification)

5.8 Galleries, Libraries, Archives and Museums: eCulture Science Gateway

The case study on Italian Libraries is based on eCSG, will offer a secure and effective long-term preservation in the Cloud to the digital collection stored in InternetCulturale. The e-Culture Science Gateway presents a model to give a transparent access to Libraries for the biggest number of researchers and it presents some tools in order to upload digital resources with related metadata.

The eCSG will allow identified users to upload, download digital collection in a massive way and will offer system for search and retrieve and tools for annotation and visualization.

The community involved are libraries from all type of Institutions like National Library, State Libraries local Libraries, and Libraries of Universities that manage digital collections. But also the large community of researchers in the sector of Digital Humanities, Cultural Heritage, Archaeology and Science that are interested in viewing the digital copies of books, manuscripts and other document for annotating re-use it. So, it is possible to define different type of users with multilevel staff privileges, from a top level for the collection manager/records keeper, down to the other staff member and simple researchers.

5.9 ELIXIR: Galaxy as a Cloud Service

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.

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.

5.10 EMSO: MOIST, Multidisciplinary Oceanic Information SysTem

The scientific management of time-increasing quantity of data has recently become a big challenge in terms of storage capacity, preservation, interoperability and data access in many disciplinary sectors.

In the environmental Science sectors, the analysis of large amount of time series (sustained measurements over time) is considered necessary for any predicting modelling in reply to urgent questions on global changes at different space and time scales.

As an example, the development and use of multiparameter seafloor and water column observatories enabling a multidisciplinary approach to investigate the deep sea processes with different time scales (from seconds to decades), has posed the need to collect, organise and maintain in the long-term a variety of long time series.

MOIST, Multidisciplinary Oceanic Information SysTem, is presently a data and metadata provider initiated within the ESONET NoE project and under implementation and development as part of EMSO (European Multidisciplinary Seafloor and water column Observatory) also in link with funded projects such as EC Genesi-DEC, EC SCIDIP-es, EC ENVRI, EC CoopEUS, EC ENVRI-Plus, EC EMSODEV.

MOIST is designed to make available scientists and users multidisciplinary data obtained by means of fixed-point observatories managed by INGV in some EMSO key-sites. The MOIST configuration underpins the observatory data flow from the sensor acquisition to the dissemination.

MOIST is developed by adopting the most common data standards (e.g., OGC, NASA, INSPIRE) organising, indexing and converting the data into a unique data scheme and supports some EMSO observatories node regardless of their specific suite of sensors and sensor configuration and operational status.

Visualisation functions are implemented to inspect and qualitatively compare the time series of different parameters and/or sensors.

Raw data quality check tools are going to be designed to enrich the metadata with additional information about data completeness, consistency and coherence.

MOIST represents a significant example of local data management systems of a typical EMSO observatory node.

5.11 DARIAH: Big Data in Arts and Humanities

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

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

5.12 Other Research Applications in EGI FedCloud

As indicated, a relevant set of Case Studies and Applications has emerged from the EGI Federated Cloud outreach activities. They are listed below, and included in the Annex prepared by EGI.eu:

- **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** 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. in 2010) originating from bacterial samples. Meanwhile is has been extended to process data generated in different experimental setups and from all domains of life.
- **JAMS** is a 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 ranging from 1km^2 to 100.000 km^2 and a temporal time step ranging from hours to months.
- **HAPPI**, developed in the 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 so called Preservation Descriptive Information.

- **INERTIA** 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 adopts the Internet of Things/Services principles to the Distribution Grid Control Operations.
- **DRIHM** is a European initiative that has been running along the last five years, aiming at providing an open, fully integrated workflow platform for predicting, managing and mitigating the risks related to extreme weather phenomena.
- **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.

Requirements have been also gathered from other communities/research infrastructures that do not yet use EGI FedCloud, like the Human Brain Project and also other EGI-Engage related projects.

The Human Brain Project aims to accelerate the 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, Neuromorphic Computing and Neurorobotics. HBP is developing an image service which manages all issues that providing multi-range solution of the use onto the data, searching of sub-volume of data, also be able to do arbitrary facing angles. HBP is looking repositories enable data services for accessing just sub-portion or different resolution view of the data wanted and ultimate be able to do analysis where that data sets are. The purpose is to leave the data in place. The two main use cases analyzed so far are one interactive and one batch processing:

- HBP is developing an image service that provides multi-range solutions for data access, searching of sub-volume of data, and provide a visualization web interface. This is a basic service for core processing like to register data and align data to to standard template brain spaces. The service is developed using Python, HDF5. Data container is currently plug into several web clients using desktop analysis software.
- HBP plans to deploy analytical software that runs either on a single thread or multiple threads to analyse, for example, to extract neuron morphology, which needs a tracing algorithm for large collaborative data neutrons which is recently launched a world-wide efforts of collection of large neuron image sets. The aim is to use automatic reconstruct algorithm to trace these neutrons and extract the object that from the image sets which them be analysed by cluster algorithm. This may require to run ideally in parallel reconstruction algorithms and to be able to compare the results.

CONFAR: Canadian Advanced Network for Astronomical Research (CANFAR) is a computing infrastructure for astronomers. CANFAR aims to provide users easy access to very large resources for both storage and processing, using a cloud based framework. Current problem related to e-Infrastructure is that CANFAR data center in Italy only have data but no computing capacity. There is a need to associate the computing facility with data cloud -- move computation to Data. CARFAR services in Canada allow replication of the data, data localization, move computing to the data using Open Stack technology. CARFAR Italian datacenter needs the similar facilities. As the first step, the immediate requirement is to enable data replication between Europe and Canada.

LoFAR is a next-generation radio telescope that is currently being developed across Europe. The data volume of LoFAR is increasing dramatically each year results in urgent needs for updating its data system. One of the current problems of LoFAR data system is the pipeline for data access is not easily reconfigured: each time a user need to stage a large volumes of data from tapes onto the disks which is highly time consuming (takes 1~2 days). Traditional a researcher need to download the data into a local machine for data processing. As the data volume become larger, this become infeasible. LoFAR has limited computational resources at the moment, when there is burst access/processing request, it's own computing capacity will not likely be able to cope with. It needs new solutions, e.g., use EGI FedCloud to federate the LoFAR private data repository, and to provide elastic computing resources on demands. In this scenario the use case is to have data replication and staging capabilities, to make the distributed scalability possible.

EGI-Engage project collaborates with eight EU high-impact research infrastructures/communities through joint development of customized services based on core EGI capabilities: the so-called Competence Centres (CCs). The communities that collaborate through EGI-Engage CC program include BBMRI-ERIC, ELIXIR, MoBrain, DARIAH, LifeWatch, EISCAT-3D and EPOS.

New use cases that emerge from these EGI-Engage Competence Centers with requirements for using Cloud resources include:

- **BBMRI-ERIC**, which aims to facilitate the implementation of big data storage in combination with data analysis and data federation by integrating technologies from the Bio-banking community, EGI and other e-Infrastructures.
- **EPOS (**the European Plate Observing System) aiming to drive the future design of the use of grid and cloud for the integrated solid Earth Sciences research. The Competence Center will:
    - (1) identify and validate authentication and authorization services
    - (2) test cloud resources and usage models, and
    - (3) provide knowledge transfer services between e-Infrastructure and EPOS communities.
- **Disaster Mitigation**, which aims 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 required to respond to natural disasters.

All the collected information is included as ANNEX P07 attached to this deliverable. Please refer to it for detailed information.

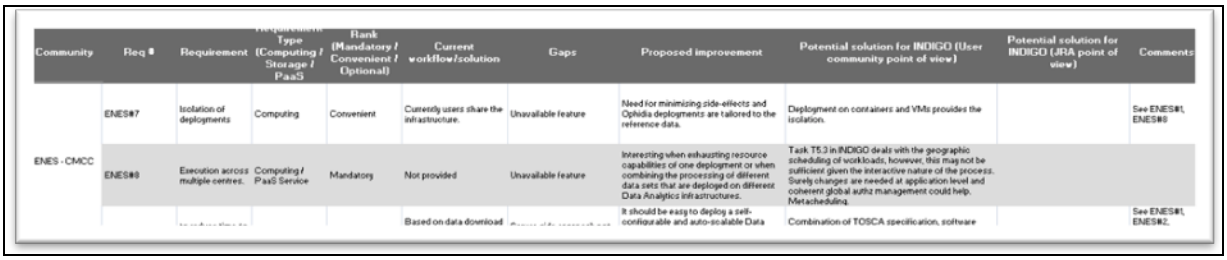
6 REQUIREMENTS

As described in section 4 on Methodology, the previous Case Studies summarized in section 5 and described in detail in the corresponding Annexes were used directly to gather the list of requirements.

6.1 Requirements gathered from Case Studies

As an initial analysis, and not precluding a more detailed and direct analysis of the Annexes directly by JRA teams, a (long) list of requirements per Case Study was compiled internally within WP2/NA2. The tables include several entries per Case Study, and detail the Research Community, Requirement enumeration (#), description, priority rank (Mandatory/Convenient/Optional), current solution, gap, etc. The requirements are classified as "Mandatory", if they are felt as such in the input provided by the corresponding Research Community. This may mean that either the requirement is satisfied by a current solution and needs to be preserved, or that it must be implemented in the new solution.

The complete list is provided in an Appendix (section 9) in this same document, for completeness, but an online version of the list is also available[3]. The format is depicted in the figure below:



Community	Req #	Requirement	Type (Computing / Storage / PaaS	Rank (Mandatory / Convenient / Optional)	Current workflow/solution	Gaps	Proposed improvement	Potential solution for INDIGO (User community point of view)	Potential solution for INDIGO (JRA point of view)	Comments
	ENES#7	Isolation of deployments	Computing	Convenient	Currently users share the infrastructure.	Unavailable feature	Need for minimizing side-effects and Ophidia deployments are tailored to the reference data.	Deployment on containers and VMs provides the isolation.		See ENES#1, ENES#8
ENES - CMCC	ENES#8	Execution across multiple centres.	Computing / PaaS Service	Mandatory	Not provided	Unavailable feature	Interesting when exhausting resource capabilities of one deployment or when combining the processing of different data sets that are deployed on different Data Analytics infrastructures.	Task T5.3 in INDIGO deals with the geographic scheduling of workloads, however, this may not be sufficient given the interactive nature of the process. Surely changes are needed at application level and coherent global authz management could help. Metascheduling.		
					Based on data download		It should be easy to deploy a self-configurable and auto-scalable Data	Combination of TOSCA specification, software		See ENES#1, ENES#2.

Figure 4: Case Studies Requirements format showing the different fields

The complete list includes around 100 identified requirements.

6.2 Common Requirements

As a next step in the proposed analysis, a new list was prepared within WP2/NA2 trying to identify in the previous list the Common Requirements.

This initial list is included below, and can be found also online (following the same link of the complete list). The requirements are classified into three categories:

Category A: Computational Requirements

Category B: Requirements linked to Storage

Category C: Requirements on Infrastructure

They are also classified according to their type (Computing / Storage / PaaS Service), and a rank (Mandatory/Convenient/Optional) assigned on the basis of the Case Studies requirements.

The list is shown in the following table:

[3] See https://docs.google.com/spreadsheets/d/10LYbk-V5Y7YeYLY9KVaLRDcgWTR293-ZUxIlsIac2oQ/edit?pli=1#gid=0

#REQ	Description	Type	Rank	Proposed Improvement
CO#1	**Deployment of Interface SaaS**	Computing / PaaS	M	A mechanism to facilitate the deployment of a customised Haddock portal and backend in system in a panoply of infrastructures with minimal intervention.
CO#2	**Deployment of Customized computing back-ends as batch queues**	Computing / PaaS	M	Each instance may have an independent software configuration, potentially incompatible with other projects or specially tailored without side-effects.
CO#3	**Deployment of user-specific software**	Computing / PaaS	M	Manual installation may be cumbersome for large-scale application involving many computing resources or when requesting users to update VMIs. This should be automated.
CO#4	**Automatic elasticity of computing batch queues**	Computing / PaaS	M	When moving to the cloud, users should be provided with the exact number and size of resources they need. Overprovisioning will produce an undesirable cost or inability to serve other requests. On the other side, underprovisioning will lower QoS.
CO#5	**Terminal access to the resources.**	Computing / PaaS service	M	This feature must be linked to the AAI
CO#6	**Privileged access**	Computing / PaaS service	M	This feature must be linked to the AAI
CO#7	**Execution of workflows**	Computing / PaaS	M	Processing done on the cloud where the outputs of the processing are stored. Orchestration of complex pipelines.
CO#8	**Provenance information**	Computing / PaaS Service	C	Very important for revision of papers and project proposals.
CO#9	**Cloud bursting**	Computing / PaaS Service	M	Supplementing the computing capacity with special instances (provided with higher individual resources, such as RAM) or with a larger number of them.
CO#10	**Data-aware scheduling**	Computing / PaaS Service	C	Currently storage and computing are highly coupled.
CO#11	**Provisioning of efficient Big Data Analysis solutions exploiting server-side and declarative approaches**	Computing / Storage / PaaS Service	M	
CO#12	**Execution across multiple centres.**	Computing / PaaS Service	M	Interesting when exhausting resource capabilities of one deployment or when combining the processing of different data sets that are deployed on different Data Analytics infrastructures.
CO#13	**On-line processing of data**	PaaS	M	Special management of post-processing jobs that could be sent to the resources hosting the data and not vice-versa. Data analytic techniques may

				benefit.
CO#14	**Special hw configuration - MPI, multicore, GPGPU**	Compute / PaaS	M	More flexibility in the way the requirements are defined and the matching with the infrastructure, as well as the automatic installation and configuration of software dependencies will be key.
SO#1	**Shared storage accessible like a POSIX filesystem**	Storage / PaaS Service	M	Limited storage and no scalability
SO#2	**Persistent data storage**	Storage	M	
SO#3	**Long-term availability of results**	Storage	M	External, long-term, self-maintained storage.
SO#4	**Local user storage**	Storage / PaaS Service	M	Separate individual volumes will increase scalability and privacy.
SO#5	**Availability of reference data**	Storage / PaaS Service	M	
SO#6	**Interoperability with application domain specific software**	Storage / PaaS Service	M	And with specific services (e.g. IS-ENES/ESGF) No improvement, keeping this feature.
SO#7	**Metadata management / Database as a Service**	Storage / PaaS Service	C	
SO#8	**Share data capabilities**	Storage/IaaS service	C	Block storage with added NFS-like capability of multiple access. API/POSIX access. Better performance object storage. The solution must support both shared access and high performances.
SO#9	**Data replication**	PaaS	M	Hide the data topology to the user, data federation, data replication capabilities
SO#10	**Distributed storage**	Storage / PaaS service	M	Cloud or grid based solutions have not proven to be efficient yet.
SO#11	**Dropbox-like storage**	Storage / PaaS service	C	Facilitate interaction with users in uploading and downloading files
PL#1	**Global-level AAI**	PaaS	M	Centralized mechanism to define general authorisation policies will give scalability and a coherent mechanism.
PL#2	**On-line access to data**	Computing / Storage / PaaS	M	Interactive access to the VMIs to avoid downloading huge amounts of data for consolidated inspection of results
PL#3	**Network configuration**	IaaS	O	Extend current standard interfaces to support network configuration, such as VPN-aaS, Firewall-aaS
PL#4	**Monitoring and operation**	PaaS	C	Keep functionality

Table 5: List of Common Requirements

This list of requirements is the main outcome of this deliverable, and it is provided as input to the work of the JRA teams within INDIGO-DataCloud project. It was used as starting point in the f2f meeting beginning of July in Valencia oriented to define INDIGO Architecture.

7 AN ANALYSIS EXERCISE: FROM REQUIREMENTS TO GENERIC SOLUTIONS

After analyzing the Case Studies and compiling the requirements, it was felt within WP2/NA2 that at least a sketch of potential generic solutions including the most relevant points found in the analysis, could be helpful as an input to JRA. This was also felt to be useful to trigger the internal discussion with the Research Communities ICT teams. Following this argument, the complete table contains also a column labelled "*Potential solution for INDIGO (User community point of view)*".

And when considering the Use Cases, it was considered important to remark the need to include different roles (final user, developer, managers) in their analysis, and to try to find general solutions to support common requirements.

Two complete generic examples were prepared by the UPV team, and are shown below. They were presented for discussion at the f2f meeting in Valencia, and are included here as reference.

7.1 Generic Solution A: User Community Computing Portal Service

A user community has an application (or set of them) that can be accessed through a portal and requires a batch queue as back-end. Among its features, an unpredictable workload and user access profile. The application consists of two main parts: the portal / scientific Gateway and the processing working nodes. The requirements imply that:

- Working nodes should scale-up and down according to the workload.
- Cloud-bursting to external infrastructures may be requested.
- Portal services should also adapt to workload.
- Users can access reference data and provide their own local data.

A solution along these lines has been requested in the Case Studies from ELIXIR, WeNMR, INSTRUCT, EGI-FedCloud, DARIAH, INAF-LBT, CMCC-ENES, INAF-CTA, LifeWatch-Algae–Bloom, EMSO–MOIST.

The figure below shows the scheme of a potential solution, to trigger the discussion within JRA and illustrate back to the Research Communities.

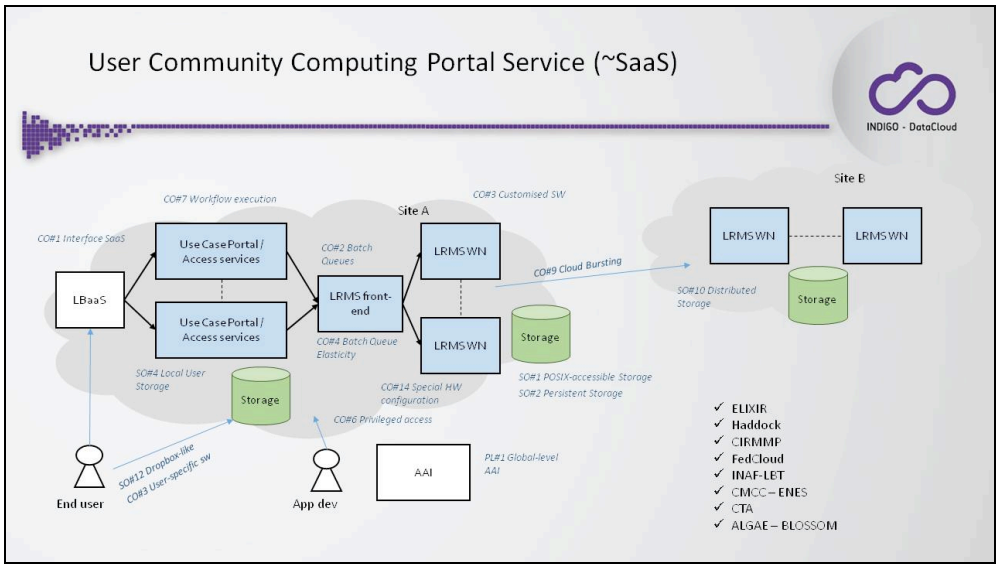


Figure 5: An example of potential generic solution: User Community Computing Portal

7.2 Generic Solution B: A Data Analysis Service

This generic solution example is based on the need of a user community that has a coordinated set of data repositories and software services to access, process and inspect them. Processing is interactive, requiring accessing a console deployed on data's premises. The application consists on a console / Scientific Gateway that interacts with the data:

- Examples include "R", Python, Ophidia
- It can be a complementary scenario from the previous one.
- It can expose programmatic services.

Requested by the Case Studies from INSTRUCT, INAF-LBT, CMCC-ENES, LifeWatch-Algae–Bloom, EMSO–MOIST.

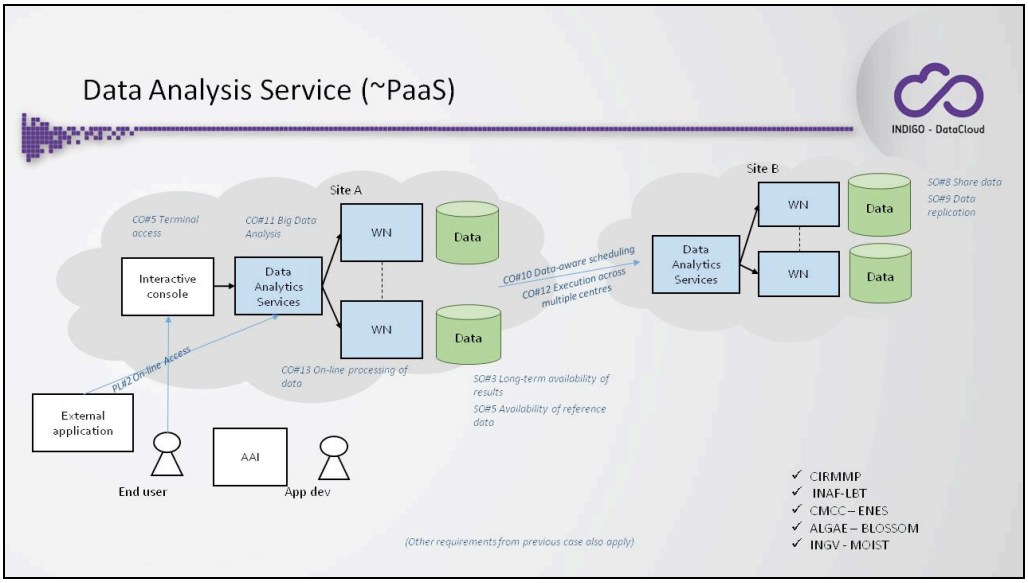


Figure 6: An example of potential generic solution: Data Analysis Service

8 NEXT STEPS

The formal submission of this deliverable to the INDIGO-DataCloud PMB (Project Management Board) triggers, as already shown in figure 3, the start of an evolving interaction among WP2/NA2 and the JRA WPs that will develop INDIGO solutions.

The next step has been already given in the INDIGO JRA-Architectural f2f meeting in Valencia on 7th and 8th July, where the results of the collection of information on the Case Studies included in this deliverable were already presented.

An initial set of additional questions have been already posed by JRA teams, that will be in turn considered and answered by the research communities, providing an initial refinement of the requirements. This work will be done along the line agreed of supporting an Agile development method. Along that same line, the current version of the Annexes of the different Case Studies will be updated and improved. Moreover, internally, each Research Community should try to identify a "Champion" person to lead this "Agile-like" effort in coordination with WP2/NA2 management and with the JRA teams, to assure both that they are taken into account and that INDIGO solutions are of interest to the Research Communities that they represent.

So the work will continue along next months, in particular also including the subtask related to the interaction with other parallel projects that are starting recently,

-Liaise with the INFRADEV-4 projects to enable synergies between the projects, and interoperability between the INDIGO outputs and the VRE to be deployed by the E-INFRA-9 projects.

This deliverable D2.1 will be updated in Month 9 (December 2015) in parallel to the preparation of deliverable D2.4, on *"Confirmation of support to initial requirements from JRA design and extended list of requirements"*.

D2.1 will be also the basis for the preparation of deliverable D2.3 "*Specifications of use cases for testing and validation purposes*", due by Month 8 (November 2015), within the task T2.3 on "*Application Test and Validation"*. The applications originally proposed are included in the list that has been presented here: HADDOCK, Molecular dynamics simulation, Climate model intercomparison analysis, astronomical pipeline for data reduction, and Galaxy.

Finally, improving on the data management requirements is another of the key objectives for the next months to be addressed within the subtasks oriented to the:

-Acquisition of procedure details/parameters (i.e., DMP, Collection, Authenticity & Provenance, Data Preservation) to elaborate the specifications for data ingestion and use in INDIGO.

-Definition of the specifications of INDIGO ingestion integrity test.

that will provide the input to deliverable D2.7 "*Specifications of data ingestion and use in INDIGO"*.

9 APPENDIX A: LIST OF REQUIREMENTS DERIVED FROM CASE STUDIES

The complete list at the time of completing this deliverable is included here for completeness; an online up-to-date version can be found at https://docs.google.com/spreadsheets/d/10LYbk-V5Y7YeYLY9KVaLRDcgWTR293-ZUxIlsIac2oQ/edit?pli=1#gid=0

Community	Req #	Requirement	Type (Computing / Storage / PaaS service)	Rank (M / C / O)	Current workflow/solution	Gaps	Proposed improvement
EuroBioImaging	EB#1	Persistent (but medium-term) data storage volumes with standard POSIX file access	Storage	M	Remote access of a shared disk through NFS and XNAT	Limited storage and no scalability	Volumes on the cloud
	EB#2	ACL in the access to data	Storage / Privacy	M	Use of linux filesystem users/groups permissions	Unflexible mechanism requiring manual intervention	A more sophisticated ACL mechanism will enable creating external access tokens.
	EB#3	Execution of data-driven and computing-intensive workflows	Computing / PaaS service	M	Use of LONI on top of a batch queue	No encapsulation, limited resources	Processing done on the cloud where the outputs of the processing are stored. Orchestration of complex pipelines.
	EB#4	Availability of customised software	Computing / PaaS service	M	Common installation of major software packages	Potentially non compatible software packages, increasing disk footprint, complex patching and management.	Each subproject may have an independent software configuration, potentially incompatible with other projects or specially tailored without side-effects.
	EB#5	Deployment of own software	Computing / PaaS service	C	Manual installation on user-space of own software	Manual intervention, user-space.	Manual installation may be cumbersome for large-scale application involving many computing resources or when requesting users to update VMIs. This should be automated.
	EB#6	Resources adaptation to workload	Computing / PaaS service	M	Unavailable now. Shared use of a computing queue.	Unavailable, this will be needed when using other external pay-as-you-go resources.	When moving to the cloud, users should be provided with the exact number and size of resources they need. Overprovisioning will produce an undesirable cost or inability to serve other requests. On the other side, under provisioning will lower QoS.

	EB#7	Terminal access to the resources.	Computing / PaaS service	M	SSH	Fulfilled now.	This feature must be preserved.
	EB#8	Online access to data	Storage / PaaS Service	C	Direct access to the infrastructure	Fulfilled now.	This feature must be preserved.
	EB#9	Management of users and groups	Security / PaaS Service	M	Manual administration of the resources	Single global management is needed.	Centralized mechanism to define general authorisation policies will give scalability and a coherent mechanism.
	EB#10	Long-term availability of results	Storage	C	Not provided, relying on users' own external storage.	Totally unavailable	External, long-term, self-maintained storage.
	EB#11	Provenance and repeatability of experiments	All	C	Unavailable now.	Totally unavailable	Very important for revision of papers and project proposals.
Elixir	EL#1	Deployment of a Galaxy instance	Computing / PaaS service	M	Single instance installed	Shared access to a cluster presents limitations described in other requirements.	Individual instances will have advantages such as isolation (EL#3).
	EL#2	Privileged access to the Galaxy instance	Computing / PaaS service	M	Through system administrator	By having multiple instances, each deployment may have a system administrator with enough privileges to operate the instance.	Each project could have its own galaxy admin, giving enhanced flexibility and higher scalability.
	EL#3	Isolation of instances	Computing	M	Cohabiting the same physical resource, isolated by galaxy and OS credentials.	Not provided currently	Isolation enables reducing side-effects due to individual failures, privacy leakages and software configuration incompatibilities.
	EL#4	Isolated storage for each instance	Storage	M	Cohabiting the same physical resource, isolated by galaxy and OS credentials.	Shared storage mapping to OS credentials	Separate individual volumes will increase scalability and privacy.
	EL#5	File-system like storage	Storage	M	Real NFS storage.	Currently shared and provided by the OS and the galaxy accounts.	This feature must be preserved.

	EL#6	Customisation of Galaxy software	Computing / PaaS service	M	Through system administrator	Different projects may require software configurations that are incompatible among them, and configuration through system administrator is normally a bottleneck.	Special users may have the privileges to install software. A catalogue of galaxy software can be provided to improve the process.
	EL#7	Cloud bursting facilities	Computing / PaaS service	C	Not available	A heavy demand of computing resources, storage and RAM may be needed, exceeding provider's capacity	Supplementing the computing capacity with special instances (provided with higher individual resources, such as RAM) or with a larger number of them.
	EL#8	Global-level user administration	PaaS Service	M	Through system administrator	Profiles of users are centrally defined at the level of the galaxy installation.	Each subproject will have its own set of users and privileges, which should be coherent within the whole platform.
	EL#9	Include own software	Computing / PaaS service	C	Through system administrator	System administrators are requested to integrate new software tools and pipelines.	Special users may have the privileges to install new software. New packages should be included in a local catalogue of galaxy software.
	EL#10	Persistent storage	Storage	M	Through the OS storage	Fulfilled now.	This feature must be preserved.
	EL#11	Minimize data transfers	Storage / PaaS Service	C	Computing and storage resources located in the same physical local area network.	Currently storage and computing are highly coupled.	This feature must be preserved.
	EL#12	Automatic elasticity	Computing / PaaS service	M	Not provided.	Galaxy is deployed on a fixed number of nodes which are accessible through a LRMS.	Working nodes should be powered on/off and added/removed to/from the infrastructure depending on the workload and transparently to the user.
	EL#13	Import data from storage facility	Storage / PaaS Service	C	Generally not provided		
	EL#14	Availability of reference data	Storage / PaaS Service	C	Through system administrator	Fulfilled now.	This feature must be preserved.

ENES - CMCC	ENES# 1	Deployment of a software framework on heterogeneous infrastructures	Computing / PaaS service	M	Dedicated installation of the solution. A prototype using DevOps recipes to install customised installations has been successfully tested.	Easy to deploy	More flexibility, higher level of supported platforms and multi-platform support. Individual specification of virtual hardware restrictions.
	ENES# 2	Provisioning of efficient Big Data Analysis solutions exploiting server-side and declarative approaches	Computing / Storage / PaaS Service	M	Mostly sequential. Currently using CDO, NCO, Grads, NCL to analyse/visualize data.	Scalable data analysis	Provisioning of big data solutions to run data analytics workflows/experiments
	ENES# 3	Interoperability with IS-ENES/ESGF	Storage / PaaS Service	M	A multi-service software architecture	New services or components addressing specific needs like data analysis should be interoperable w.r.t. security & computing interface.	No improvement, keeping this feature and work taking into account the activities & roadmap of the ESGF working groups on security and compute parts.
	ENES# 4	Workflow management systems	Computing / PaaS service	M	This involves three steps: definition of the experiment from a repository of available workflows and operators; running the experiment; and inspection of intermediate results, which could imply remote visualization and publication.	Workflow management systems are not exploited to run data analysis experiments. Bash scripts are usually prepared by scientists (client-side analysis)	WfMS could be exploited in the "general experiment workflow" to orchestrate/support data analysis experiments
	ENES# 5	Flexible deployment	Computing / PaaS service	C	No solution	Currently deployments are fixed.	Deployment must be tailored to the specific requirements of the data to be analysed.
	ENES# 6	Efficient, elastic and dynamic	Computing / PaaS service	M/ O	Current solutions are quite static.	No possibility of growing or decreasing except for prototype results in EUBrazilCC	Capability to add / remove computing and storage on demand. Dynamicity can be considered Mandatory, whereas elasticity can be considered Optional.

	ENES#7	Isolation of deployments	Computing	C	Currently users share the infrastructure.	Unavailable feature	Need for minimising side-effects and Ophidia deployments are tailored to the reference data.
	ENES#8	Execution across multiple centres.	Computing / PaaS Service	M	Not provided	Unavailable feature	Interesting when exhausting resource capabilities of one deployment or when combining the processing of different data sets that are deployed on different Data Analytics infrastructures.
	ENES#9	to reduce time-to-solution and complexity	PaaS Service	M	Based on data download preliminary step, client-side analysis, sequential analysis	big data approaches with parallel and server-side capabilities not available.	It should be easy to deploy a self-configurable and auto-scalable Data Analytics cluster with all the software and the console / graphical user interface without system administration background.
	ENES#10	Reusability of final products, intermediate results and workflows	Storage / PaaS Service	M	Re-usability is not well addressed.	Need for provenance and literature information as well as for sharing and re-use of existing "experiment templates"	Availability of papers and workflows (for provenance and repeatability). Marketplace solutions for sharing workflows. PID support taking into account existing work in the area (e.g. RDA).
	ENES#11	Metadata management.	Storage / PaaS Service	C	The system uses Thredds for managing catalogues/meta data, Solr index for indexing datasets	Available	Keep feature
	ENES#12	Authentication and Authorization	Security / PaaS Service	M	Federated identity based on OpenID mechanism. Myproxy servers are also available.	Available for data sharing only	it should be extended to big data analysis facilities for running intercomparison experiments.
	ENES#13	External restricted access	Security / PaaS Service	C	Anonymous access to web portals and scientific gateways	Not identified	Specific deployment with limited data analysis functionalities could serve for demo, training, dissemination.
	ENES#14	Interactive processing	Computing / PaaS Service	C	Interactive processing is available client-side	Server-side approach should provide interactive processing capabilities	To be made available in a distributed, server-side processing/analysis scenario. Software like Ophidia and IPython deal with interactive data analysis aspects.
	ENES#15	Easy-to-use environment	Security/Computing	M/O	Set of tools for data analysis, processing,	No scientific gateways tailoring data analytics	-Data Analytics Gateways for complex experiments/workflows and high resolution data for

					visualization to be available on the client machine	experiments at large scale for CMIP5 experiments are available	scientific research (M), - mobile apps for simplified access to reduced datasets (e.g. indicators, timeseries), for dissemination or demo purposes (Optional).
EGI Federated Cloud requirements	FC#1	Share data capabilities	Storage/IaaS service	C	Ad hoc solutions,NFS over block storage, Object storage	Object storage can be accessed in a shared way, but lacks of performances, and has no file-system like access. Block storage cannot be shared, and can be accessed through a VM only.	Block storage with added NFS-like capability of multiple access. API/POSIX access. Better performance object storage. The solution must support both shared access and high performances.
	FC#2	PaaS deployment orchestration	IaaS	M	OCCI contextualization plus ad hoc configuration of services	No standard/based orchestration mechanisms for federated clouds.	Better orchestration support in the IaaS standard interface, and native APIs of major Cloud Management System.
	FC#3	Integrate orchestrators for specific communities in INDIGO services	PaaS	O	Some user communities have specific orchestrators: Arvados, Curoverse	These orchestrators should be able to interact with the new services developed by INDIGO	These orchestrators should be able to interact with the new services developed by INDIGO
	FC#4	Data aware brokering	PaaS	C	Currently community must autonomously know where their data is and deploy VMs accordingly	In a distributed infrastructure with distributed/replicated datasets users need to know where the datasets are available to submit their computational tasks close to the data.	Broker capabilities using data discoverability services
	FC#5	Data replication	PaaS	C	Data stored in community repositories, copied by users locally when necessary	No easy way to access the data, locate the data, and replicate it	Hide the data topology to the user, data federation, data replication capabilities
	FC#6	Network configuration	IaaS	O	Little network configuration capabilities, not extensively supported in the standard interfaces	Standard API does not allow complex network configuration, such as VPN-aaS, Firewall-aaS	Extend current standard interfaces to support network configuration

Haddock Portal	HD#1	Deployment / undeployment of a portal solution	Compute / PaaS	M	A haddock deployment consist on a portal instance connected to a back-end batch queue system that executes the jobs of a workflow run from the portal.	Fixed installation and lack of resources that is complemented through external grid infrastructures.	A mechanism to facilitate the deployment of a customised Haddock portal and backend in system in a panoply of infrastructures with minimal intervention.
	HD#2	Batch queue back-end	Compute / PaaS	M	The portal runs on a back-end which can be a dedicated batch queue or a grid executing service.	Installation of the back-end at local premises is fixed in size and resource availability (e.g. the number and type of dedicated resources are fixed on the installation time). Customized images for each platform.	A flexible mechanism for defining the processing back-end will enable portal administrators to deploy different back-ends with different configurations in terms of resources and capabilities. Users could bring their own credentials for their own clouds
	HD#3	Back-end customisation	Compute / PaaS	C	The configuration of the back-end is manually defined by a system administrator.	Multiple versions of the same components or data in a single deployment cannot coexist easily. Customization is manual.	A flexible mechanism for defining the processing back-end will enable portal administrators to better manage specific configuration, multiple portals or specific deployments for special users who could bring own-cloud credentials. Isolation of version could be used to enrich Haddock workflows with versioning information and enhanced provenance.
	HD#4	Privileged user access	PaaS	M	Privileged (root-level) users can login the portal and the back-end and perform operation maintenance and configuration.	Administrative access is not managed at central level.	Central-level management could unify the way different infrastructures can be accessed and maintained.
	HD#5	Regular user access	PaaS	M	Users of the portal are identified using SSOX or WeNMR Virtual Research Community credentials	No additional features	Feature must be preserved

	HD#6	Cloud bursting	Compute / PaaS	M	Users or administrators can define which is the rightmost queue to submit jobs (local / grid). Grid queues have a larger amount of resources.	No real cloud bursting which could derive an existing job workload to a different infrastructure.	The availability of redirect workload to external infrastructures gives the opportunity of defining different QoS and maximizing the usage of local infrastructures.
CIRM MP	CIR#1	Deployment of interfaces for simulation and post-processing	Compute / PaaS	M	An instance of the processing service involves two steps: simulation and analysis, which have different interfaces and are exposed in differents sections of the portal.	Deployment of the system and connection to the back-end is manually performed	Need for automatic customisation and interoperability with different infrastructures
	CIR#2	Deployment of back-ends based on batch queues	Compute / PaaS	M	Processing nodes are orchestrated through a batch queue.	Deployment of the system and connection to the back-end is manually performed	Need for interoperability with different infrastructures. No explicit mentioning to cloud bursting, although it may be of high benefit considering the approach followed.
	CIR#3	Customisation of back-ends with specific software	Compute / PaaS	M	Processing nodes have a specific software configuration which depends on a specific project.	Currently the configuration is static and manually set up. Shared by different projects running on the infrastructure.	Need for automatic customisation and isolation of multiple different virtual infrastructures with different requirements.
	CIR#4	Privileged user access	PaaS	C	Privileged (root-level) users can login the portal and the back-end and perform operation maintenance and configuration.	Administrative access is not managed at central level.	Central-level management could unify the way different infrastructures can be accessed and maintained.
	CIR#5	Regular user access	PaaS	M	Regular users access the platform through the portal / the WeNMR SSO.	No additional features	Feature must be preserved
	CIR#6	Long-term storage of output data	Storage	M	The portal does not provide long-term storage	Not fulfilled	The ability of having a service for long-term storage will open many possibilities for data reusing and post-processing analysis.

	CIR#7	On-line processing of data	PaaS	M	Post-processing is performed both when the final results are available or during the execution of the experiment to check its progress. A dedicated application in the portal is used. Lower consumption of resources and shorter processing time.	Seem to be linked to the final stage of jobs.	Special management of post-processing jobs that could be sent to the resources hosting the data and not vice-versa. Data analytic techniques may benefit.
	CIR#8	Special hw configuration - MPI, multicore, GPGPU	Compute / PaaS	M	The Amber suite can exploit HPC resources including multi core, MPI and GPGPUs	Manually defined by the administrators who should identify the resources, configure special software requirements on on-premise resources and select the rightmost grid queues and define the proper job requirements.	More flexibility in the way the requirements are defined and the matching with the infrastructure, as well as the automatic installation and configuration of software dependencies will be key.
	CIR#9	Monitoring and operation	PaaS	C	The portal and jobs must be monitored by privileged users to guarantee the QoS.	Specific system monitoring services	Keep functionality
DARIAH	DA#1	Distributed storage	Storage / PaaS service	M	DARIAH Bit preservation is a distributed storage architecture based on iRods offering https interface and a metadata database for indexation.		
	DA#2	Centralised IAA	PaaS service	M	Currently based on SAML		

Lifewatch - TRUFA	LWT#1	User access portal as SaaS	PaaS service	C	Single web access portal.	Access portal does not have strong requirements (different from computing backends) but it may be overloaded if a high number of concurrent user accesses.	Provide mechanisms for implementing scalable SaaS portals.
	LWT#2	Deployment of computing back-end	Compute / PaaS	M	The pipeline is explicitly installed in the computing nodes.	Need for a mechanism to deploy computing nodes easily on any infrastructure.	Improved capability on the access to resources and flexibility.
	LWT#3	Scalable back-end	Compute / PaaS	M	Scalability is currently managed through the LRMS	Valid for a static deployment where the computing nodes are shared through a batch queue, but inC if the resources are deployed exclusively for the portal.	Power-on/off virtual resources on demand.
	LWT#4	User authentication	PaaS Service	M	Local authentication in the portal	Valid for a static deployment but it may be insufficient for delegation in coordinated resources	Availability of a global IAA management.
	LWT#5	Local storage for users	Storage / PaaS service	M	Modified version of simogeo	Relies on local storage accessed through a web file system. It will need coordination if portal is a SaaS.	Ensure that the scalability of the storage will be sufficient for an increased workload.
	LWT#6	Dropbox-like storage access	Storage / PaaS service	C	Not provided	Not provided	Facilitate interaction with users in uploading and downloading files
	LWT#7	On-line postprocessing	Compute / PaaS	C	Not provided	Not provided	Capability of interactively accessing data online, without downloading the data and executing "R" or Python scripts on top of it.
	LWT#8	Output data management	Storage / PaaS service	C	Not provided	Not provided	Facilitate downloading large files in an asynchronous model.

Lifewatch - ALGAE BLOOM	LWAB #1	Model processing (hydrodynamic, water quality)	Compute / PaaS	M	Running in Altamira HPC, EGI FedCloud or local resources	Not scalable for many models	Find the best solution through parametric exploration. It requires to run the model several times changing some parameters values.
	LWAB #2	Distributed storage (dropbox like)	Storage / PaaS service	M	Replicated data in different resources. Sharing data with SugarSync (Dropbox like)	Cloud or grid based solutions have not proven to be efficient yet.	Cloud back-end will facilitate the deployment on a wider range of infrastructures.
	LWAB #3	Online postprocessing	Compute / PaaS Service	M	Post-processing is done by Delft3D tool called Quickin based on Matlab. This tool is installed locally or in a Windows Virtual Machine.	Graphical User Interface need Windows to be used. Post Processing alternatives could be developed.	Developed post processing online tools based on R or python
	LWAB #4	Data & Metadata Management	Storage / PaaS service	C	Data is stored in local resources (on site database and replica). Metadata is planned but not yet implemented		Store data in resource provider. Implement metadata management system.
	LBT#1	Data policy	PaaS Service	M	Data policy managed by a database with fixd terms	Configuratipn on cloud authorization policy	Link between data policy db and cloud configuration for data distribution and access. Not already federated clouds common configuration is required.
INAF.LBT	LBT#2	Distributed storage	Storage / PaaS service	M	An existing middleware (TANGO) requiring special network parameters.	Cloud or grid based solutions have not proven to be efficient yet.	Cloud back-end will facilitate the deployment on a wider range of infrastructures.
	LBT#3	User authentication	PaaS Service	M	Portal based and managed locally or through the VO-DANCE (Virtual Observatory).	No additional features	A SSO mechanism is needed with regular user registration and linked to an authorization mechanism.
	LBT#4	Data retrieval	Storage / PaaS service	M	A database is exposed and searched through a Java application that also downloads data. Virtual Observatory services also access this database.	More efficient mechanisms are needed.	An enhanced on-line processing will reduce the need for downloading, and other mechanisms of asynchronous synchronization may help.

	LBT#5	Data reduction	Compute / PaaS	M	Data is preprocessed and calibrated, generating additional data and metadata Processing is done locally and then distributed to PI.	Deployment is manual and fixed.	The availability of a variable range of resources will reduce this processing stage and make it more versatile.
	LBT#6	On-line processing of data	Compute / PaaS	M	Data is inspected but not fully processed remotely. Data is downloaded locally and processed using the pipeline tools (written in java) provided by the community.	Fully online processing is missing.	It will reduce the requirements on the client side.
	LBT#7	Subset managment	PaaS	O	Not yet implemented		Identify a set of data and/or metadata included into a paper available to the public access.
INAF-CTA	CTA#1	Scientific GateWay	SaaS	M	Many services/applications are currently implemented and already in use in CTA, and will evolve into a single web-oriented global platform of services.	current technology is limited in supporting user to easily access and manipulate the huge amount of CTA data	A web-based community-specific set of tools, applications, and data collections that are integrated together via a web portal, providing access to resources and services from a distributed computing infrastructure. The Gateway aims at supporting workflow handling, virtualisation of hardware, visualization as well as resource discovery, job execution, access to data collections and applications and tools for data analysis. It may potentially host all monitoring services of data operations and some remote control or monitoring applications for instruments and devices. It shall be a software repository for development tools, version track services and software validation test benches.
	CTA#2	Single Sign-On(AAI)	PaaS	M	is under development	is required	based on each user's profile and category (e.g., basic, advanced users, managers, collaboration users etc.)

	CTA#3	Easily handle & manage huge amount of data	SaaS/PaaS	M	Grid solutions, use of DIRAC to integrate heterogeneous resources	handling and archiving the large-amount of data generated by the instruments and delivering scientific products according to astrophysical standards is one of the challenges	
	CTA#4	Effective long-term storage	PaaS	M	CTA Archive System	avoid silent corruption in long-term storage system, which is the worst type of errors because are unnoticed and propagated (tests shown that every 15 mins a silent corruption rise in a long term storage) resulting in cascading failures	Maximize Consistency, Availability & Partition tolerance
EMSO - MOIST	MO#1	Distributed data storage	Storage / PaaS service	M	Locally deployed solution	Real multi-site distributed storage is not provided	Higher scalability and improved performance.
	MO#2	Metadata storage	Storage / PaaS service	M	Multiple databases and user community formats	Real multi-site distributed storage is not provided	Interoperability
	MO#3	Efficient data retrieval	Storage / PaaS service	M	Network requirements	More efficient mechanisms are needed.	Increase availability, robustness and reduce actual transferring time.
	MO#4	On-line processing of data	Compute / PaaS Service	C	Downloading of data and further processing on user's' computer.	Not available - it will reduce impact on data retrieval.	On-line access to "R" and similar problem solving environments will reduce the need of downloading data and improved accessibility to data.
	MO#5	Tracking user access and authorisation	PaaS Service	M	Local authentication in the portal	Increased Cybersecurity	SSO and tracking user access. Improved capability of response to attacks
ICCU	ICCU# 1	Large data storage	Storage	M	A system called MAGTeca is provided to the community for managing catalogues.	Local and different solutions at each library.	Improved capability of management big amounts of data

	ICCU# 2	On-line access to data	Storage / PaaS service	M	free-text google-like searches in less than 0,2 seconds for large dataset. Unclear if they have/will have on-line processing.	Local and different solutions at each library.	Need for a unique gate access to different libraries
	ICCU# 3	Metadata management and interoperability	Storage / PaaS service	M	the data of the community follows specifications such as MAG, METS, PREMIS and uses OAI-PMH for interoperability. There are other requirements (e.g. a semantic search engine) which should be at application level.		Improved management of metadata
	ICCU# 4	Centralised IAA	PaaS Service	M	ECSG implements an authentication schema based on Identity Federations and in the DCH-RP project was realized the Identity Provider (idP) by ICCU, thanks to GARR, the Italian NEREN.	Enhanced security is required, although no clear statements are found in the document.	keep functionality and provided increased authorization mechanisms.