INDIGO-DataCloud

INITIAL REQUIREMENTS FROM RESEARCH COMMUNITIES ANNEX 1.*P0*: SELECTED CASE STUDY FROM MEDICAL IMAGING BIOBANKS – POPULATION IMAGING (EUROBIOIMAGING)

INPUT TO EU DELIVERABLE: D 2.1

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

Abstract

This report summarizes the findings of T2.1 and T2.2 **for partner P0** along the first three months of the project. It is an integrated document including a general description of the research communities involved and the selected Case Studies proposed, in order to prepare deliverable D2.1, where the requirements captured will be prioritized and grouped by technical areas (Cloud, HPC, Grid, Data management) etc. The report includes an analysis of DMP (Data Management Plans) and data lifecycle documentation aiming to identify synergies and gaps among different communities.

I. COPYRIGHT NOTICE

II. DELIVERY SLIP

	Name	Partner/Activity	Date
From	<<**The author/editor in P0**>>	**P0**/WP2	
Reviewed by	**Moderators:** P.Solagna, F.Aguilar, J.Marco **Internal Reviewers:** <<To be completed by project office on submission to PMB>>		
Approved by	**PMB** <<To be completed by project office (no submission)>>		

III. DOCUMENT LOG

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

TABLE OF CONTENTS

0 INTRODUCTION AND CONVENTIONS

PLEASE, READ CAREFULLY BEFORE COMPLETING THE ANNEX:

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

There will be around 10 Annexes, for example Annex 1.P0 for partner 0 in WP2 (i.e. CSIC), will cover Case Studies from LifeWatch research community (an ESFRI initiative).

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

Some relevant definitions:

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

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

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

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

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

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

RESEARCH MANAGERS

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

RESEARCHERS

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

APPLICATION DEVELOPERS AND INTEGRATORS WITHIN THE RESEARCH COMMUNITIES

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

MIDDLEWARE DEVELOPERS AND E-INFRASTRUCTURE MANAGERS

-Sections 7, 8: should be discussed with them

---

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

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

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

Other conventions and instructions for this document:

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

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

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

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

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

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

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

1 EXECUTIVE SUMMARY ON THE CASE STUDY

Summarize the research community applications/plans/priorities (max length 2 pages).
To be completed after sections 2-6. Supervision by a senior researcher is required.

Comentario [I1]: It depends. I think that it can be filled-in after section 2. Not necessary to go through sections 3-6.

If we go through technical details we may lose the interest by the senior reader.

Indeed senior supervision

1.1 Identification

- *Community Name*: Medical Imaging Biobanks –Population Imaging (EuroBioImaging)
- *Institution/partner representing the community in INDIGO:* UPVLC
- *Main contact person*: Ignacio Blanquer
- *Contact email: iblanque@i3m.upv.es*
- *Specific Title for the Case Study:* Medical Imaging Biobanks

1.2 Brief description of the Case Study and associated research challenge

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

Comentario [I2]: I think that this is also referr in section 2.3. I suggest removing "timeline".

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

According to the definition from the EuroBioImaging web site, Population imaging is 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 significant. Three fundamental barriers arose. 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 database 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 is needing 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 signging a Joint Research Unit Agreement for regulating their participation in projects. The main mandate of BIM-CV is to create such population imaging node.

1.3 Expectations in the framework of the INDIGO-DataCloud project

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

The main expectation is the consideration of a software framework that can be deployed on state-of-the-art infrastructures for acquiring, managing, processing and analysing medical imaging data. We envisage that cloud computing technologies could provide the proper isolation, computing capacity and flexibility to deploy customized environments for population imaging to develop.

INDIGO could address the requirements of data access, privacy management, automation of virtual infrastructure configuration and deployment that a research infrastructure for population imaging must address.

1.4 Expected results and derived impact

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

Several types of projects that we could find are:

1 *Massive processing of a large set of data using a pipeline of existing tools.*

2 *Searching on a massive database for specific context-aware data related to a suspicious of diagnosis.*

3 *Assessment of the benefits of new processing tools by comparing to gold standards.*

4 *Large-scale training of data classifiers.*

1.5 References useful to understand the Case Study

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

More information can be found in:

[1] Ignacio Blanquer, Miguel Caballer, Luis Martí-Bonmatí, Ángel Alberich-Bayarri, Jacobo Martínez , Maria de la Iglesia, "A Cloud Infrastructure for Scalable Computing on Population Imaging Databanks". International Journal on Image Mining, accepted and pending for publication.
https://www.researchgate.net/publication/277307625_A_Cloud_Infrastructure_for_Scalable_Computing_on_Population_Imaging_Databanks

[2] María de la Iglesia-Vayá, José María Salinas, Gonzalo M Rojas, Juan Carlos Pérez Cortés, Rafael Llobet, Miguel Angel Cazorla, Jacobo Martínez, Luis Martí-Bonmatí, Ignacio Blanquer, Manuel Regaña and José Miguel Puig, "BIMCV: Synergy between Peta Bytes of data in population medical imaging, computer aided diagnosis and AVR". Studies in health technology and informatics 01/2015; 210:987-9.

[3] Kick-off meeting of the initiative (in Spanish) https://www.acim.lafe.san.gva.es/acim/?p=1056

2 INTRODUCTION TO THE RESEARCH CASE STUDY

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

2.1 Presentation of the Case Study

Describe the Case Study from the research point of view

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.

2.2 Description of the research community including the different roles

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

Describe the research community specifically involved in this Case Study

The community is composed of three level of researchers:

1 *Practitioners (medical doctors), who need data and processing tools to prove their hypothesis on diagnosis and treatment.*

2 *Pharmacy, who XXXX*

3 *Medical informatics, who develop new biomarkers, knowledge extraction systems and other image processing tools and want to extensively validate on a large cohort of data and compare the results to those obtained by reference tools and procedures.*

2.3 Current Status and Plan for this Case Study

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

[3] http://dicom.nema.org

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

BIM-CV was positively evaluated in the call for Expressions of Interest to become a EuroBioImaging node. The process of nominating official EuroBioImaging nodes will be opened in the short future.

BIM-CV is willing to develop a pilot Virtual Biobank service in the coming months for evaluation purposes. If INDIGO takes into account requirements that address the issues in BIM-CV, BIM-CV will define the architecture taking into account the software pieces to be developed and integrated.

The exploitation of the Virtual Biobank will come through the participation in EuroBioImaging, which will open the access to the infrastructure to a wider public.

BIM-CV Virtual Biobank is an ongoing activity. Currently it has a limited support and lacks from flexibility to customize the environment to different users. In its current status, it exposes an XNAT portal connected to a LONI pipeline which executes in a batch back-end.

The availability of a prototype by the end of 2015, planning to enter in low-scale production along 2016 will be key. From the interest letters received, we envisage to support more than 20 subprojects per year, both at national and international level.

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

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

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

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

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

2.5 General description of e-Infrastructure use

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

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

The infrastructure will be supported by the HUPLF and the UPV. UPV Federated Cloud site is willing to contribute with resources to the use case. Currently, it is not using any external infrastructure and executions are performed on bare-metal configurations.

2.6 Description of stakeholders and potential exploitation

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

Describe the exploitation plans related to this Case Study

After presenting the visions and mission of BIM-CV, a number of research institutions across Europe have shown their interest in the facilities and the population imaging data that could be provided. The number of expressions of interest have been more than 30 with 15 project defined at national level and 15 projects at international level. In summary, there are 10 expressions of interest that present projects in the area Neurosciences, 6 in the area of Cardiology, 6 in the area of Oncology, 5 dealing with Biomarkers and 8 led by TIC centres. A summary is provided below.

In the medical research, there are three main areas of research identified from the LoIs received:

- *Cardiology, such as the cases from Univ. of Sheffield, INRIA, Univ. Pompeu Fabra, Hospital Clinic de Barcelona, and Kings College London and Sant Tomas Hospital).*
- *Oncology, in the diagnosis (in a general scope – Fund. Champalimaud, Univ. di Roma, Maastricht Univ, etc. and in particular areas, such as Breast Cancer by INEGI), and therapy, such in the simulation of radiotherapy doses (CESGA).*
- *Neurosciences, with interest on mental disorders (e.g. CIBERSAM, Univ. Cambridge, Max Planck Institute, Nijmegen University) and neurodegenerative diseases (e.g. CIPF, INCLIVA, Univ. of Alicante, Hosp. Vall d'Hebron).*

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

Area	***Impact Description***	***KPI Values***
Access	*Increased access and usage of e-Infrastructures by scientific communities, simplifying the "embracing" of e-Science.*	• *Number of ESFRI or similar initiatives adopting advanced middleware solutions* ESFRIs: *1 (EuroBioImaging)* • *Number of production sites supporting the software 3*
Usability	*More direct access to state-of-the art resources, reduction of the learning curve. It should include analysis platforms like R-Studio, PROOF, and Octave/Matlab, Mathematica, or Web/Portal workflows like Galaxy.* *Use of virtualized GPU or interconnection (containers).* *Implementation of elastic scheduling on IaaS platforms.*	• *Number of production sites running INDIGO-based solutions to provide virtual access to GPUs or low latency interconnections 1* • *Number/List of production sites providing support for Cloud elastic scheduling 3* • *Number of popular applications used by the user communities directly integrated with the project products: 10* • *Number of research communities using the developed Science Gateway and Mobile Apps: 2 (medical imaging, and medical informatics)* • *Research Communities external to INDIGO using the software products: 2, same as before*
Impact on Policy	*Policy impact depends on the successful generation and dissemination of relevant*	• *Number of contributions to roadmaps, discussion papers: 0.*

Comentario [I3]: Typo: Matlab

	knowledge that can be used for policy formulation at the EU, or national level.	
Visibility	*Visibility of the project among scientists, technology providers and resource managers at high level.*	• *Number of press releases issued: 1 per year* • *Number of download of software from repository per year: 40* • *List of potential events/conferences/workshops: 20-30 by the end of the project, referencing INDIGO* • *Number of domain exhibitions attended 25* • *Number of communities and stakeholders contacted If considering research societies, 10 by the end of the project.*
Knowledge Impact	*Knowledge impact creation: The impact on knowledge creation and dissemination of knowledge generated in the project depends on a high level of activity in dissemination to the proper groups.*	• *Number of journal publications: 5 referencing INDIGO* • *Number of conference papers and presentations: 20-30 referencing INDIGO*

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

3 TECHNICAL DESCRIPTION OF THE CASE STUDY

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

3.1 Case Study general description assembled from User Stories

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

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

3.2 User categories and roles

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

The platform will require several types of user categories:

- *Sub-project users. Researchers that access a customized platform and data. They should be able to access the resources of the infrastructure associated to a specific subproject. Radiologists, medical informatics, and disease-specific researchers will be users at this level.*
- *Platform operator. This refer to ICT experts who operate and monitor the global services to ensure the proper performance of the system.*
- *External users who would access the platform for awareness.*

3.3 General description of datasets/information used

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

The main dataset to be used is the PACS and RIS of the HUPLF. This information will be complemented with the Vendor Neutral Archive (VNA) repository of the Valencia Regional Authorities. This dataset covers the retrospective information of the last 10 years from a population of nearly 300.000 inhabitants. The VNA will gather the whole population of the Valencia Region (5 million people).

3.4 Identification of the different Use Cases and related Services

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

One suitable target application of population imaging is the study of highly prevalent diseases with a long sub-clinic phase that can be diagnosed using imaging. Four cases are of interest have been depicted by the experts in the field:

- *Non-alcoholic fatty liver disease.*
- *Neurodegeneration.*
- *Pulmonary emphysema.*
- *Osteoporosis.*

Osteoporosis is a major degenerative process that has a great impact on the population. The process remains silent (without external evidences) until a fracture is produced. Hip joint fracture is an habitual consequence of osteoporosis. In the advent of a hip joint fracture, a third of the population dies in the following months, and another third suffers from permanent disabilities. Early diagnosis is key for guiding treatments and healthy habits.

An important tool will be the creation of osteoporosis atlases characterized by population and other concurrent factors. This can be obtained by processing a large amount of CT images from a whole population. Millions of abdominal CT studies are performed daily in the world for multiple purposes. The radio-opacity of the spine bones can be used to assess the risk of suffering an osteoporosis event. Moreover, restrospective studies have proven to be a good surrogate of the osteoporosis diseases. If an automatic method for the segmentation and analysing of the spine bones would be available, practitioners could have a population-level map of the biological age of bones. This information can be used to trigger alarms and advice practitioners when dealing with a concrete patients' health.

The map will require a huge amount of CT information, an automatic segmentation system and an automatic analysis mechanism.

- *Huge data is currently available. Modern abdominal CT images have a high resolution. Abdominal CT exams are huge (around 1 GB per study), and include also clinical metadata. Anonymized data in nifty format (http://nifti.nimh.nih.gov/nifti-1) can be used.*
- *Images need to be pre-processed to extract quality indicators, filtered and de-noised. The use of GPU-based software will be desirable.*
- *Automatic segmentation in population-level data is currently using deep learning techniques. This requires large sample data and non-trivial computing capabilities. A tool called Caffe is currently used for this purpose. Zebra imaging is currently using it.*
- *Bone density analysis is done with researcher analysis methods written in Matlab and python. Reports are written in Jasper.*
- *Visualization can be done in web-based viewers, such as dcm4chee.*

3.5 Description of the Case Study in terms of Workflows

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

The interaction with the Virtual Biobank for Population Imaging is defined in the following steps:

1. *Definition of a subproject. A subproject will comprise four main blocks of information: data sample request (e.g. MRIs from males above 30 with an early onset of dementia); software requirements (e.g. Caffe, dcm4chee, Matlab, python); computational requirements (in terms of maximum simultaneous processing units – each subproject will have a quota); access credentials.*

2. *Sub-repository population. Data will be extracted and properly anonimized from clinical practice repositories. These anonymized data will be uploaded on a persistent working repository that must allow POSIX access.*
3. *Deployment of the interaction portal and access. A portal (XNAT + LONI + other tools) will be deployed. Each subproject consists of at least two types of instances: Front-end and processing nodes. Processing nodes may be increased as computing is required. Interaction can be done through batch queues. Deployment should be platform-agnostic, and eventually including cloud-bursting.*
4. *Data visualization and inspection. The portal will include a graphical user interface that will allow browsing the data and the results.*
5. *Workflow execution. Along with the tools, pre-existing workflows (pipelines) will be available. Those pipelines will have default values that could help naïve users. Workflows will be launched from the front-end panel and executed on a set of Working Nodes, which must not be powered-on unnecessary.*
6. *Uploading new processing components and workflow update. Workflows may be cloned and updated to enable integrating own tools. This is the use case for Medical Imaging developers.*

Additionally, other administrative use cases are envisaged

1. *Access control and accounting. Subprojects may have an access quota or could use public infrastructures. In such cases accounting is critical for the user.*
2. *Infrastructure monitoring. Virtual infrastructure should offer a high degree of reliability and different actions should be taken in case of failure (notifying special user, automatic redeployment, etc.).*

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

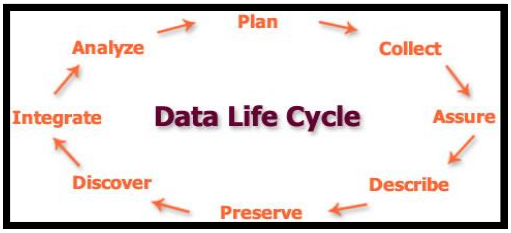
Indicate the current deployment framework (cluster, Grid, Cloud, Supercomputer, public or private) and the relevance for the different Use Cases of the access to those resources.

Currently, the use case is not in a fully operational status. Other sub-projects are already available and are deployed on a local cluster. This impede to scale up and to expose a customizable environment.

4 DATA LIFE CYCLE

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

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



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

4.1 Data Management Plan (DMP) for this Case Study

According to EU H2020 indications[4], following UK DCC tool indications

4.1.1 Identification of the DMP

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

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

Principal Researcher: **<input here>**

DMP Manager: **<input here>**

Description: **<input here>**

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

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

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

For each data set provide:

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

Data set reference and name HUPLF PACS data

Data set description DICOM images from routine studies, properly anonymised and annotated with radiology reports

Standards and metadata mainly DICOM

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

Connection to Instrumentation,

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

Data is produced by Computer Tomography scanners, and fed into the PACS systems. There is no direct relation of the use case and the instrumentation apart from registering model and manufacturer within the metadata.

Vocabularies and Ontologies

Are they relevant? Internal vocabularies related to the specific fields. RDA groups.
(pending definitive form, see next sections)

The use case requires three types of information: Medical Images, associated clinical data and provenance information. Standards are used for the three types of data.
Medical Images are stored in DICOM (http://dicom.nema.org/). DICOM standard defines a tag-based specification to code the metadata and actual data of medical images. Despite that DICOM is accepted by all current manufacturers of medical devices, there are extensions and specificities inherent to each manufacturer. DICOM also defines communication protocols with medical scanners, although communication in this use case will be made mainly with the PACS systems, which typically expose a relational DB interface and store the objects in the native DICOM format.
Associated clinical data comes from three main sources: DICOM header metadata (which may include demographic information and acquisition parameters); Radiology reports, which can be found in plain text or coded in a structured format such as RadLex (http://bioportal.bioontology.org/ontologies/RADLEX), depending on when the data was produced; and specific data from the Electronic Health Records (diagnosis and other terms properly coded).

Finally, provenance information will be recorded to annotate the workflows and runs. Up to now, there is no information collected and no agreement on the specific format to be used.

Data Capture Methods

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

• *How will the data be created? Extracted from regional PACS and hospital PACS and complemented with the associated metadata.*

• *What standards or methodologies will you use? DICOM and IHE recommendations*

• *How will you structure and name your folders and files? DICOMDIR*

• *How will you ensure that different versions of a dataset are easily identifiable? Each subproject will be provided with a pre-processed anonymised copy of the data of interest from the repository.*

Metadata

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

• *How will you capture / create the metadata? Metadata is associated to the DICOM objects in the DICOM header, and complemented with radiology reports and selected items from the Electronic Health Records.*

• *Can any of this information be created automatically? Not yet. It requires a data collector who merges the relevant data from the*

• *What metadata standards will you use and why? IHE recommendations for DICOM data. ICD 9 and RADL terminologies whenever possible.*

Data sharing

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

The original dataset comes from an clinical practice and therefore cannot be used directly for data sharing. However, these data are the source for the creation of the individual, sharable, and pseudo-anonymised data subrepositories. In the use case proposed for this requirement analysis, there is no risk of re-identification from the plain information of anonymised

medical images (abdominal area). In the case of neuroimaging, which could be exposed to facial recognition, images will be pre-segmented removing cranial features. Data is stored in an institutional repository of a public agency.

Access to data extracted from the BioBank will be given under non-disclosure conditions. Users who have access granted must acknowledge BIMCV on their publications and results of the research derived from the use of the data and tools provided by the node. Users must fill-in brief periodic progress reports and do not perform any commercial exploitation of the data.

BIMCV will get track of the data released to each subproject to guarantee the traceability of the results that may arise from the analysis, guided by ethical principles defined in their general ethical plan.

Method for Data Sharing

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

• *How will you make the data available to others? Through the specific request by means of competitive sub-project calls. Then, data will be extracted, pre-processed and exposed as a separate repository.*

• *With whom will you share the data, and under what conditions? Subprojects positively evaluated in the competitive calls will have access granted to the data. Free access will be given to public research institutions. Processing requirements, additional consultancy and support and other services will be properly accounted and billed if they imply incurring in external costs not covered from the platform. This will be clearly declared in the service catalogue of the centre.*

Restrictions on Sharing

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

• *Are any restrictions on data sharing required? e.g. limits on who can use the data, when and for what purpose. Personal data is excluded. Data and services will be made available to non-commercial research positively evaluated by an accession committee. IPRs for knowledge arising from the sub-projects will be attained by the sub-project participants. The node will only require a share of the IPRs if a significant involvement in the subproject takes place (e.g. customization of tools, experimentation and parameter tuning, provision of models, etc.).*

• *What restrictions are needed and why? Data cannot be reused for purposes different from the ones stated in the sub-project objectives nor released to third-parties, requiring an explicit authorisation from the BIMCV node. However, renewal of projects to continue using*

the data will be facilitated. Moreover, the IP of the sub-project will be entitled to keep a copy of the data beyond the scope of the sub-project, to facilitate additional actions that the long-term process of scientific publication may require.

• *What action will you take to overcome or minimise restrictions? Meaningful public research will have access granted. The principles of the accession committee are open and will only depend on the node capability to meet the expected demand.*

Data Repository

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

• *Where (i.e. in which repository) will the data be deposited? Currently, the platform uses its own resources for storing the temporal sub-project repositories. Clearly, this is not scalable and other alternatives are being analysed.*

Archiving and preservation (including storage and backup)

Questions to consider before answering:

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

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

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

• What additional resources are needed to deliver your plan?

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

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

• Will charges be applied by data repositories?

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

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

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

Indeed, this is a key aspect for the sustainability of the research. Currently, the preservation of the raw data is guaranteed by healthcare provider legal obligations. The sub-projects data sets will require:

- *A local storage closely accessible from a computing to efficiently apply the processing services of the node.*
- *A longer-term storage for keeping the outcomes and results of the sub-project, as well as to back-up the raw data in case that further sub-projects are implemented. Agreements with third party repositories is envisaged.*

In order to facilitate the long-term preservation, BIM-CV envisages a model in which data and processing services can be embedded in the form of a Virtual Appliance.

Indicate how long the data should be preserved, what is its approximated end volume, what the associated costs are and how these are planned to be covered. The sub-project data is expected to be kept beyond the end of the sub-project and to deal with the necessary actions required during the publication process of the results obtained. No strict deadline will be given, although data by itself may become obsolete after a long period of time (new techniques, new devices and new protocols will surely arise). The end-volume may reach few TBs and preservation costs are on the sub-project budget.

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

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

Questions to consider:

• How will potential users find out about your data?

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

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

*Are the data and associated software produced and/or used in the project discoverable (and readily located), identifiable by means of a standard identification mechanism (e.g. **Digital Object Identifier**)? It is envisaged that the main source of requests will come through the EuroBioImaging ESFRI. The ESFRI will publicise the sub-project open calls. BIM-CV node will advertise its data and features.*

SCIENTIFIC RESEARCH DATA SHOULD BE **ACCESIBLE**

Questions to consider:

• Who owns the data?

• How will the data be licensed for reuse?

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

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

State who will own the copyright and IPR of any new data that you will generate. For multi-partner projects, IPR ownership may be worth covering in a consortium agreement. If purchasing or reusing existing data sources, consider how the permissions granted to you affect licensing decisions. Outline any restrictions needed on data sharing e.g. to protect proprietary or patentable data. See the DCC guide: How to license research data.

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

The ownership of the data will be the Valencian Health Authorities. Data will be licensed for the specific research objectives of the sub-projects. Extension of this license must require an explicit request and consent by the BIM-CV. Reuse of the data by third parties will require extending the sub-project and the evaluation of the access committee and the ethic body. The IPR of the results obtained in a sub-project belong to the sub-project partners. Each sub-project will include a preliminary exploitation plan defining the mechanisms for sharing the IPRs.

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

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

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

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

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

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

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

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

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

Guidance on Metadata:

• How will you capture / create the metadata?

• Can any of this information be created automatically?

• What metadata standards will you use and why?

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

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

Are the data and associated software produced and/or used in the project useable by third parties even long time after the collection of the data? e.g. is the data safely stored in certified repositories for long term preservation and curation; is it stored together with the minimum software, metadata and documentation to make it useful; is the data useful for the wider public needs and usable for the likely purposes of non-specialists? See item related to accessibility of data in this sub-section.

INTEROPERABLE** TO SPECIFIC QUALITY **STANDARDS

• What format will your data be in?

• Why have you chosen to use particular formats?

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

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

See the UKDS Guidance on recommended formats

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

The data of a sub-project will be packaged into the formats and terminologies stated before (DICOM, Nifty, IHE, ICD, etc.). They are the de-facto standards in research and industry since more than 20 years ago.

Quality is a key factor in imaging data. Quality depends on multiple factors, and strongly affects the results, which sometimes highly depend on subtle variations in the images. All the images released will go through an automatic quality analysis procedure which will check the ratio signal-noise and other key parameters before being released. Images below the quality thresholds will not be released.

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

4.2.1 General description of data levels

Indicate if the DATASETS are organized into different levels (LEVEL-0, 1, 2, 3, 4) and if so what are the relevant definitions and how DOI are provided. Two levels of datasets are defined. Source reference dataset, which includes personal information and it is not directly accessible. Second, each sub-project will have a separate anonymised sub-dataset.

4.2.2 Collection/Acquisition

Gathering RAW data

Specify how do you gather/collect your data (e.g. sensors, observations, satellites, etc.)? Routine clinical data.

How do you pre-process, transfer and store your RAW data? Data is extracted manually from the clinical repository and enriched with additional clinical metadata (diagnosis and treatment mainly). Then, data is pseudoanonymised, keeping locally at the hospital the relation between the identifiers.

From RAW Data to Calibrated Data

Describe the processes applied for Data Calibration, Validation, Filtering, etc. Data filtering and validation is performed as part of the post-processing, including denoising, intensity homogenisation and quality control.

4.2.3 Access to external data

Describe the identification and access to External Data. Users may bring their own complementary data, although there is no procedure for accessing reference external datasources.

Indicate if there is a procedure for validation of External Data. Same filtering and quality control as local data, as explained in previous section.

4.2.4 Data curation

Specify any automatic check applied, like completing series, detecting outlier. Quality control procedures are applied to compute the noise/signal ratio.

Describe manual quality checks. Potentially by visual inspection from the experts

Are there quality flags applied to the data? Not planned.

4.2.5 Data ingestion / integration

Describe transformations applied to data taking into account ontologies/metadata. Indicate also if there is any "harmonization procedure" (to share/integrate data) and how linking internal and external data is made if relevant. Medical Imaging data is complemented with ICD diagnostic information. No structured radiology reports are available, although it is planned to be included in the future. In this case RADLEX will be used.

4.2.6 Further data processing

Describe, if relevant, the different additional processing steps (and the associated software and resources) applied to the (collected/curated) datasets to provide a "final" dataset collection that can be used in the analysis Described in the following section.

4.3 Analysis

4.3.1 Basic analysis and standard analysis suites

Describe usual examples of basic analysis in the Case Study

The basic analysis go through several steps. First, the data selected for a sub-project needs to be pre-processed to guarantee anonymization and quality.

- *Anonymisation and format conversion.*
- *Quality analysis (homogeneity, noise/signal ratio, filtering).*

Second, the analysis go through the following steps

- *Automatic segmentation of spine bone through deep learning techniques. This process requires a set of training steps to adjust the different neuron levels before the model is able to recognize bone tissues in new images.*
- *Analysis of the bone density using own image processing functions.*
- *Visualization.*

Specify if software packages/tools like MATLAB, R-Studio, iPython,etc. are used

The following is a list of the tools and software packages required:

- *DICOM data processing (anonymisation, format conversion, quality): dcm2nii (MRIcron package, http://www.mccauslandcenter.sc.edu/mricro/mricron/), dcmtk (http://dicom.offis.de/dcmtk.php.en), DICOM Cleaner (http://www.dclunie.com/pixelmed/software/webstart/DicomCleanerUsage.html).*
- *Massive segmentation: caffe (http://caffe.berkeleyvision.org/) a deep learning framework made with expression, speed, and modularity in mind.*
- *MATLAB, to execute the processing functions that compute a bone age estimation based on the features of the region of interest (spine bone) previously segmented.*
- *Data organization and visualization: XNAT (http://www.xnat.org/), dcm4chee (http://www.dcm4che.org/)*
- *JasperReports for presenting information from the database (http://community.jaspersoft.com/project/jasperreports-library).*

All except Matlab are open-source tools. Matlab compiled code may be used without additional license.

4.3.2 Data analytics and Big Data

Describe relevant examples of advanced analysis in the Case Study (like for example application of neural networks, series analysis, etc.) This use case plans to use deep learning for the automatic segmentation of spine bones.

Specify the resources and additional software required. The use case will use caffe (http://caffe.berkeleyvision.org/) and a yet not estimated amount of resources.

Identify analysis challenges that can be classified as "Big Data". The problem has some features in common with Big Data challenges. Data volume is huge, but not due to a massive amount of individual records, but to the huge size of the individual images. Non-conventional knowledge extraction is needed.

List Big Data driven workflows. Not yet defined.

4.3.3 Data visualization and interactive analysis

Indicate the need for data and analysis results visualization The visualization of medical images can be performed using web-based applications, such as dcm4chee.

Indicate how visualization is made and if interactivity/steering is needed. Final results can be visualized on Matlab compiled applications. Final results are exposed as reports.

Specify the User Interfaces (web, desktop, mobile, etc.) Web-based interfaces are preferred.

4.4 Data Publication

Describe the information flow from the analysis to the publication The result of the analysis will be stored in databases and will be the basis for the scientific articles.

Indicate the requirements from publishers/editors to access data, and how it is made available (open data?) Not all the data needs to be made available, even no data at all (enabling users to provide own data), for the use case defined in the document.

5 SIMULATION/MODELLING

Describe the Simulation/Modelling requirements in this Case Study

5.1 General description of simulation/modelling needs

Describe the different models used (including references) The use case will use deep learning classification models.

Indicate the type and quantity of simulations needed in the Case Study, and how they are incorporated in the general workflow of the solution Experiments are currently in the way. Not yet defined.

5.2 Technical description of simulation/modelling/postprocessing software

For each simulation package:

Identify the simulation/postprocessing software. Described in section 4.3.1.

Provide a link to its documentation, and describe its maturity and support level. All the applications are mature and maintained. Links are included in section 4.3.1.

Indicate the requirements of the simulation software (hardware: RAM, processor/cores, extended instruction set, additional software and libraries, etc.). Described in section 7.

Tag the simulation software as HTC or HPC. HTC, but embedded in a dataflow.

List the input files required for execution and how to access them. DICOM image files.

Describe the output files and how they will be stored. Reports and postprocessed images.

Reference an existing installation and performance indicators. See section 4.3.1

Specify if the simulation software is parallelized (or could be adapted). Partially.

Specify if the simulation software can exploit GPUs. Yes, Caffe can.

Specify how the simulation software exploits multicore systems. Yes in some of the cases

Specify if parametric runs are required Orchestrated by the pipeline engine.

Estimate the use required of the resources (million-hours, # cores in parallel, job duration, etc) Described in section 7.

5.3 Simulation Workflows

Describe if there are workflows combining several (HTC/HPC) simulations or simulations and data processing. Yes, all the processing steps are orchestrated by pipeline software. They combine processing steps with different memory and computing requirements. At least one of the processing steps will benefit from using GPUs.

6 DETAILED USE CASES FOR RELEVANT USER STORIES

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

6.1 Identification of relevant User Stories

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

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

For each relevant User Story:

Case 1: The platform operator creates a subprepository for a new subproject to be inserted in the platform		
Draft Use Case	*Sup. definition*	*Requirements & Acceptance*
The platform operator extracts the data from the clinical database and exposes them to the anonymiser tool, which will go through all those data item and will create pseudoanonymised copies of the input files. The output files will include a unique id that will link to the original images and the relation will be kept locally. Pseudoanonymised data will be validated and then transfer to the sub-repository storage.	*Flow diagrams.*	*Pseudo-anonymization and data index generation must be performed on a resource located in a safe area[5] (such as a DMZ or similar). Pseudoanonymised data should be stored in a way that can be accessible as any other POSIX filesystem.*

Case 2: The platform operator selects the rightmost tools to be used in the subproject		
Draft Use Case	*Sup. definition*	*Requirements & Acceptance*
The platform operator choose the software and virtual hardware configuration of the Virtual Appliance that will process the data. The Virtual Appliance will be composed of several virtual machines and the software configuration depends on the subproject. The software must be installed automatically on the machines and it	*Mock-ups, software diagrams.*	*Installation should be independent of the back-end, so minimal preconfiguration is required. Configuration should be self-scalable if possible and should be able to be monitored.*

[5] http://ec.europa.eu/justice/data-protection/index_en.htm

should expose an interface to orchestrate the executions.		

Case 3: The sub-project user is provided with access to a customized virtual appliance with access to data and tools		
Draft Use Case	*Sup. definition*	*Requirements & Acceptance*
A user of the sub-project will be able to access a virtual appliance front-end and the console of the working nodes, to browse data and run the software prepared for the sub-project. This Virtual Appliance should automatically "mount" the data of the sub-project and provide a persistent repository for the results.	*Example architectural diagram*	*Console access to the Virtual appliances and XNAT portals. This will require the proper management of authorization credentials. Automatic file-system-like availability of the data. Availability of a persistent storage for the sub-project results.*

Case 4: A sub-project user can execute a processing pipeline on the platform, defining specific parameters		
Draft Use Case	*Sup. definition*	*Requirements & Acceptance*
The processing of data will involve executing a dataflow involving different requirements and concurrency degrees. This should be transparent to the user, who should be only required to initiate and monitor the workflow, using GUI or CLI tools. Users should be able to define specific configuration options. Executions are not interactive and should be preserved on different sessions.	*Workflow definitions*	*Workflow enactment systems can be provided in the software catalogue, although customisation for efficiently running on the platform may be requested. In the case of using batch queues as back-ends, platform should adapt itself to the actual workload. Persistence will likely depend only on the workflow engine.*

Case 5: A sub-project user can visualize the results obtained		
Draft Use Case	*Sup. definition*	*Requirements & Acceptance*
The users should be able to explore the results produced by the executions without requiring downloading the actual results. A user may decide to keep or remove the results obtained by a specific experiment.	*Sample output products*	*The Virtual Appliance must allow the access to Graphical interfaces. Network configuration must allow such type of traffic and*

		firewall rules.

Case 6: A sub-project user can inject own-specific software to be integrated in the pipeline		
Draft Use Case	*Sup. definition*	*Requirements & Acceptance*
A user may would like to include his/her own software in the processing pipeline. This will require the user to know the specific workflow language configuration and to modify the content of the VM disk.		*First, console access to all the VMs of the VA should be provided. Second, VM disks should be persistent, just in case that the VM is powered off, the changes persist.*

Case 7: A sub-project user can receive consolidated results from the analysis		
Draft Use Case	*Sup. definition*	*Requirements & Acceptance*
Final results of the whole processing need to be preserved for further analysis and research. Those results may be reused and are subject of IPR restrictions. Moreover, those results could be downloaded or stored somewhere else to guarantee the long-term persistence		*Block transferring of data should be facilitated, as well as the possibility of linking to existing data infrastructures.*

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

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

7 INFRASTRUCTURE TECHNICAL REQUIREMENTS

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

7.1 Current e-Infrastructures Resources

Start from the current use of e-infrastructures.

7.1.1 Networking

Describe the current connectivity. Currently, the systems are connected by means of two networks: Arterias, the health network of Valencia Region, and RED-IRIS, the Spanish national research and academic network.

Describe the key requirements (availability, bandwidth, latency, privacy, etc) The main requirement is the bandwidth. Medical Images are of large size (500MB per case)

Specify any current issue (like last mile, or access from commercial, etc) Bridge between health and academic networks requires special configurations due to security reasons.

7.1.2 Computing: Clusters, Grid, Cloud, Supercomputing resources

Describe the current use of each of these type of resources: size and usage. A cluster of 16 nodes.

Indicate if there is any mode of "orchestration" between them Shared disk and LONI pipeline.

7.1.3 Storage

Describe the current resources used 5 TB storage

Discuss the key requirements (I/O performance, capacity, availability, reliability, any other QoS indicator) 500 MB/s disk access on a SAN in a local area network.

7.2 Short-Midterm Plans regarding e-Infrastructure use

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

7.2.1 Networking

Describe the proposed connectivity Connected to GEANT.

Describe new/old key requirements (availability, bandwidth, latency, QoS, private networking, etc) Mainly bandwidth.

Specify any potential solution/technique (for example SDN) No complex restrictions.

7.2.2 Computing: Clusters, Grid, Cloud, Supercomputing resources

Describe the evolution expected: which infrastructures, total "size" and usage. Nearly 3K CPU hours for the proposed sub-project. We expect to host ~30 subprojects per year (around 10 simultaneous projects).

Detail potential "orchestration" solutions Beyond current pipelines (LONI) and web portals (XNAT)

7.2.3 Storage

Describe the resources required. Moderate for the specific subproject (lower than 10 TB). We expect to host ~30 subprojects per year (around 10 simultaneous projects).

Discuss the key requirements (I/O performance, capacity, availability, reliability, any other QoS indicator) Data bandwidth could be the limitation, since single studies involve 500MB and need to be loaded in memory quickly.

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

Sample questions to capture details of a support case

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

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

- *Which NGIs are interested in supporting this case? (Question to the NGIs)*

7.3 On Monitoring (and Accounting)

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

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

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

Availability of services is controlled by using Nagios.

7.4 On AAI

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

Describe the current AAI status of your community/research infrastructure

• *Does your community/research infrastructure already use AAI solutions? Basic user/password. Interest on moving to the support of external identity providers.*

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

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

- *Type of IdP to be integrated (e.g. institutional IdP part of national federations and eduGAIN or non federated, social media credentials, dedicated research community catch-all IdP, ...) No current plans, but willing to integrate European-wide authentication mechanisms.*
- *Preferred authentication technology, and requirements for support of multiple technology and credential translation services (e.g. SAML -> X509 translation) User / password and / or access tokens (sub-accounts).*
- *Community level authorization/attribute based authorization to support different authorization levels for the users. Per Sub-project.*
- *Web access and/or non-web access. Web-based access.*
- *Need for delegation (e.g. execute complex workflows on behalf of the user). Yes, potentially using access tokens.*
- *Support for different level of assurance credentials, and need to use the information about users with lower level of assurance credentials to limit their capability. Potential use of access tokens or sub-accounts.*
- *Requirements for high level of assurance credentials (e.g. to access confidential/sensitive data) Data is anonymized. Access to sub-project data should be granted only to people in the sub-project.*

7.5 On HPC

Describe any specific issue related to the use of supercomputers.

GPU or MPI-based distributed computing may be required for extended versions of the use case or similar sub-projects.

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

Software used	***Software/applications/services required, configuration, dependencies (Describe the software/applications/services name, version, configuration, and dependencies needed to run the application, indicating origin and requirements.)*** - *DICOM data processing: dcm2nii (MRIcron package), dcmtk, DICOM Cleaner.* - *Deep learning: caffe.* - *Bone quality: MATLAB runtime.* - *Data organization and visualization: XNAT, dcm4chee.* - *Database presentation: JasperReports.*
Operating system requirements	*Linux.*
Run libraries requirements	***Run API/libraries requirements (e.g., Java, C++, Python, etc.)*** *Python, Java, C++ and Matlab Runtime*
CPU requirements (multithread,MPI, "wholenode")	*Multithread, also benefiting GPUs. MPI for future cases*
Memory requirements	*<16GB RAM per node*
Network requirements	*Fast communication between storage and computing nodes. No need for low-latency networks.*
Disk space requirements (permanent, temporal)	***Include the requirements for data transferring (upload and download of data objects: files, directories, metadata, VM/container images, etc.)*** *permanent storage on the order of 5 TB per sub-project, temporal storage may double it.*
External data access requirements	*No external access, once the data is selected.*

Typical processing time	*3h per study (3K hours for the sub-project).*
Other requirements	***Requirements for data synchronization*** ***Requirements for data publication*** ***Requirements for depositing data to archives and referring them*** ***Requirements for mobile application components for data storage and access*** ***Requirements for data encryption and integrity control-related functionality*** *It will make profit of GPUs.*
Other comments	**<input here>**
Relevant references or URLs	**<input here>**

8 CONNECTION WITH INDIGO SOLUTIONS

<To be filled by INDIGO JRA >

8.1 IaaS / WP4

8.2 PaaS / WP5

8.3 SaaS / WP6

8.4 Other connections

9 FORMAL LIST OF REQUIREMENTS

<this will be further edited within WP2>

10 REFERENCES

R 1	
R 2	
R 3	
R 4	
R 5	