INDIGO-DataCloud

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

INPUT TO EU DELIVERABLE: D 2.1

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

Abstract

This report summarizes the findings of T2.1 and T2.2 **for partner P6 (CIRMMP)** 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	Antonio Rosato	**P6**/WP2	
Reviewed by	**Moderators:** P.Solagna, F.Aguilar, J.Marco **Internal Reviewers:** <<To be completed by project office on submission to PMB>>		
Approved by	**PMB** <<To be completed by project office (no submission)>>		

III. DOCUMENT LOG

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

TABLE OF CONTENTS

0 INTRODUCTION AND CONVENTIONS

PLEASE, READ CAREFULLY BEFORE COMPLETING THE ANNEX:

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

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

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

Some relevant definitions:

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

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

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

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

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

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

RESEARCH MANAGERS

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

RESEARCHERS

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

APPLICATION DEVELOPERS AND INTEGRATORS WITHIN THE RESEARCH COMMUNITIES

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

MIDDLEWARE DEVELOPERS AND E-INFRASTRUCTURE MANAGERS

-Sections 7, 8: should be discussed with them

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

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

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

Other conventions and instructions for this document:

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

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

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

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

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

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

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

1 EXECUTIVE SUMMARY ON THE CASE STUDY

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

1.1 Identification

- *Community Name*: **Structural Biology (INSTRUCT)**
- *Institution/partner representing the community in INDIGO:* **CIRMMP; U. Utrecht**
- *Main contact person*: **Antonio Rosato**
- *Contact email:* **rosato@cerm.unifi.it**
- *Specific Title for the Case Study:* **Molecular dynamics simulations**

1.2 Brief description of the Case Study and associated research challenge

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

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

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, which can be summarized as follows:

1. The setup of simulations needs to be adjusted to the specific system under investigation. Adjustments can be done automatically but the corresponding tools are not widespread or easy to use;
2. The computational power to run the long dynamics that are currently state-of-the-art in the field is not always available especially to researchers in small laboratories; multi-core processors and/or GPGPU-based systems are required, as well as a suitable implementation of MPI libraries
3. The simulations produce large datasets that must be stored for subsequent analysis using dedicated tools. Online analysis of so-called molecular dynamics trajectories is hindered by the large size of the data. Again, a researcher in the field presently needs to locally install and maintain his/her own version of the software, and will experience a high barrier to adopt more recent tools.

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.

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 objective is the implementation of interfaces that allow users to run multi-threading as well as MPI-based molecular dynamics simulations depending on specific needs, provide access to the simulated data (trajectories) and perform standardized analyses of the trajectories.

Cloud technologies would be instrumental both to obtain access to different types of computational resources and to allow the post-simulation analysis of trajectories without the need to move the data from the researchers' lab.

1.4 Expected results and derived impact

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

On-line processing of molecular dynamics data

Development of improved methods for the setup of molecular dynamics simulations

Development of improved methods for the comparison between experimental data and the output of molecular dynamics simulations

Improved characterization of the effect of macromolecular dynamics on key experimental parameters

1.5 References useful to understand the Case Study

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

<input here>

2 INTRODUCTION TO THE RESEARCH CASE STUDY

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

2.1 Presentation of the Case Study

Describe the Case Study from the research point of view

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.

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 involves researchers from academia as well as, potentially, pharma and biotech companies. The researchers will include:

- Biologists, who want to validate hypotheses on functional mechanisms relying on macromolecular dynamics;
- Pharmacologists, especially for ligand docking applications;
- Experts in MD simulations, typically with a background in physics or physical chemistry and an interest in the development of MD methods and their validation through experimental data

The first two profiles will often have only a coarse understanding of the complexity of MD simulations and little to none skills in cloud/distributed computing.

2.3 Current Status and Plan for this Case Study

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

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

Simple portals for the setup and execution of MD simulations on the EGI computational infrastructure have been developed within the WeNMR project (www.wenmr.eu). The applications supported are free dynamics and MD-based refinement of protein structures. Data management is not implemented: users only have the option to download their data. Very limited capabilities for analysis of trajectories are provided.

A development which is being already pursued is the implementation of portals executing simulations on GPGPU computational resources, which can significantly speed up calculations and thus allow users to perform longer simulations, up to more biologically relevant time scales.

In the short/medium term, it is expected that these portals will be endowed with a larger set of pre-defined applications and also will be able to cope with different architectures of the computing elements.

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

- 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.

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 currently available portals are supported by the WeNMR e-infrastructure. They exploit the computational resources provided by EGI, using the EMI middleware.

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

This Case Study is relevant to the INSTRUCT ESFRI Infrastructure, as macromolecular dynamics is a crucial aspect of the mechanism of function of biological molecules. These activities will be supported also by the West-Life Virtual Research Environment, which will provide an electronic infrastructure operating in close contact with INSTRUCT to support the integrated usage of multiple experimental techniques and of data from multiple experimental facilities to obtain information on the large macromolecular machineries of cells.

The portals will also be used within the iNEXT (infrastructure for NMR, EM and X-rays for Translational research) project. iNEXT will provide high-end structural biology instrumentation and expertise, facilitating expert and non-expert European users to translate their fundamental research into biomedical and biotechnological applications. It consists of a consortium of 23 partners and brings together leading European structural biology facilities under one interdisciplinary organizational umbrella. It includes synchrotron sites for X-rays, NMR centers with ultra-high field instruments, and, for the first time, advanced electron microscopy and light imaging facilities.

Ligand docking applications are of interest also to the EuOpenScreen ESFRI Infrastructure, because of their potential use in biomedicine.

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: **3** • *Number of production sites supporting the software* **5**
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* **1** • *Number of popular applications used by the user communities directly integrated with the project products:* **6** • *Number of research communities using the developed Science Gateway and Mobile Apps:* **2** • *Research Communities external to INDIGO using the software products*: **2**
Impact on Policy	*Policy impact depends on the successful generation and dissemination of relevant knowledge that can be used for policy formulation at the EU, or national level.*	• *Number of contributions to roadmaps, discussion papers:* **1**

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:* **0** • *List of potential events/conferences/workshops:* **10-15 per year (average)** • *Number of domain exhibitions attended* **10** • *Number of communities and stakeholders contacted* **Does community mean "Research center"? 10**
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** • *Number of conference papers and presentations*: **20**

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.

The following steps define the operation of the case study:

1. Registration of users, via a central web portal
2. Definition of a project. The project will comprise of input structural data, specific simulation parameters if needed (e.g. force field parameters for non-standard organic molecules), selection of the desired application. Depending on the application selected and the size of the input system, appropriate computational requirements will be established
3. Deployment of the user interface and of the computing nodes. Deployment should be platform-agnostic
4. Automated setup of simulation and validation of input data
5. Execution of MD simulation, which produces a so-called trajectory i.e. atomic coordinates at different time points of the simulation
6. Definition of post-simulation analysis of the trajectory. A second interface will provide the users with options for trajectory analysis, depending on the application selected in 2.
7. Deployment of the computing nodes and execution of the trajectory analysis. This step will typically not be very computational intensive
8. Visualization of results

Administrative actions

1. Access control and accounting
2. Infrastructure monitoring
3. Update of protocols; implementation of new applications

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.)

We distinguish essentially two broad user categories:

- **Platform users**, who however can have different background in computational sciences and the theoretical foundations of MD simulations. In fact, we can distinguish two subgroups, consisting respectively of researchers who exploit predefined protocols without modification and more advanced users who can cooperate with the developers to provide improved or tailored versions of the protocols;
- **Platform developers**, who deploy optimized protocols on the infrastructure for different applications, also upon interaction with the users

A further related role is that of the administrators who monitor usage and take care of accounting.

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 database used is the Protein Data Bank, which stores all known 3D structures of biological macromolecules and adducts thereof. Input file are usually text files in the so-called PDB format. Additional input maybe force field parameters for non-standard molecules, which are also in text format.

Experimental data can provided to MD simulations using a variety of text formats.

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

The initial use cases will focus on the provision of an interface and computational resources to perform two types of MD simulations:

- MPI-based
- Multi-threading, using multi-core CPUs

Storage needs will depend on system size and various simulation parameters, such as total duration and saving frequency. The typical sizes of the output trajectories would be several tens of gigabytes per day.

3.5 Description of the Case Study in terms of Workflows

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

Input data (plain text, in different formats) are provided by the users to the system via the dedicated interface. These can include experimental data, to be used to steer the simulation or simply for comparison purposes (e.g. agreement with experiment as a function of time/motion).

Simulations produce trajectories (i.e. atomic coordinates at different time points of the simulation) in binary format. These data are stored for post-simulation analysis.

Analysis tools access to the trajectories and can produce a significant variety of data, ranging from tabular to visual data, depending on users demands. Such data are typically used for comparison with experimental data, to direct new experimental work, or in publications (or a combination of these).

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.

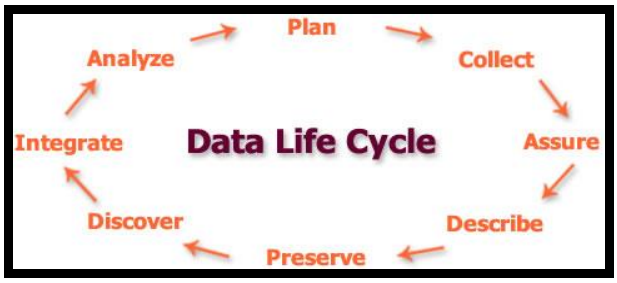
Current implementations have been demonstrated to run effectively both on multi-core CPU units and on HPC resources. Currently available web portals allow users to access to grid (EGI) resources.

More advanced applications will need the use of MPI.

4 DATA LIFE CYCLE

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

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



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

4.1 Data Management Plan (DMP) for this Case Study

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

According to the common practice in Structural Biology, as implemented also in the data policy of the INSTRUCT infrastructure, it is the users' responsibility to archive, maintain and possibly release in the public domain all data associated with his/her research. No specific repositories exist for the dynamics data addressed in this Use case. The experimental data and any condensed information extracted from the simulated data are typically included in the supporting information of publications. The simulated data themselves are typically unregistered and remain available only within the user's laboratory.

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

4.1.1 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>**

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

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

For each data set provide:

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

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

Data set description **The data generated by the simulations consist of atomic Cartesian coordinates at fixed timepoints along the simulation. They are stored as binary files. Many different data types, from tabular to visual, can be derived from the original simulated data, depending on the specific study.**

Standards and metadata **There are no community metadata standards, although some ML-based systems to store metadata have been proposed (e.g. MSML). The simulated data themselves can be stored in a restricted number of different formats, depending on the software used.**

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

Connection to Instrumentation,

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

<input here>

Vocabularies and Ontologies

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

<input here>

Data Capture Methods

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

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

Metadata

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

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

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

Data sharing

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

<input here>

Method for Data Sharing

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

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

Restrictions on Sharing

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

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

• *What action will you take to overcome or minimise restrictions?* **<input here>**

Data Repository

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

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

Archiving and preservation (including storage and backup)

Questions to consider before answering:

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

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

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

• What additional resources are needed to deliver your plan?

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

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

• Will charges be applied by data repositories?

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

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

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

<input here>

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

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

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

Questions to consider:

• How will potential users find out about your data?

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

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

Are the data and associated software produced and/or used in the project discoverable (and readily located), identifiable by means of a standard identification mechanism (e.g. ***Digital Object Identifier****)?* **The data addressed in this use case are typically used to generate condensed information that may eventually be compared with experimental data to validate the simulation. Such information, along with the relevant experimental data if available, are typically included in publications' supporting materials.**

SCIENTIFIC RESEARCH DATA SHOULD BE **ACCESIBLE**

Questions to consider:

• Who owns the data?

• How will the data be licensed for reuse?

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

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

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

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

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

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

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

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

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

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

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

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

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

Guidance on Metadata:

• How will you capture / create the metadata?

• Can any of this information be created automatically?

• What metadata standards will you use and why?

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

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

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

INTEROPERABLE** TO SPECIFIC QUALITY **STANDARDS

• What format will your data be in?

• Why have you chosen to use particular formats?

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

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

See the UKDS Guidance on recommended formats

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

<input here>

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

4.2.1 General description of data levels

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

4.2.2 Collection/Acquisition

Gathering RAW data

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

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

From RAW Data to Calibrated Data

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

4.2.3 Access to external data

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

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

4.2.4 Data curation

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

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

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

4.2.5 Data ingestion / integration

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

4.2.6 Further data processing

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

4.3 Analysis

4.3.1 Basic analysis and standard analysis suites

Describe usual examples of basic analysis in the Case Study **Derivation of predictions for experimental observables such as NMR chemical shifts, bond order parameters, various structural properties**

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

4.3.2 Data analytics and Big Data

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

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

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

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

4.3.3 Data visualization and interactive analysis

Indicate the need for data and analysis results visualization **Data visualization is common practice**

Indicate how visualization is made and if interactivity/steering is needed **VMD (http://www.ks.uiuc.edu/Research/vmd/) is one of the post popular programs for the task. It entails interactive tasks.**

Specify the User Interfaces (web, desktop, mobile, etc.) **Web; desktop**

4.4 Data Publication

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

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

5 SIMULATION/MODELLING

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

5.1 General description of simulation/modelling needs

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

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

5.2 Technical description of simulation/modelling software

For each simulation package:

Identify the simulation software **Amber14**

Provide a link to its documentation, and describe its maturity and support level **http://ambermd.org/; highly mature, well supported**

Indicate the requirements of the simulation software (hardware: RAM, processor/cores, extended instruction set, additional software and libraries, etc.) **libraries: CUDA 6.5, openMPI Hardware: >5Gb RAM; multi CPU and/or multi GPU**

Tag the simulation software as HTC or HPC **HTC for molecular refinement; HPC for free simulations**

List the input files required for execution and how to access them **Text files that contain the MD directive of the simulation, cartesian coordinates of protein atoms, and amino acid topology information, force field parameters. Coordinates can be retrieved from the PDB database http://www.rcsb.org/; the other input files can be generated using the package amber14 tools**

Describe the output files and how they will be stored **The output file contains all snapshot of the system coordinates saved during time simulation, the output can be saved as text or netCDF binary format**

Reference an existing installation and performance indicators **http://ambermd.org/gpus/**

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

Specify if the simulation software can exploit GPUs **YES**

Specify how the simulation software exploits multicore systems **MPI can be used**

Specify if parametric runs are required **NO**

Estimate the use required of the resources (million-hours, # cores in parallel, job duration, etc) **Job duration depends critically on number of atoms(system size), can take from a few days up to a few weeks**

5.3 Simulation Workflows

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

6 DETAILED USE CASES FOR RELEVANT USER STORIES

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

6.1 Identification of relevant User Stories

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

Implementation

Portal developers will access to Virtual Machines with Scientific Linux, MPI libraries, Python, Http support, a queuing system such as torque. It will be necessary to allow the definition of different hardware configurations, such as systems with a single multi-core worker node or including multiple nodes (Virtual cluster).

The developers will deploy on the virtual machine or master node of the virtual cluster the specific software tools for running the simulations (e.g. the AMBER package), the appropriate html/python interface(s) for the setup of the simulation, and post-processing software, such as VMD, as requested. The interface will embed a simulation protocol that minimizes the number of parameters that the final user must define, in order to facilitate usage of the tool. Different protocols will be made available as separate VMs (e.g. structure refinement; free MD simulation in water).

The portal developers will define users on the system as needed.

The configured systems will be stored on a central repository. The configuration should also define to some extent the hardware required/suggested.

Execution

A portal operator wants to execute a simulation.

- He/she will access a specific interface to configure the hardware of the virtual machine or cluster hosting the simulation. The system automatically takes care of updating the related settings (e.g. nodes defined in the queuing system, host list, …).
- The operator will select available resources (or some brokering system automatically assigns available resources (should be limited to one site for efficient communication between nodes)).
- The requested virtualized system is launched
- Each protocol is tested
- Credentials of authorized end users are updated using some AAI system or the SSO module of WeNMR.
- The portal is put into production mode and made accessible to the users.

Operation

MD simulations can take from 24 hours to several weeks. Suitable points to interrupt a simulation and restart it later can be defined. It is important that system failures do not compromise the status of the simulation being executed, in order to avoid data loss. The protocols will in any case include periodic saving of simulation status to allow the user to restart from the last saved point in case of failure.

Post-simulation analysis require only few hours typically, but may entail interactive sessions. A mechanism to maintain stored data available to users also after the end of the simulation and shut-down of the virtual system is crucially needed.

It is entirely possible that the data are analysed using a different hardware than that used to execute the simulation. VMD analyses can be performed on essentially any type of platform.

Permanent data storage is not needed, as the user will eventually want to remove his/her data.

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

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

7 INFRASTRUCTURE TECHNICAL REQUIREMENTS

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

7.1 Current e-Infrastructures Resources

Start from the current use of e-infrastructures.

7.1.1 Networking

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

7.1.2 Computing: Clusters, Grid, Cloud, Supercomputing resources

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

7.1.3 Storage

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

7.2 Short-Midterm Plans regarding e-Infrastructure use

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

7.2.1 Networking

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

7.2.2 Computing: Clusters, Grid, Cloud, Supercomputing resources

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

7.2.3 Storage

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

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

Sample questions to capture details of a support case

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

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

7.3 On Monitoring (and Accounting)

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

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

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

<input here>

7.4 On AAI

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

Describe the current AAI status of your community/research infrastructure

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

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

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

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

7.5 On HPC

Describe any specific issue related to the use of supercomputers.

<input here>

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

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

8 CONNECTION WITH INDIGO SOLUTIONS

<To be filled by INDIGO JRA >

8.1 IaaS / WP4

8.2 PaaS / WP5

8.3 SaaS / WP6

8.4 Other connections

9 FORMAL LIST OF REQUIREMENTS

<this will be further edited within WP2>

10 REFERENCES

R 1	
R 2	
R 3	
R 4	
R 5	