

# INDIGO-DataCloud

## INITIAL REQUIREMENTS FROM RESEARCH COMMUNITIES ANNEX 1.**P10**: SELECTED CASE STUDY FROM **HADDOCKPORTAL**

### INPUT TO EU DELIVERABLE: D 2.1

---

Document identifier:	INDIGO-WP2-D2.1-ANNEX-1P0-V7
Date:	<b>04/06/2015</b>
Activity:	<b>WP2</b>
Lead Partner:	<b>EGL.eu</b>
Document Status:	<b>DRAFT</b>
Dissemination Level:	<b>CONFIDENTIAL (INTERNAL)</b>
Document Link:	

---

#### Abstract

This report summarizes the findings of T2.1 and T2.2 **for partner P10** 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

Copyright © Members of the INDIGO-DataCloud Collaboration, 2015-2018.

## II. DELIVERY SLIP

	Name	Partner/Activity	Date
<b>From</b>	Alexandre Bonvin	P10/WP2	
<b>Reviewed by</b>	<b>Moderators:</b> P.Solagna, F.Aguilar, J.Marco <b>Internal Reviewers:</b> <<To be completed by project office on submission to PMB>>		
<b>Approved by</b>	<b>PMB</b> <<To be completed by project office (no submission)>>		

## III. DOCUMENT LOG

Issue	Date	Comment	Author/Partner
1	3-June-2015	First draft, v0	A. Bonvin, U. Utrecht
2	4-June-2015	v1: Added use case scenario in section 6	A. Bonvin, U. Utrecht
2	25-June-2015	v3: Further edits and update of some figures, completed sections 5, 7.2	A. Bonvin, U. Utrecht

## TABLE OF CONTENTS

<b>0</b>	<b>INTRODUCTION AND CONVENTIONS</b>	<b>5</b>
<b>1</b>	<b>EXECUTIVE SUMMARY ON THE CASE STUDY</b>	<b>7</b>
1.1	Identification	7
1.2	Brief description of the Case Study and associated research challenge	7
1.3	Expectations in the framework of the INDIGO-DataCloud project	7
1.4	Expected results and derived impact	7
1.5	References useful to understand the Case Study	7
<b>2</b>	<b>INTRODUCTION TO THE RESEARCH CASE STUDY</b>	<b>8</b>
2.1	Presentation of the Case Study	8
2.2	Description of the research community including the different roles	9
2.3	Current Status and Plan for this Case Study	11
2.4	Identification of the KEY Scientific and Technological (S/T) requirements	11
2.5	General description of e-Infrastructure use	11
2.6	Description of stakeholders and potential exploitation	12
<b>3</b>	<b>TECHNICAL DESCRIPTION OF THE CASE STUDY</b>	<b>14</b>
3.1	Case Study general description assembled from User Stories	14
3.2	User categories and roles	15
3.3	General description of datasets/information used	16
3.4	Identification of the different Use Cases and related Services	16
3.5	Description of the Case Study in terms of Workflows	16
3.6	Deployment scenario and relevance of Network/Storage/HTC/HPC	17
<b>4</b>	<b>DATA LIFE CYCLE</b>	<b>18</b>
4.1	Data Management Plan (DMP) for this Case Study	18
<b>5</b>	<b>SIMULATION/MODELLING</b>	<b>19</b>
5.1	General description of simulation/modelling needs	19
5.2	Technical description of simulation/modelling software	19
5.3	Simulation Workflows	21
<b>6</b>	<b>DETAILED USE CASES FOR RELEVANT USER STORIES</b>	<b>22</b>
<b>7</b>	<b>INFRASTRUCTURE TECHNICAL REQUIREMENTS</b>	<b>24</b>
7.1	Current e-Infrastructures Resources	24
7.1.1	Networking	24
7.1.2	Computing: Clusters, Grid, Cloud, Supercomputing resources	24
7.1.3	Storage	24
7.2	Short-Midterm Plans regarding e-Infrastructure use	25
7.2.1	Networking	27
7.2.2	Computing: Clusters, Grid, Cloud, Supercomputing resources	27
7.2.3	Storage	27
7.2.4	<i>SPECIFIC QUESTIONS REGARDING USE OF EGI.eu (FROM EGI DOC 2478)</i>	28
	<i>Sample questions to capture details of a support case</i>	28
7.3	On Monitoring (and Accounting)	29

7.4	On AAI .....	29
7.5	On HPC .....	31
7.6	Initial short/summary list for “test” applications (task 2.3) .....	31
<b>8</b>	<b>CONNECTION WITH INDIGO SOLUTIONS.....</b>	<b>32</b>
8.1	IaaS / WP4 .....	32
8.2	PaaS / WP5 .....	32
8.3	SaaS / WP6 .....	32
8.4	Other connections .....	32
<b>9</b>	<b>FORMAL LIST OF REQUIREMENTS.....</b>	<b>33</b>
<b>10</b>	<b>REFERENCES.....</b>	<b>34</b>

## 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<sup>1</sup>.*

*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<sup>2</sup> 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*

---

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

<sup>2</sup> <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:
- Institution/partner representing the community in INDIGO:
- Main contact person:
- Contact email:
- Specific Title for the Case Study:

## 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*

## 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?*

## 1.4 Expected results and derived impact

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

## 1.5 References useful to understand the Case Study

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

## 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.*

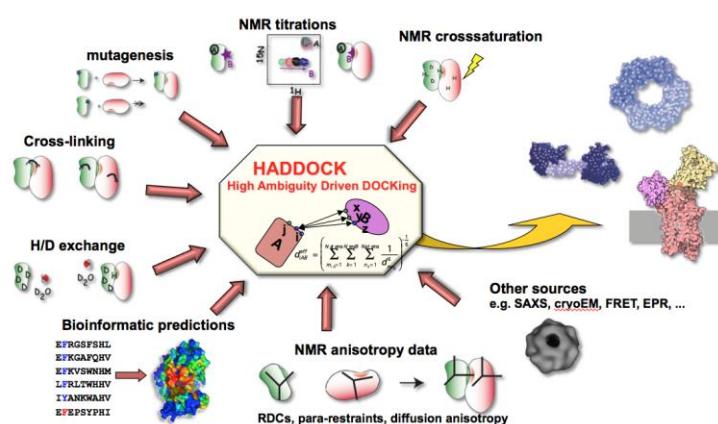
### 2.1 Presentation of the Case Study

*Describe the Case Study from the research point of view*

#### Integrative modelling of macro-molecular assemblies with HADDOCK

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

The Utrecht partner (P10) has developed for over ten years now the integrative, information-driven docking approach, HADDOCK (<http://www.bonvinlab.org/software/haddock2.2/haddock.html>). It supports the incorporation of a large variety of data from NMR and other biophysical methods (Figure 1). HADDOCK has demonstrated a strong performance in the blind docking experiment CAPRI.



**Figure 1:** Illustration of data types supported by HADDOCK for integrative modelling of biomolecular complexes.

The software is made available through a user-friendly web interface, offered both on local resources and within the context of the WeNMR Virtual Research Community (<http://www.wenmr.eu> ; <http://haddock.science.uu.nl/enmr/services/HADDOCK2.2> ). It has attracted a large user community worldwide (>5500 users; see <https://www.targetmap.com/viewer.aspx?reportId=39366> ), submitting a sustained number of computations to HTC infrastructures like the EGI (>6M jobs per year) and resulted in over 120 deposited structures of complexes in the PDB.



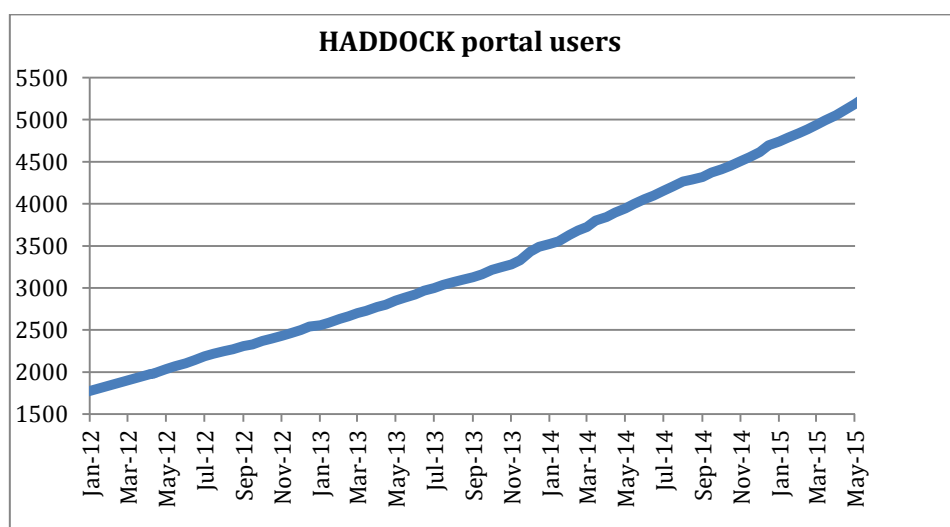
## 2.2 Description of the research community including the different roles

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

*Describe the research community specifically involved in this Case Study*

P10, U. Utrecht, is the main developer of HADDOCK. The Bonvin group is developing and distributing the software (<http://bonvinlab.org/software>), and also operating the web portals (both local versions running on the group computing resources and the grid-enabled version offered via the WeNMR VRC).

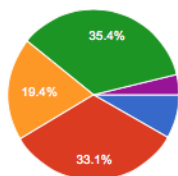
End users of HADDOCK consist of a large scientific community worldwide with different backgrounds and expertise, ranging from bachelor students to experienced researchers and even commercial companies. They mostly interact with the web portal front end and/or run a local version of the software. The user community shows a sustained growth rate (Figure 2).



**Figure 2:** HADDOCK portal user growth.

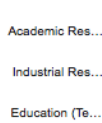
A recent survey of the community and application areas of HADDOCK can be found [here](#). A few highlights are provided in the following based on 317 respondents at the date of writing this document.

### What is your current position?



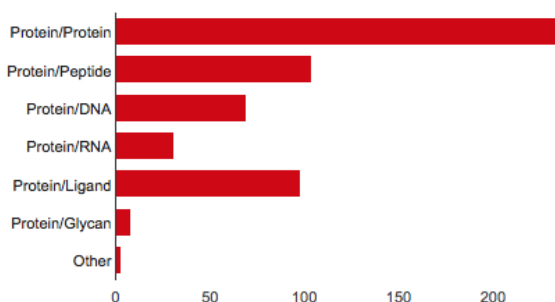
Student (Bachelor/Master)	26	8.3%
Ph. D Student	104	33.1%
Postdoctoral Researcher	61	19.4%
Researcher / Staff Member	111	35.4%
Other	12	3.8%

### In which category does your usage of HADDOCK fit best?



Academic Research (non-profit)	308	97.2%
Industrial Research (for-profit)	8	2.5%
Education (Teaching)	23	7.3%

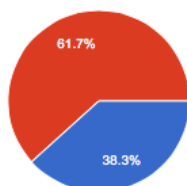
### What kind of systems do you study with HADDOCK?



Protein/Protein	235	74.6%
Protein/Peptide	104	33%
Protein/DNA	69	21.9%
Protein/RNA	31	9.8%
Protein/Ligand	98	31.1%
Protein/Glycan	8	2.5%
Other	3	1%

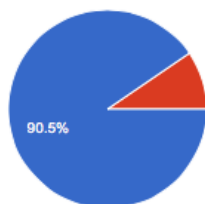
## Software & Usage

### Do you use a local HADDOCK installation?



Yes	121	38.3%
No	195	61.7%

### Do you use the HADDOCK web server?



Yes	285	90.5%
No	30	9.5%

## 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*

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

## 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 science requirements should be left to the software developers of HADDOCK, i.e. the Utrecht group and are an active part of the research of the group.

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

## 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.*

We currently distinguish three versions of our HADDOCK portal:

- 1) A local version running on local resources (e.g. linux cluster with masternode and >600 CPU cores based on 48cores nodes)
- 2) A grid-enabled version making use direct submission to EGI and OSG resources (including the federated desktop grid resources) via glite commands
- 3) A grid-enabled version making use of DIRAC4EGI to submit jobs to EGI resources.

The volume of job submission for versions 2) and 3) is in the order of 8 million jobs per year (Figure 3) (one user run generating typically several hundred jobs) with the largest volume currently being handled by DIRAC4EGI (Figure 4).

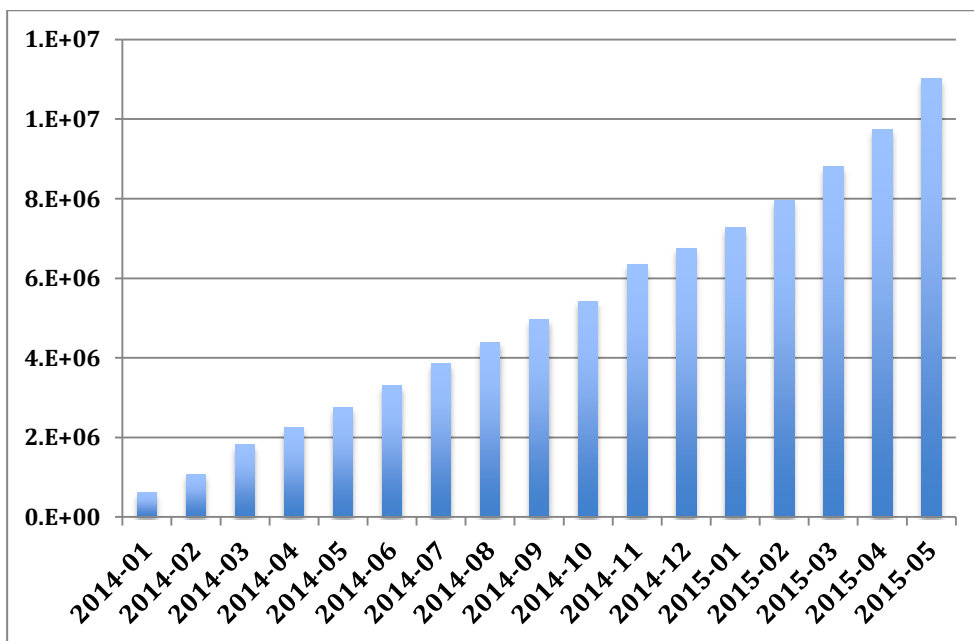


Figure 3: HADDOCK cumulative job statistics

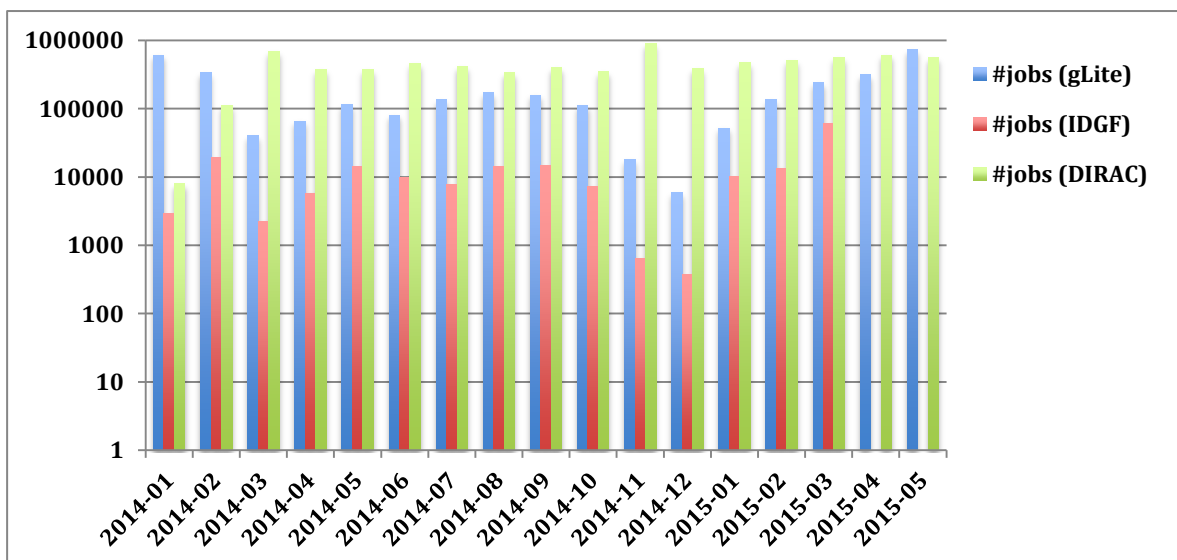


Figure 4: HADDOCK monthly job distribution for direct glite submission, DIRAC4EGI and desktop grid.

## 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*

The stakeholders for this Case Study are:

- 1) The HADDOCK software developers - easier portal management, disconnection from the physical infrastructure – fall back solutions
- 2) The large HADDOCK user community (see section 2.3) – for them transparent use of the portals – details of the implementation are irrelevant
- 3) Pharmaceutical industry and small biotech companies that might run a virtualized server with computing resources on their internal cloud

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

<b>Area</b>	<b>Impact Description</b>	<b>KPI Values</b>
<b>Access</b>	<i>Increased access and usage of e-Infrastructures by scientific communities, simplifying the “embracing” of e-Science.</i>	<ul style="list-style-type: none"> <li>• <i>Number of users of the HADDOCK web portals</i></li> <li>• <i>Number of runs handled by the server</i></li> </ul>
<b>Usability</b>	<i>Simplifying deployment of the web portals in cloud resources</i>	<ul style="list-style-type: none"> <li>• <i>Number of cloud instances of the portal initiated</i></li> </ul>
<b>Impact on Policy</b>	<i>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.</i>	<ul style="list-style-type: none"> <li>• <i>N/A</i></li> </ul>
<b>Visibility</b>	<i>Visibility of the project among scientists, technology providers and resource managers at high level.</i>	<ul style="list-style-type: none"> <li>• <i>Number of citations of the HADDOCK software (sustained increased of ~ 50 citations a year)</i></li> <li>• <i>Advertisement at events/conferences/workshops: at least 10 conferences/workshops a year</i></li> </ul>
<b>Knowledge Impact</b>	<i>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.</i>	<ul style="list-style-type: none"> <li>• <i>Number of journal publications acknowledging the project: 5</i></li> </ul>

*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.)*

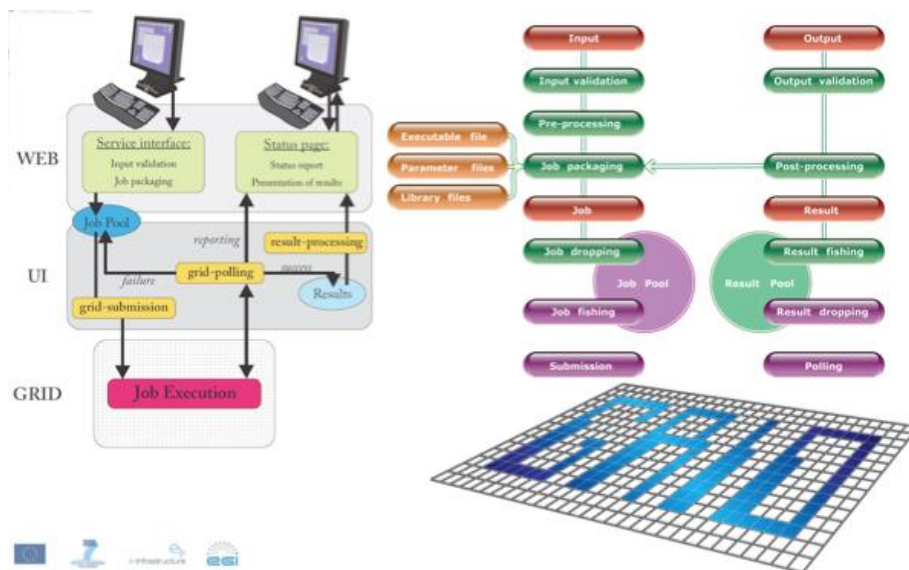
*Assemble it using preferably an AGILE scheme based on User Stories.*

#### 3.1 Case Study general description assembled from User Stories

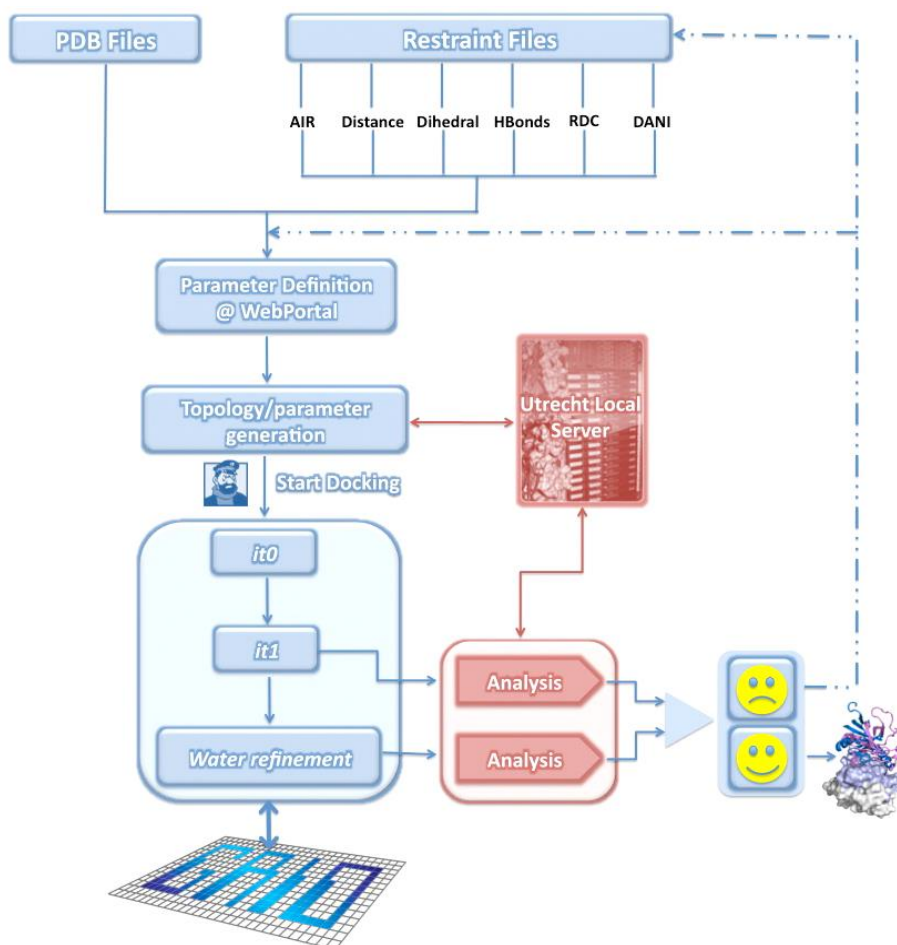
*Please describe here globally the Case Study. If possible use as input “generic” User Stories built according to the scheme: short-description (that fits in a “card”) + longer description (after “conversation” with the research community). Provide links to presentations in different workshops describing 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 HADDOCK portal effectively implements a complex workflow in which user data are first validated and processed before HADDOCK computations are launched. Each HADDOCK run correspond to a complex workflow, orchestrated by a master python script that manages the workflow, generates jobs for submission to local queues (e.g. via torque) or grid resources, and monitors the results. A generic portal workflow is summarised in Figure 5 and the specifics of the HADDOCK portal in Figure 6.



**Figure 5:** Generic workflow of structural biology portals within the WeNMR VRC.



**Figure 6:** Workflow of the HADDOCK portal.

### 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 here mainly two different user categories (for details refer to section 2.2):

- 1) The HADDOCK software developers and operators of the associated web portals. These should basically be able to access the cloud machines (clusters) at the root level, to manage and operate the portals
- 2) The HADDOCK user community, which is mainly interaction at the web interface level with the portals and are mainly interested in getting their results back in a reasonable time. Several levels of access are provided depending on the expertise and needs of the users. Current users are either registered directly with the HADDOCK portal, obtaining their credentials in the form of username and password for job submission, or can make use of their WeNMR VRC credentials for submitting jobs to the web portals through the single-sign-on mechanism for

external services developed under the WeNMR project (see for details: <http://www.wenmr.eu/wenmr/wenmr-sso-module>).

And additional category might be system administrators are commercial companies that might install a local version of the virtual HADDOCK portal in the future (subject to special licensing conditions). Operation of the portal is however a rather complex undertaking that typically also required scientific expertise.

### **3.3 General description of datasets/information used**

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

HADDOCK supports a variety of input data, with the main and required type of input being 3D coordinates of molecules. These are plain text files in PDB (Protein Data Bank) format. The PDB is the main database repository for 3D structures of biomolecules (see <http://www.rcsb.org> and <http://www.pdbe.org>). All other input data for HADDOCK are simple text files. The format is described in a Nature Protocol 2010 publication.

S.J. de Vries, M. van Dijk and A.M.J.J. Bonvin The HADDOCK web server for data-driven biomolecular docking. *Nature Protocols*, 5, 883-897 (2010). Download the final author version here.

The online HADDOCK manual provides further details.

See: <http://www.bonvinlab.org/software/haddock2.2/manual.html>

### **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 HADDOCK use case is about PAAS, namely providing a virtualized HADDOCK web portal with all its required computational power, with as minimal requirements:

- Master node with 8 cores, 32 GB memory minimum, ideally 1TB disk space, http daemons to serve the portal, queuing system in place (e.g. torque)
- X Compute nodes with 0.5 GB memory per core and 250 GB disk space (25GB tmp space), for a total of at least 100 CPU cores
- NFS mount of the home partition on all nodes
- Connectivity between nodes and master of at least 1GB via switch
- All components running Scientific Linux, with Python version 2.7 or higher (within the 2.X range)

### **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.*

HADDOCK is already in itself a workflow – no external workflow solutions required in first instance, although in the long term we might consider workflows to manage a large volume of submission to the web portal, irrespective of where the portal will run (local, grid, cloud...). Within a related Center of



Excellence project we will possibly build workflow to connect HADDOCK to other resource, like Gromacs for molecular dynamics simulations. This will be at a higher level, building links between portals.

### **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.*

HADDOCK has already been running on HPC, local clusters and grid resources. The computations are embarrassingly parallel (no MPI). The most relevant scenario for INDIGO-DataCloud is the use case described under 3.4, i.e. deployment of virtual cluster on which the portal will operate.

## 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.*

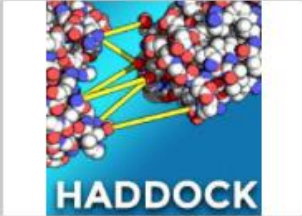
### 4.1 Data Management Plan (DMP) for this Case Study

*According to EU H2020 indications<sup>3</sup>, following UK DCC tool indications*

Data for use in HADDOCK are various and belong to the users (policy of the ESFRI INSTRUCT project), so are the data returned by the web portals. As such no specific DMP is required. Models obtained with HADDOCK might be deposited on open repositories or provided as supplementary material associated with publications. A recent example of HADDOCK data made public by the Utrecht group can be found here:

<http://data.sbggrid.org/dataset/131>

### HADDOCK Docking Models



Data DOI: [10.15785/SBGRID/131](https://doi.org/10.15785/SBGRID/131) | ID: 131  
Bonvin Laboratory, Utrecht University  
Release Date: May 26, 2015

Other data (3D structures) are typically deposited into public databases by end users, e.g. the Protein Data Bank:

- <http://www.rcsb.org>
- <http://www.pdbe.org>

---

<sup>3</sup> *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).*

## 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) N.A.*

*Indicate the type and quantity of simulations needed in the Case Study, and how they are incorporated in the general workflow of the solution N.A.*

### 5.2 Technical description of simulation/modelling software

*Identify the simulation software:*

**HADDOCK** with all its associated software (managed by the HADDOCK software developers – i.e. no need for external support since required the expertise of the developers).

*Provide a link to its documentation, and describe its maturity and support level*

<http://www.bonvinlab.org/software/haddock2.2/haddock.html>

Over 10 years of development, >1000 local installations, >5500 web portal users

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

See Section 3.4

*Tag the simulation software as HTC or HPC*

HTC primarily – embarrassingly parallel

Has run on HPC resources under various HPC-Europe(2) projects.

*List the input files required for execution and how to access them<input here>*

Over 500 parameters can be accessed at the guru level of the web portal. And a variety of input data can be provided.

Refer to the **HADDOCK online manual** at:

<http://www.bonvinlab.org/software/haddock2.2/manual.html>

*Describe the output files and how they will be stored*

Results are returned to the user via a web page which provide links to 3D models in PDB format and a zipped tar archive of the complete run. Those are typically only stored on the server for a maximum of 2 weeks after which they are automatically deleted. For an example output see:

<http://haddock.science.uu.nl/enmr/services/HADDOCK2.2/Files/E2A-HPr-demo/index.html>

*Reference an existing installation and performance indicators*

Refer to usage statistics presented in section 2.5. Further statistics on the number of registered users and number of processed runs can be found at:

<http://haddock.science.uu.nl/enmr/services/HADDOCK2.2/haddock.php>

Usage statistics as of June 24<sup>th</sup> 2015:

#### HADDOCK WEBSERVER STATISTICS

Server statistics generated on: 2015-06-25 11:36:36

Number of running requests on milou: 59, of which 47 on the WeNMR grid

Number of queued requests on milou: 93

Total number of served requests as of June 1st 2008: 102354 , of which 28028 on the eNMR grid

Number of registered users: ( 4595 easy / 289 expert / 570 guru)

Number of registered users for the [grid-enabled portal](#): 5732

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

Embarrassingly parallel – a large number of single jobs are sent to grid or local clusters.

It does not make sense to parallelize the computational engine used (CNS, see [www.cns-online.org](http://www.cns-online.org)).

CNS does support openmp though, but considering the large volume of jobs, the best throughput is obtained with a large number of single core jobs.

*Specify if the simulation software can exploit GPUs*

NO

*Specify how the simulation software exploits multicore systems*

By running multiple concurrent jobs

*Specify if parametric runs are required*

NO

*Estimate the use required of the resources (million-hours, # cores in parallel, job duration, etc)*

See Section 3.4. A typical run would require about 1 hour on 100 cores to run the complete HADDOCK workflow. But depends on the input data and parameter settings. The server itself is managing multiple runs with consequently large CPU usage (see statistics above and Figure 3 showing the number of single jobs sent to the grid by the HADDOCK portal (not counting all runs performed on our local cluster resources).

### **5.3 Simulation Workflows**

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

N.A. – HADDOCK is running its own workflow making use of both local and grid resources (see Figure 6).

## 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.*

The technical requirements for the HADDOCK use case are defined in section 3.4. In short, pre-configuration of a HADDOCK portal on a virtual machine, with all the necessary software pre-installed, queuing system and compute nodes.

The users of this use case will be mainly the software developers and the operators of the portals.

We foresee the following scenario

### 1) Configuration/installation of the system:

- a) A preconfigured VM with Scientific Linux, http support, queuing system (e.g. torque) and a minimal number of nodes (should be configurable at a later stage to expand the system) is made available by INDICO-DataCloud.
- b) The software developers/portal operator install all software and services on the master node (scientific software, http configuration, ...). Where possible this should be automated (e.g. DOCKER? Rsync..? Guidance/advice will be required here).
- c) The software developers add all required users on the system (only those managing the computations – not the end users).
- d) The fully configured system is stored on some repository (EGI AppDB? Community repository).
- e) There should be a control on who is allowed to access this VM since there are software licensing issues.

### 2) Launching of a virtualized HADDOCK VM with compute capabilities

- a) A portal operator decides to launch a new virtualized portal with compute capabilities.
- b) He/she configures in some interface the number of CPU cores required (which defines the number of nodes required). The system automatically takes care of updating the related settings (e.g. nodes defined in the queuing system, host list, ...).
- c) He/she select available resources (or some brokering system automatically assigns available resources (should be limited to one site for efficient communication between nodes)).
- d) The virtualized cluster is launched
- e) Credentials of authorized users for HADDOCK are automatically updated (currently users can either use their WeNMR VRC credentials using our SSOSX module, or an internal database (simple text file) is used). The SSOSX module does however required authorization on the WeNMR VRC side for a specific IP to be added – so not so simple. Or some INDICO-DataCloud AAI solution is adopted.
- f) A test run is performed on the virtualized server to make sure the portal works fine.
- g) The portal is put into production mode and made accessible to users (we might need here an automated mechanism to advertise the IP of the portal on some website).

### 3) Operation of a virtualized HADDOCK VM with compute capabilities

- a) Since a typically HADDOCK run on say 100 CPU cores might take between 1 and 10 hours depending on the data provided by the user, the virtual cluster should be up and running for several days (or even weeks).
- b) **Results of the computations should be stored and made available to the end users on an external system (with http capabilities to serve the result pages) so that data can be accessed after the lifetime of the virtual cluster.** Since the user is the owner of the data, results can be stored for a limited period of time (e.g. a few weeks).

#### 4) Shutting down of a virtualized HADDOCK VM with compute capabilities

- a) If a VM hits the real limits on the site where it is running, and the service is marked as continuous operation, there should be a mechanism to freeze the virtual cluster and migrate it to another available site to ensure smooth and continuous operation.
- b) Turning off the virtualize cluster should mean closing the portal for end users, but allowing current runs to complete so that users do not experience lost jobs.

## 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*

Local clusters with various connectivities, from 100MB to Infiniband. Gigabites uplink to the “world”

*Describe the key requirements (availability, bandwidth, latency, privacy, etc)*

High availability and reliability both for external access (key to the end user) and within the compute infrastructure for execution.

*Specify any current issue (like last mile, or access from commercial, etc)*

#### 7.1.2 Computing: Clusters, Grid, Cloud, Supercomputing resources

*Describe the current use of each of these type of resources: size and usage*

HADDOCK portals are currently running on both local resources (e.g. 200 CPU core cluster with PBD queuing system and 700 CPU core cluster with torque queueing system) and making use of grid resources (via glite or DIRAC4EGI submission). HADDOCK has also been running on HPC resource via HPC-Europe(2) project (the Utrecht lab is a host lab associated with the SURFSara resources).

*Indicate if there is any mode of “orchestration” between them*

The grid-enabled HADDOCK portals do require also local resources for pre- and post-processing.

#### 7.1.3 Storage

*Describe the current resources used*

A typical HADDOCK run might generate up to 10-50 GB of data while running. The results of the computations are presented to the user in web interface with an option to download the complete archive of the run, which typically represents in the order of a few 100MB to 5-10 GB. Data are stored only for a maximum of two weeks, after which they are deleted. It is up to the user to save his/her data.

*Discuss the key requirements (I/O performance, capacity, availability, reliability, any other QoS indicator)*

A fast connection between nodes and master node is required during execution (1GB or higher ideally).



## 7.2 Short-Midterm Plans regarding e-Infrastructure use

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

The following e-Infrastructure requirements have been recently submitted to the EGI-Engage project as input for possible SLA with NGIs. We are including all those requirements here, although some are not relevant for the HADDOCK use case (those are gray shaded)

In order to ensure the continued successful operation of the services offered by both WeNMR and N4U in the context of the MoBrain CC under EGI-Engage, but also to include the activities of the future VRE project West-Life, which effectively builds upon the WeNMR achievements and broaden its activities to server the entire structural biology community, we foresee the following e-Infrastructure requirements for the coming years (note that these numbers are a baseline estimate for the first year and will need to be adapted expecting a growth in the future):

<b>Estimated CPU requirements (HEPSPEC06):</b>	<b>3000 CPU years</b> (current WeNMR usage is 2600 CPU years)
<b>Estimated storage space:</b>	<b>100 TB</b> (current N4U usage is 10TB, WeNMR usage is negligible). However, the cryo-EM use case within MoBrain (related to Objective 2 above) will require quite some storage space)
<b>Estimated GPGPU requirements</b>	<b>Test infrastructure in first instance</b> to develop and test portals (linked to objective 3 above). Future requirements will have to be defined as a later stage (over one year) once portals become operational.
<b>Queue configuration requirements</b>	<ol style="list-style-type: none"><li>1) As indicated in the enmr.eu VO card, because of the large variety of applications and their different CPU requirements, we would ideally need several queue with different time limits to ensure an efficient execution of jobs of various time requirements.</li><li>2) Some applications can make use of multithreading parallelism (e.g. Gromacs for molecular dynamics). Getting access to some resources (nodes) with a larger number of cores (e.g. 24 or 48) would benefit those applications. These can be selected using standard JDL requirements.</li></ol>

<b>Cloud requirements</b>	Activities in both the MoBrain CC (objective 2 above) and the future West-Life VRE will be requiring cloud resources. <b>At this time the only requirements are for a test bed infrastructure.</b> Once the cryo-EM cloud project will have been completed we foresee a large increase in cloud resources use (to be defined at a later stage – over one year).
<b>Services requirement: CVMFS</b>	<b>CMVFS is currently our preferred way of remotely deploying the software</b> we are managing. Currently about 1/3 of the sites supporting us have CVMFS in place for our VO. This also applies to the OSG sites on which CVMFS repos are replicated in the OASIS system.
<b>Services requirement: DIRAC4EGI</b>	Only one of our portal is currently making use of DIRAC4EGI, but this one is handling a large job volume. <b>DIRAC4EGI is a very efficient submission mechanism which we intend to extend to other portals in the future</b> (an unfunded task in MoBrain, but operation will be supported in the West-Life VRE)
<b>Services requirement: WMS</b>	Most portals are still using a direct gLite-based submission. Since we do not pre-define the sites on which our jobs should be running, but are using software tags and timing/availability requirements, <b>the WMS system is crucial for a smooth execution.</b>
<b>Services requirements: VOMS</b>	<b>VOMS support is required</b> to keep operating our VO, which will also be used to support the broader INSTRUCT structural community.
<b>Services requirements: accounting</b>	The <b>EGI accounting portal</b> is an important tool to collect statistics for reporting.
<b>Open data preservation and storage</b>	Although the processed / analysed data in structural biology are typically deposited in public databases (e.g. <a href="http://www.pdbe.org">www.pdbe.org</a> , <a href="http://www.bmrbl.wisc.edu">www.bmrbl.wisc.edu</a> ), the raw data are in most case only stored locally (often without clear policies or metadata). We foresee a need to long term preservation of raw experimental data. Next

	<p>to experimental data, there is also a need for open repositories for modelling data (i.e. results of simulations rather than experiments). This can be within institutional repositories (often not yet present), EUDAT or related initiatives. <b>Sharing / preserving data (and making them citable) in the context of a federated data cloud under EGI is a scenario that will need to be investigated.</b></p>
--	---

Currently the WeNMR enmr.eu VO has access to ~110'000 CPU cores distributed over 42 different sites, including OSG resources, which provide enough variety to ensure a smooth operation. See <http://gstat.egi.eu/gstat/summary/VO/enmr.eu> .

### 7.2.1 Networking

*Describe the proposed connectivity*

Fast network connecting the nodes and the master (at least 1GB), fast connecting to the outside world (at least 100MB, ideally 1GB) for data transfer (upload/download)

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

Internal network between virtualized cluster components must be fast and reliable

Connection to outside world must be 100% available in order to server users

### 7.2.2 Computing: Clusters, Grid, Cloud, Supercomputing resources

*Describe the evolution expected: which infrastructures, total "size" and usage*

See section 3.4

### 7.2.3 Storage

*Describe the resources required*

See section 3.4

*Discuss the key requirements (I/O performance, capacity, availability, reliability, any other QoS indicator)*

See section 3.4 and 7.2.1

Reliability and availability are key in order to server the user community! Which means ~100% availability and reliability (with solutions to migrate the complete virtualized cluster/web portal to other resources when a site becomes unavailable). Data should not be lost.

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

The HADDOCK portal has been under operation for more than 7 years, making use of EGI resources over the entire period, under two former FP7 e-Infrastructure projects (eNMR and WeNMR). This use case is well documented and known under EGI. Refer to previous relevant documents.

### **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/constraints 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.*

A simple way to monitor the load on HADDOCK platform is monitor the number of running and queuing jobs, while also monitoring the load of the system (e.g. via ganglia). See for an example:

<http://haddock.science.uu.nl/stats.php>

For the grid usage statistics, we keep track of the job submissions on the portal side (end user), at the grid submission level on the portal site, and we also make use of the EGI accounting portal to monitor the number of jobs and CPU usage. By making use of roles, we can have an application specific accounting.

### **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?*

As described in section 3.2, our current users are either registered directly with the HADDOCK portal (see the registration form at: <http://haddock.science.uu.nl/services/HADDOCK2.2/signup.html> - requires manual approval of the user), obtaining their credentials in the form of username and password for job submission, or can make use of their WeNMR VRC credentials for submitting jobs to the web portals through the single-sign-on mechanism for external services (SSOSX) developed under the WeNMR project (see for details: <http://www.wenmr.eu/wenmr/wenmr-ssso-module> ).

This module can also make use of EDUGAIN, but currently only Dutch Universities are supported through the SURFConext modules of SURFSara. In the past other EU institutions were also supported. Getting credential via EDUGAIN requires however a one to one negotiation with each site, which is way too much work to be a viable solution at this time.

- *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*

Users are not required to have a personal X509 certificate to use the HADDOCK portal. Registration directly on the portal page, or via the WeNMR VRC is sufficient.

The grid-enabled version of the HADDOCK portal makes use of a X509 robot certificate to send jobs to the grid.

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

Please refer to previous relevant sections. All information has already been provided above.  
 Several points below seem completely irrelevant (e.g. the connection between AAI and multithreading!)

<b>Software used</b>	<i>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.)</i> <input here>
<b>Operating system requirements</b>	<input here>
<b>Run libraries requirements</b>	<i>Run API/libraries requirements (e.g., Java, C++, Python, etc.)</i> <input here>
<b>CPU requirements (multithread,MPI, “wholenode”)</b>	<input here>
<b>Memory requirements</b>	<input here>
<b>Network requirements</b>	<input here>
<b>Disk space requirements (permanent, temporal)</b>	<i>Include the requirements for data transferring (upload and download of data objects: files, directories, metadata, VM/container images, etc.)</i> <input here>
<b>External data access requirements</b>	<input here>
<b>Typical processing time</b>	<input here>
<b>Other requirements</b>	<i>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</i>

	<input here>
<i>Other comments</i>	<input here>
<i>Relevant references or URLs</i>	<input here>

- *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.

N/A

## 7.6 Initial short/summary list for “test” applications (task 2.3)

It will be the HADDOCK portal developers and operators that will take care of installing the necessary software provided a basic linux version with a proper queuing system and standard compilers (gnu c, c++ and gfortran) is provided. Please refer to section 3.4. Other “virtual hardware requirements” have already been defined under section 3.4. The HADDOCK software is described on its online manual at <http://bonvinlab.org/software> (as already indicated previously in sections 2.2 and 3.3).

## **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**

<b>R 1</b>	
<b>R 2</b>	
<b>R 3</b>	
<b>R 4</b>	
<b>R 5</b>	