Milestone M4.1

Project Title:	World-wide E-infrastructure for structural biology	
Project Acronym:	West-Life	
Grant agreement no.:	**675858**	
Title	Inventory of available resources and testbed setup	
WP No.	4	
Lead Beneficiary:	4: MU	
WP Title	Operation and maintenance of the computing and data infrastructure	
Contractual delivery date:	2/2016	
Actual delivery date:	2/2016	
WP leader:	Aleš Křenek	4: MU
Contributing partners:	INFN, Luna	

1 Executive summary

In the first months of the project WP4 gathered detailed information on technology used by 13 scientific portals operated by project partners. The information is presented in consistent form, and it will become input to the definition of the project common architecture (D4.1).
Project testbed was set up by negotiating access to cloud resources (3 project partners and additiona EGI site). The cloud resources are complemented by grid resources, which provide access to the community via the enmr.eu VO. It has been clarified that there is no interference between the testing and production traffic on these resources, hence there is no need to set up and maintain separate testbed. All the testbed resources are monitored by standard EGI tools.

2 Project objectives

With this milestone, the work done contributed to the following project objectives:

No.	Objective	Yes	No
1	**Provide analysis solutions for the different Structural Biology approaches**	x	
2	**Provide automated pipelines to handle multi-technique datasets in an integrative manner**	x	
3	**Provide integrated data management for single and multi-technique projects, based on existing e-infrastructure**		x
4	**Foster best practices, collaboration and training of end users**		x

3 Detailed report on the milestone

3.1 Survey of existing portals and technology used

Information on existing scientific portals operated by the project partners was gathered in collaboration with WP5 (work on D5.2), it's initial phase was done by circulating a questionnaire focused on technology used by the implementation. The main question categories were:

- quantitative – number of user requests, size of data uploaded/downloaded, amount of CPU time etc.,
- security – how the users are registered, what are the mechanisms to enforce authentication and authorization etc.
- job management – what software is used to keep control on the running calculations, are jobs handled in a local cluster or submitted remotely,
- resources – what storage is used, what are the computing resources locally attached to the portal, what are the grid sites supporting it,
- application software – what software is used, and how is it distributed (installed and updated) in the infrastructure

The full text of the questinnaire is attached to this document as well as a table with complete answers. Altogether 13 answers were received, covering all the project partners who operate scientific portals, and giving representative samlple of the technology.

The gathered data will be used extensively in designing the common infrastructure model (D4.1). At this point in time we make preliminary observation:

Data uploaded on job submission. Generally up to a few tens of MB of input data is required, exception is Scipio (CryoEM), the input can be up to 2TB.

Data downloaded as the result. Generally from MBs to GBs, exceptions is CPP4online (Balbes/MrBUMP) – up to tens of GB (but it is not typical use case

Background data. Generally portals do not require any background data, CPP4online uses local copies of PDB and HHpred, downloaded and processed before.

Usage of local and remore resources. 6 portals use remote resources, (EGI grid + gLite, Haddock sends some jobs via Dirac), simultaneous use of local resources for

pre/postprocessing and grid jobs for the main payload is typical. 5 portals use only local resources.

User submissions and grid jobs. Hundreds or small thousands of user submissions translate into thousands or tens of thousands grid submissions. The ratio is higher for Haddock – 25k user submissions jobs translates into 7,5M grid jobs, and CS-Rosetta3 – 67 jobs translates into 189k grid jobs.

MPI and shared filesystem. Not used/required with the exception of Scipion and AutoRickshaw.

3.2 Testbed setup

The project testbed provides resources to develop and deploy prototype of the consolidate portal architecture emerging from the project. According to the workplan the current testbed is expected to become part of the production infrastructure by November 2016, and a new generation of the testbed will be set up.

3.2.1 Resource inventory

Cloud resources

The following table summarizes the cloud resources available for the testbed in February 2016:

Cloud name	Cloud framework	Institute	Phys.Cores	Phys.RAM	GPUs	Block Storage
INFN-PADOVA-STACK	OpenStack/Juno	INFN	144	283 GB	0	3.7 TB
CESNET-MetaCloud	OpenNebula	MU	80 (+400 best effort)	512 GB (+2.5 TB)	4x M2090 (shared)	5 TB (+25 TB)
IISAS-GPUCloud	OpenStack/Kilo	IISAS	96	384 GB	12x K20 (shared)	6 TB

In particular, *CESNET-MetaCloud* committed to provide resources by signing a SLA with EGI-Engage MoBrain competence centre, which collaborates with the project closely.

For the time being, MU collaborates with CESNET to provide access to the resources as a single site. We count MU resources available to Westlife here, not the whole FedCloud site. A standalone MU site is planned by the end of 2016.

In addition SurfSara provides cloud machines to test Scipion, but not via FederatedCloud.

Grid resources

We consider the existing EGI resourcess accessed in the legacy ways and accessible to the enmr.eu VO to be part of the Westlife testbed because they can accept jobs of the community portals.

On the other hand, it is not worth building a separate parallel infrastructure in this way. List of the sites in the time of writing this document is attached as separate file. Altogether more than 120,000 CPU cores is available. In particular, the following sites signed the SLA with MoBrain, committing resources to the community. INFN-PADOVA (Italy), RAL-LCG2 (UK), TW-NCHC (Taiwan), SURFsara (The Netherlands), NCG-INGRID-PT (Portugal), NIKHEF (The Netherlands).

Further, CIRMMP provides a prototype site with GPU support available through the grid interface – 3 nodes (2x Intel Xeon E5-2620v2) with 2 NVIDIA Tesla K20m GPUs per node and GPU-enabled AMBER and GROMACS.

3.2.2 Access

All the sites provide access through the enmr.eu virtual organization. However, effectively only robot certificates are used to submit jobs, not user ones. Therefore we plan restrict access to the sites to the robot certificates only. Technically this will be done by a dedicated VOMS role in the existing VO.

3.2.3 Monitoring and accounting

Because the testbed resources are fully integrated in EGI, we leverage the standard EGI monitoring and accounting tools. The following links provide access to the current data:

Grid Resources from EGI Monitoring tools:

- Data from Gstat
- Data from Vapor

Cloud Resources from EGI Monitoring tools:

- Status from ARGO
- Monthly Availability from ARGO
- CESNET-MetaCloud Status from Nagios
- INFN-PADOVA-STACK Status from Nagios

Grid Resources from EGI Accounting Portal (last 12 months data):

- Site vs #jobs
- Site vs CPU time (HEPSPEC06.hours)

Appendices

The following additional files are attached

- M4_1_grid_resources.xlsx – list of grid resources supporting enmr.eu, hence available for the project testbed too
- M4_1_survey.pdf – questionnaire on portal technologies circulated among project partners
- M4_1_survey_results.xlsx – full results of the survey