



Deliverable 3.3

Request for Enhancements of gLite to support bio-NMR applications

SEVENTH FRAMEWORK PROGRAMME Research Infrastructures

INFRA-2007-1.2.2 - Deployment of eInfrastructures for scientific
communities

**Grant agreement for: Combination of Collaborative projects &
Coordination and support actions**

Proposal/Contract no.: 213010 – **e-nmr**

Project full title: Deploying and unifying the NMR e-Infrastructure in System Biology

Project coordinator: Prof. Dr. Harald Schwalbe

Project website: <http://www.enmr.eu/>

Period covered: from **01-11-2007 to 30-4-2009**

Table of contents

1. INTRODUCTION	4
1.1. PURPOSE	4
1.2. DOCUMENT ORGANISATION	4
1.3. REFERENCES	4
1.4. TERMINOLOGY	4
2. EXECUTIVE SUMMARY.....	6
3. SURVEY OF APPLICATIONS CURRENTLY AVAILABLE IN E-NMR GRID	7
3.1. APPLICATION DESCRIPTION	7
3.2. APPLICATIONS CHARACTERISTICS	9
3.2.1. <i>Is the application parallel (MPI) or sequential?</i>	9
3.2.2. <i>Is the application CPU intensive, data intensive or both?</i>	10
3.2.3. <i>Is the application interactive or batch oriented?</i>	11
3.2.4. <i>Does the application have security requirements?</i>	11
3.2.5. <i>Does the application have encryption requirements for data storage or transfer?</i>	12
3.2.6. <i>Does the application make use of grid enabled data storage for input and/or output data?</i> 12	12
3.2.7. <i>What is the typical size of the application software package?</i>	13
3.2.8. <i>Is the application installed locally in each cluster or is it downloaded with the job submission?</i>	14
3.3. RESOURCE REQUIREMENTS.....	14
3.3.1. <i>What is the typical CPU time consumption per run?</i>	14
3.3.2. <i>What is the amount of RAM required at run and compile time?</i>	15
3.3.3. <i>What is the typical number of concurrent running/queued jobs?</i>	16
3.3.4. <i>Disk space requirements per run?</i>	16
3.3.5. <i>Does the application require direct network connectivity inbound/outbound or both?....</i>	17
3.3.6. <i>Typical input data size</i>	18
3.3.7. <i>Typical output size</i>	19
3.4. SOFTWARE DEPENDENCIES.....	19
3.4.1. <i>Operating system and version?</i>	20
3.4.2. <i>Compilers and versions?</i>	20
3.4.3. <i>Databases and versions?</i>	21
3.4.4. <i>Other required libraries?</i>	21
3.4.5. <i>System or general tools required at run or compile time?</i>	22
3.4.6. <i>Does the application need party commercial software to run?</i>	23
3.5. GLITE REQUIREMENTS	23
3.5.1. <i>Which grid elements other than CE/WNs your application make use of? (e.g. WMS, SE, LFC, AMGA, HYDRA, MYPROXY, CREAM)</i>	23
3.5.2. <i>Does the application use standard Data management functionality? (i.e. file registration into LFC file catalogue, file access through LFN, file transfer/replication through standard lcg- utils)</i> 24	24
3.5.3. <i>How many files per run are registered with the LFC file catalogue?</i>	25
3.5.4. <i>Does the application use advanced functionality for Data Encryption? (EDS / Hydra servers)</i>	26
3.5.5. <i>Are the normal users submitting the jobs from a UI or through a web portal?</i>	26
3.5.6. <i>Does the application use WS, Java or C/C++ APIs for job submission and/or data management, or simply glite-wms-*, lcg-*and lfc-* command line tools?</i>	27
3.5.7. <i>Does the application make use of Collection, Parametric or DAG job type functionalities?</i> 27	27
3.5.8. <i>Did you find any limitation on gLite current functionalities?</i>	28
3.5.9. <i>Did you find any limitation on gLite current performances?</i>	28
3.5.10. <i>Did you expect any functionality from gLite that was completely missing in the current release, and which could be of benefit for your application?</i>	29
4. SURVEY ABOUT DATA SECURITY ISSUES	30

4.1.	DO YOU WORK FOR?	30
4.2.	DO YOU KNOW WHAT GRID INFRASTRUCTURES ARE?	30
4.3.	DO THE POLICIES OF YOUR ORGANIZATION ALLOW YOU TO SEND YOUR DATA OVER THE INTERNET TO BE USED AS INPUT FOR CALCULATIONS BY WEB SERVERS?	31
4.4.	IF YOUR ANSWER TO THE PRECEDING QUESTION WAS 4, WHAT IS YOUR MAJOR CONCERN?	31
4.5.	WHAT KIND OF MEASURES WOULD YOU NEED TO SATISFY YOUR CONFIDENTIALITY REQUIREMENTS? 32	
4.6.	WHAT KIND OF MEASURES WOULD YOU NEED TO SATISFY YOUR DATA AVAILABILITY REQUIREMENTS?	32
4.7.	IF YOUR ANSWER TO 4.3 WAS NO, WOULD YOU BE INTERESTED IN OBTAINING A VERSION OF THE E-NMR PLATFORM FOR LOCAL INSTALLATION?	33
5.	SUMMARY	34

1. Introduction

1.1. Purpose

This document is the project deliverable D3.3 due by Month 18. It aims at reporting the activities of the task T3.2: Enhancements of grid middleware for e-NMR support, assigned to the Work Package 3 of the e-NMR project.

1.2. Document organisation

The document is organised as follows:

Section 1 contains the purpose of the document, its references and a glossary of terms and acronyms;

Section 2 summarizes the content of the document;

Section 3 reports the results of a survey distributed inside the consortium partners about the bio-NMR applications currently deployed on the e-NMR grid;

Section 4 reports the results of a survey distributed to the wider bio-NMR user community outside the consortium, about their requirements with respect to data security.

Finally, Section 5 tracks the conclusions.

1.3. References

[R1]	Public link not available	e-NMR Annex I
[R2]	glite.web.cern.ch	gLite middleware web page
[R3]	www.e-nmr.eu	e-NMR Project web page
[R4]	http://knowledge.eu-egi.eu/knowledge/index.php/UMD	EGI / UMD Knowledge Base
[R5]	http://www.jspg.org/wiki/VO_Portal_Policy	JSPG VO Portal Policy

1.4. Terminology

This subsection provides the definitions of terms, acronyms, and abbreviations required to properly interpret this document.

Term	Definition
ACL	Access Control List
AMGA	gLite Metadata Catalogue service
BCBR	Bijvoet Centre for Molecular Research, University of Utrecht, The Netherlands
BMRZ	Centre for Biomolecular Magnetic Resonance, Goethe University, Frankfurt, Germany
CE	Computing Element
CIRMMMP	Interuniversity Consortium for Magnetic Resonance on Metalloproteins, Florence, Italy
CREAM	Computing Resource Execution And Management
DAG	Direct Acyclic Graph, basic job workflow implemented by gLite
DoW	Description of Work
EGEE	Enabling Grids for e-Science
EDS	Encrypted Data Storage
EGI	European Grid Initiative
gLite	Codename of the middleware software suite developed by EGEE
Hydra	Keystore server for encrypted file storage solution
IGI	Italian Grid Infrastructure
INFN	National Institute of Nuclear Physics, Italy
JSPG	Joint Security Policy Group
LCG	LHC Computing Grid
LFC	LCG File Catalogue
LFN	Logical File Name
NGI	National Grid Initiative
NMR	Nuclear Magnetic Resonance
SE	Storage Element
UI	User Interface
UMD	Unified Middleware Distribution
VO	Virtual Organization
VOMS	Virtual Organisation Membership Service
WMS	Workload Management System
WP	Work Package

2. Executive summary

As stated in the DoW [R1], the main objective of the task T3.2 is

“to identify the enhancements required by the selected NMR applications that need to be integrated and shared in the European NMR Research Infrastructure and in the general European e-Infrastructure.

[...]

Based on NMR system requirements and the EGEE existing and planned features, we will define the research developments required to enhance EGEE for fully enabling NMR applications. The EGEE middleware deployed and configured on partner sites as part of the Service activity will be useful to test and evaluate the current release of EGEE in a bio-NMR setting. Using the e-NMR infrastructure, a study for the NMR framework integration with the EGEE middleware will be performed. This will allow us to evaluate which parts of the middleware need to be enhanced.

[...]

This deliverable will explicitly take into consideration also the security issues that can arise from the interaction with users and industrial stakeholders. In particular, the relevant communities will be surveyed for their requirements with respect to data security.”

This document thus describes the requirements of the bio-NMR applications already deployed on the e-NMR grid at Month 18, and more in general the requirements related to data security coming from bio-NMR community.

The following methodology was adopted:

- A survey with 31 questions about characteristics of the applications currently available on the e-NMR platform was circulated within the partners of the consortium. The questions were grouped into 5 items: Application description, Application characteristics, Resource requirements, Software dependencies and gLite requirements.
- An on-line survey with 7 questions about security needs of NMR applications was made available at the project web portal (see <http://www.enmr.eu/grid>) and circulated among the bio-NMR user community outside the consortium. Two weeks after the advertising of the target community, a total of 55 answers were collected.

Concerning the internal survey, it turned out that the applications currently deployed on the e-NMR grid had not very stringent requirements and could quite easily been interfaced with the latest gLite middleware exploiting its very basic functionalities. For some of the applications there is a plan for the second phase of the project to exploit more advanced gLite capabilities like MPI support and one-shot job collections submission.

Concerning the survey about data security issues, 96% of the answers came from Academia and others no-profit organizations, while only 4% from private industrial sector. From the small minority of respondents who indicated security as an issue, about one third come from organizations having policies preventing to send sensitive data over internet to be used as input for NMR calculations by the applications web servers. For this small minority, confidentiality of data handling is the major concern, to be achieved via encryption and ACL-based access.

3. Survey of applications currently available in e-NMR grid

3.1. APPLICATION DESCRIPTION

NMRPipe

NMRPipe is an extensive software system for processing, analyzing, and exploiting NMR spectroscopic data. The software has been deployed on the e-NMR grid as a requirement for CS-ROSETTA (see below). For details see: <http://spin.niddk.nih.gov/NMRPipe>

PROSA

The program package PROSA is an efficient implementation of the common data processing steps for multi-dimensional NMR spectra. PROSA performs linear prediction, digital filtering, Fourier transformation, automatic phase correction, baseline correction, and other common NMR data processing methods. High efficiency is achieved by avoiding disk storage of intermediate data and by the absence of any graphics display, which enables calculation in the batch mode and facilitates porting the PROSA software package to a variety of different computer systems.

INFIT

INFIT is a method to determine scalar coupling constants from in-phase multiplets, focusing on spectra which are routinely recorded with a good signal-to-noise ratio during the course of protein structure determination by NMR, e.g. homonuclear [1H-1H]-NOESY spectra.

CNS versions 1.2 and 1.2-para

Crystallography & NMR System (CNS) is the result of an international collaborative effort among several research groups. The program has been designed to provide a flexible multi-level hierarchical approach for the most commonly used algorithms in macromolecular structure determination. Highlights include heavy atom searching, experimental phasing (including MAD and MIR), density modification, crystallographic refinement with maximum likelihood targets, and NMR structure calculation using NOEs, J-coupling, chemical shift, and dipolar coupling data. For details see <http://www.cns-online.org>

The 1.2-para version contains additional paramagnetic restraints routine written at CERM in Florence.

CYANA

The program package CYANA (Combined assignment and dYnamics Algorithm for Nmr Applications) for efficient calculation of three-dimensional protein and nucleic acid structures from distance constraints and torsion angle constraints collected by nuclear magnetic resonance (NMR) experiments performs simulated annealing by molecular dynamics in torsion angle space and uses a fast recursive algorithm to integrate the equations of motions. Torsion angle dynamics can be more efficient than molecular dynamics in Cartesian coordinate space because of the reduced number of degrees of freedom and the concomitant absence of high-frequency bond and angle vibrations, which allows for the use of longer time-steps and/or higher temperatures in the structure calculation.

XPLOR-NIH Version 2.21

X-PLOR is a program system for computational structural biology. X-PLOR stands for exploration of conformational space of macromolecules restrained to regions allowed by combinations of empirical energy functions and experimental data. But it also stands for exploration of modern concepts of structured programming in macromolecular simulation.

HADDOCK version 2.1

HADDOCK is a data-driven biomolecular docking program. It allows the modelling of a wide variety of biomolecular complexes such as protein-protein, protein-nucleic acids and protein-ligand complexes. A unique feature of HADDOCK is its capability to encode a wide variety of experimental information (such as for example from NMR, SAXS, mutagenesis) into highly ambiguous restraints to drive the docking process. The computational engine used in HADDOCK is CNS, a molecular dynamics and energy minimization program used in the NMR and Xray communities. For details see <http://www.nmr.chem.uu.nl/~haddock>.

CS-ROSETTA

CS-ROSETTA is a protocol which generates 3D models of proteins, using only the ¹³CA, ¹³CB, ¹³C', ¹⁵N, ¹HA and ¹HN NMR chemical shifts as input. Based on these parameters, CS-ROSETTA uses a SPARTA-based selection procedure to select a set of fragments from a fragment-library (where the chemical shifts and the 3D structure of the fragments are known). The fragments are assembled using the ROSETTA protocol. The generated models are rescored based on the difference between the back-calculated chemical shifts of the generated models and the input chemical shifts. For more information see <http://spin.niddk.nih.gov/bax/software/CSROSETTA>.

GROMACS version 3.3.3

Gromacs is a suite of programs for setting up, performing and analyzing molecular dynamics simulations. Gromacs is open source and distributed under GPL. It can be compiled with MPI support to be run on multiple processors. For more information see <http://www.gromacs.org/>. The current deployment on the grid enables running single Gromacs jobs, which are invoked from the UI using a simple command line interface. The same mechanism will be used to set up a simple web interface that allows setting up simulations and perform short runs for biomolecules.

AMBER Version 10

Amber is the collective name for a suite of programs that allow users to carry out molecular dynamics simulations, particularly on biomolecules. None of the individual programs carries this name, but the various parts work reasonably well together, and provide a powerful framework for many common calculations.

The programs used in grid are:

sander: Simulated annealing with NMR-derived energy restraints. This allows for NMR refinement based on NOE-derived distance restraints, torsion angle restraints, and penalty functions based on chemical shifts and NOESY volumes. Sander is also the "main" program used for molecular dynamics simulations, and is also used for replica-exchange, thermodynamic integration, and potential of mean force (PMF) calculations. Sander also includes QM/MM capability.

tLEaP: tLEaP is a program that provides for basic model building and Amber coordinate and parameter/topology input file creation.

Amber suit are used to refine the best structure calculated with XPLOr-NIH.

MAPPER

NMR sequence-specific resonance assignment

MDDNMR

MDDNMR is a program for processing of non-uniformly sampled (NUS) multidimensional NMR spectra. The package contains also a routine to produce NUS schedule that can be used to run two to four dimensional NUS NMR experiments. Potentially any pulse sequence can be run in the NUS mode. In the NUS acquisition, only a fraction of full (conventional) data set is recorded. MDDNMR works by replenishing missing data points in the full matrix followed by regular FT processing of the complete data.

3.2. APPLICATIONS CHARACTERISTICS

3.2.1. Is the application parallel (MPI) or sequential?

PROSA

Both

INFIT

Sequential

CNS

Sequential

CYANA

Both. CYANA structure calculations have been shown to parallelize with nearly ideal speedup up to a number of processors that equals the number of protein conformers that are calculated by the algorithm using the same input data from NMR experiments but different random initial conformations of the molecules. Benchmark calculations using a Linux cluster system with 20 nodes, each having 2 Intel Xeon E5462 quad-core CPUs, 2.80 GHz, 16 GB memory, Ubuntu 8.04 Linux, Intel Fortran compiler, OpenMPI have shown to scale very well.

HADDOCK

Sequential with parallelism in the form of a large amount of distributed parallel jobs

XPLOr-NIH

The application can be compiled with MPI, but our implementation is sequential.

The calculations of protein structures are done using a simulated annealing algorithm. To calculate 200 structures is better to run 10 jobs with 20 structures than run 200 structure using 10 parallel processors.

AMBER

Sander can be compiled using MPI, but normally we run a single job for any best structure calculated with XPLOR-NIH

CS-ROSETTA

Sequential with parallelism in the form of a large amount of distributed parallel jobs

GROMACS

Although Gromacs is MPI enabled, the MPI support from gLite has not been exploited yet. This is scheduled for the second phase.

MAPPER

Sequential

MDDNMR

Sequential with parallelism in the form of a large amount of distributed parallel jobs. MPI version of the program is available in the case of dense (not NUS) input data.

3.2.2. Is the application CPU intensive, data intensive or both?

PROSA

Both

INFIT

CPU intensive

CNS

CPU intensive

CYANA

Both

HADDOCK

CPU intensive

XPLOR-NIH

CPU intensive

AMBER

CPU intensive

CS-ROSETTA

CPU intensive

GROMACS

CPU intensive

MAPPER

CPU intensive

MDDNMR

CPU intensive

3.2.3. Is the application interactive or batch oriented?

PROSA

Both, mainly batch

INFIT

Both, mainly interactive

CNS

Batch oriented

CYANA

Both, mainly batch

HADDOCK

Batch oriented

XPLOR-NIH

Batch oriented

AMBER

Batch oriented

CS-ROSETTA

Batch oriented

GROMACS

Batch oriented

MAPPER

Both, mainly batch

MDDNMR

Batch oriented

3.2.4. Does the application have security requirements?

PROSA

Yes, it can be used only by licensed users

INFIT

No

CNS

No

CYANA

Yes, it can be used only by licensed users. Currently a demo version of the program that can be accessed freely is installed at the e-NMR sites.

HADDOCK

No

XPLOR-NIH

No

AMBER

No

CS-ROSETTA

No

GROMACS

No

MAPPER

No

MDDNMR

No

3.2.5. Does the application have encryption requirements for data storage or transfer?

None of the applications considered have such requirement.

3.2.6. Does the application make use of grid enabled data storage for input and/or output data?

PROSA

No, but it could be considered in the future.

INFIT

No

CNS

Yes, for input data

CYANA

No, but it could be considered in the future.

HADDOCK

Yes, for input data

XPLOR-NIH

Yes, for output data

AMBER

Yes, for output data

CS-ROSETTA

Yes, for input data

GROMACS

No

MAPPER

No

MDDNMR

Not yet

3.2.7. What is the typical size of the application software package?

PROSA

10 MB

INFIT

3 MB

CNS

The total size of the package is 500 MB, 250MB for each version (1.2 and 1.2-para).

CYANA

50 MB

HADDOCK

The size of the package is 52MB

XPLOR-NIH

The total size of the package is 382 MB

AMBER

The total size of the package is 856 MB

CS-ROSETTA

The total size of the deployed packages to run CS-Rosetta is ~3.5GB. Three packages need to be installed in order to run CS-Rosetta:

- nmrPipe (712 MB)
- Rosetta 2.3.0 (2.7 GB)
- CS-Rosetta (267 MB)

GROMACS

The total size of the package is 250 MB

MAPPER

2 MB

MDDNMR

The total size of the package is 100 MB

3.2.8. Is the application installed locally in each cluster or is it downloaded with the job submission?

For all of the applications considered the software is installed locally on each CE.

3.3. RESOURCE REQUIREMENTS

3.3.1. What is the typical CPU time consumption per run?

PROSA

30 minutes

INFIT

6 minutes

CNS

CPU requirements vary depending on the type calculations performed (from a few minutes to several days). Within HADDOCK, typical run times per CNS job are around 30 minutes.

CYANA

2-3 CPU hours

HADDOCK

A typical user run requires around 130 CPU hours distributed over 250 grid jobs. In addition, about 5 hours are needed for its preparation and analysis on a local cluster.

XPLOR-NIH

From 0.5 to 4 hours for 10 structures calculation per core (the time depends on the size of the protein)

AMBER

From 1 to 70 hours per core (depending on the numbers of residues in the protein)

CS-ROSETTA

A typical run will require around 10000 CPU hours distributed over 2500 grid jobs of each 4 hours + about 10 hours preparation and analysis on a local cluster

GROMACS

CPU time depends heavily on the size of the system in numbers of particles, and runs can vary from a few hours to weeks.

MAPPER

10-15 CPU hours

MDDNMR

It varies from 10 CPU minutes to 300 CPU hours

3.3.2. What is the amount of RAM required at run and compile time?

PROSA

4 GB at run time, 1 GB at compile time

INFIT

500 MB

CNS

1 GB RAM

CYANA

2 GB at run time, 1 GB a compile time

HADDOCK

A few 100MB RAM

XPLOR-NIH

From 100 MB to 600 MB, depending on the size of the protein

AMBER

50 to 500 MB depending on the number of residues

CS-ROSETTA

1 GB RAM

GROMACS

In most cases 1 GB RAM is sufficient

MAPPER

1 GB at run time, 500 MB at compile time

MDDNMR

It varies from 128 to 512 MB RAM

3.3.3. What is the typical number of concurrent running/queued jobs?

PROSA

2

INFIT

20

CNS

Depends on the way the user submits the jobs: from 1 to a few 100

CYANA

5

HADDOCK

50 or 100 depending on the stage in the protocol

XPLOR-NIH

Depending on the request of calculation: from 1 to 15

AMBER

Depending on the request of calculation: from 1 to 30

CS-ROSETTA

500 to 2500 depending on parameter settings

GROMACS

Between 1 and 10

MAPPER

2

MDDNMR

From 10 to 30

3.3.4. Disk space requirements per run?

PROSA

5 GB

INFIT

500 MB

CNS

A few hundred MB on the local cluster, about 20-40 MB on the WN's

CYANA

2 GB

HADDOCK

1 to 2 GB on the local cluster, about 20-40 MB on the WN's

XPLOR-NIH

Depending on the request of calculation: from 10 to 300 MB

AMBER

Depending on the request of calculation: from 1 to 5 GB

CS-ROSETTA

60-80 GB on the local cluster, about 20-40 MB on the WN's

GROMACS

Disk usage depends on the system size, as well as on parameters controlling the frequency of writing and the part of the system to be included in the output. At present jobs are set up to produce no more than several hundreds of MB, but demands for storage are likely to rise to several GB per run

MAPPER

1 GB

MDDNMR

It varies from 100 MB to 10 GB depending on data size

3.3.5. Does the application require direct network connectivity inbound/outbound or both?

PROSA

No

INFIT

No

CNS

Only for getting the input data from a SE

CYANA

No

HADDOCK

Only for getting the input data from a SE

XPLOR-NIH

No

AMBER

No

CS-ROSETTA

Only for getting the input data from a SE

GROMACS

No

MAPPER

No

MDDNMR

No

3.3.6. Typical input data size

PROSA

2 GB

INFIT

2 MB

CNS

10-20 MB

CYANA

500 MB

HADDOCK

10-20 MB

XPLOR-NIH

10-20 MB

AMBER

10-20 MB

CS-ROSETTA

10-20 MB

GROMACS

A simulation system description of several MB or even a single PDB file of several KB

MAPPER

100 MB

MDDNMR

It varies from 100 MB to 10 GB depending on data size

3.3.7. Typical output size

PROSA

1 GB

INFIT

2 MB

CNS

5-10 MB

CYANA

1 GB

HADDOCK

5-10 MB

XPLOR-NIH

Depending on the request of calculation: from 10 to 300 MB

AMBER

Depending on the request of calculation: from 1 to 5 GB

CS-ROSETTA

5-10 MB

GROMACS

At present jobs are set up to produce no more than several hundreds of MB, but demands for storage are likely to rise to several GB per run

MAPPER

1 GB

MDDNMR

It varies from 100 MB to 10 GB depending on data size

3.4. SOFTWARE DEPENDENCIES

3.4.1. Operating system and version?

PROSA

Linux, SL4, Ubuntu, Fedora Core

INFIT

Linux, SL4, Ubuntu, Fedora Core

CNS

Scientific Linux on 32 or 64 bits Intel or opteron processors

CYANA

Linux, SL4, Ubuntu, Fedora Core

HADDOCK

Scientific Linux on 32 or 64 bits Intel processors

XPLOR-NIH

Scientific Linux on 32 or 64 bits Intel processors

AMBER

Scientific Linux on 32 or 64 bits Intel processors

CS-ROSETTA

Scientific Linux on 32 or 64 bits Intel processors

GROMACS

Scientific Linux on 32 or 64 bits Intel processors

MAPPER

Linux, SL4, Ubuntu, Fedora Core

MDDNMR

Any Linux flavour

3.4.2. Compilers and versions?

PROSA

Intel Fortran

INFIT

Gcc

CNS

CNS (version 1.2) is compiled in static mode using Intel compilers (ifort).

CYANA

Intel Fortran

HADDOCK

The computational engine CNS (version 1.2) is compiled in static mode using Intel compilers (ifort).

XPLOR-NIH

Intel compilers, GNU compilers

AMBER

Intel compilers, GNU compilers

CS-ROSETTA

The computational engine Rosetta (version 2.3.0) is compiled using GNU compilers (gcc and g++).

GROMACS

The compilation of Gromacs can be done with a number of different compilers, although GNU gcc version 4.1.x results in a broken installation

MAPPER

Intel Fortran

MDDNMR

Gcc 4.0

3.4.3. Databases and versions?

None of the applications considered makes use of databases.

3.4.4. Other required libraries?

PROSA

No

INFIT

No

CNS

No

CYANA

No

HADDOCK

No

XPLOR-NIH

No

AMBER

No

CS-ROSETTA

No

GROMACS

For compilation it is recommended to have the FFTW library available

MAPPER

No

MDDNMR

Glibc

3.4.5. System or general tools required at run or compile time?

PROSA

No

INFIT

No

CNS

Tar, gzip, gunzip

CYANA

No

HADDOCK

Tar, gzip, gunzip

XPLOR-NIH

Tar, gzip, gunzip

AMBER

Tar, gzip, gunzip

CS-ROSETTA

Tar, gzip, gunzip

GROMACS

Tar, gzip, gunzip

MAPPER

No

MDDNMR

Tar, gzip, gunzip, gawc, gcc

3.4.6. Does the application need party commercial software to run?

PROSA

No

INFIT

No

CNS

No (for non-profit organizations only)

CYANA

No

HADDOCK

No (for non-profit organizations only)

XPLOR-NIH

No

AMBER

No

CS-ROSETTA

No (for non-profit organizations only)

GROMACS

No, Gromacs is open source and distributed under GPL

MAPPER

No

MDDNMR

No

3.5. GLITE REQUIREMENTS

3.5.1. Which grid elements other than CE/WNs your application make use of? (e.g. WMS, SE, LFC, AMGA, HYDRA, MYPROXY, CREAM)

PROSA

UI, WMS

INFIT

UI, WMS

CNS

UI, WMS, SE, LFC

CYANA

UI, WMS

HADDOCK

UI, WMS, SE, LFC

XPLOR-NIH

UI, WMS, LB, SE, LFC

AMBER

UI, WMS, LB, SE, LFC

CS-ROSETTA

UI, WMS, SE, LFC

GROMACS

UI, WMS

MAPPER

UI, WMS

MDDNMR

UI, WMS

3.5.2. Does the application use standard Data management functionality? (i.e. file registration into LFC file catalogue, file access through LFN, file transfer/replication through standard lcg-utils)

PROSA

No

INFIT

No

CNS

Yes

CYANA

No

HADDOCK

Yes

XPLOR-NIH

Yes

AMBER

Yes

CS-ROSETTA

Yes

GROMACS

No

MAPPER

No

MDDNMR

No

3.5.3. How many files per run are registered with the LFC file catalogue?

PROSA

None

INFIT

None

CNS

In UI mode (the user submits manually the jobs), typically only one.

CYANA

None

HADDOCK

For the complete HADDOCK workflow, between 300 (minimum) and 2200 (maximum) files are registered with the LFC catalogues, depending on the user access level run parameters settings.

XPLOR-NIH

Normally one tar that contain all output directory

AMBER

Normally one tar that contain all output directory

CS-ROSETTA

Depending on the user access level, between 500 (minimum) and 2500 (maximum) files are registered with the LFC catalogues.

GROMACS

None

MAPPER

None

MDDNMR

None

3.5.4. Does the application use advanced functionality for Data Encryption? (EDS / Hydra servers)

None of the applications considered makes use of such functionality.

3.5.5. Are the normal users submitting the jobs from a UI or through a web portal?

PROSA

The use of a web portal is planned

INFIT

The use of a web portal is planned

CNS

End users submit grid jobs manually from a UI. CNS is also the computational engine beyond HADDOCK. HADDOCK launches CNS jobs via a web portal (see www.enmr.eu/webportal/index.html).

CYANA

The use of a web portal is planned

HADDOCK

Web portal (see www.enmr.eu/webportal/index.html).

CS-ROSETTA

Web portal (see www.enmr.eu/webportal/index.html).

XPLOR-NIH

UI, Web portal (see www.enmr.eu/webportal/index.html, requires a personal certificate installed in the browser)

AMBER

UI, and the refinement of the XPLOR-NIH calculated structure using the Web portal (see www.enmr.eu/webportal/index.html, requires a personal certificate installed in the browser)

GROMACS

From the UI, web interface is under development

MAPPER

The use of a web portal is planned

MDDNMR

From the UI

3.5.6. Does the application use WS, Java or C/C++ APIs for job submission and/or data management, or simply glite-wms-*, lcg-*and lfc-* command line tools?

All applications make use of the command line tools.

3.5.7. Does the application make use of Collection, Parametric or DAG job type functionalities?

PROSA

No

INFIT

No

CNS

No

CYANA

No

HADDOCK

No

XPLOR-NIH

No but in development.

AMBER

No

CS-ROSETTA

Not at this time but will most likely use parametric jobs in the future

GROMACS

No, but the development of parametric Gromacs jobs is planned in the second phase

MAPPER

No

MDDNMR

No

3.5.8. Did you find any limitation on gLite current functionalities?

All applications answered no.

3.5.9. Did you find any limitation on gLite current performances?

PROSA

No

INFIT

No

CNS

The performance relies very much on the reliability of the GRID (disappearing or failing jobs cause severe delays and reduction of performance).

CYANA

No

HADDOCK

The performance relies very much on the reliability of the grid (disappearing or failing jobs cause severe delays and reduction of performance).

XPLOR-NIH

No

AMBER

No

CS-ROSETTA

The performance relies very much on the reliability of the grid (disappearing or failing jobs cause severe delays and reduction of performance).

GROMACS

No

MAPPER

No

MDDNMR

Yes, slow job submission

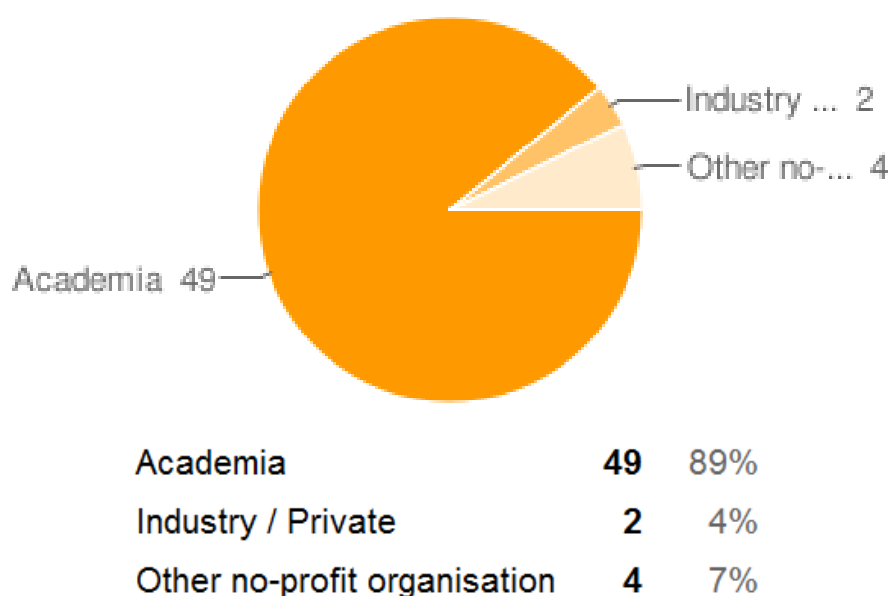
3.5.10. Did you expect any functionality from gLite that was completely missing in the current release, and which could be of benefit for your application?

All applications answered no, at the current stage.

4. Survey about data security issues

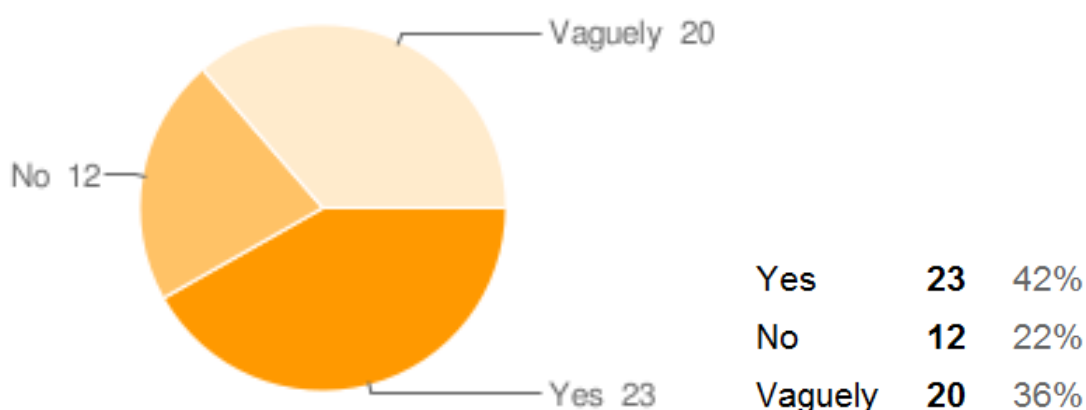
An anonymous on-line survey was proposed to the registered HADDOCK user community (~900 users counting both the registered groups and haddock portal users). These cover a wide range of techniques within the structural biology community. In addition, the survey announcement was posted to the NMR mailing list (nmr@listes.sc.univ-paris-diderot.fr) which counts more than 1430 registered members to date. The survey was also sent to a number of biotech and pharmaceutical companies.

4.1. Do you work for?

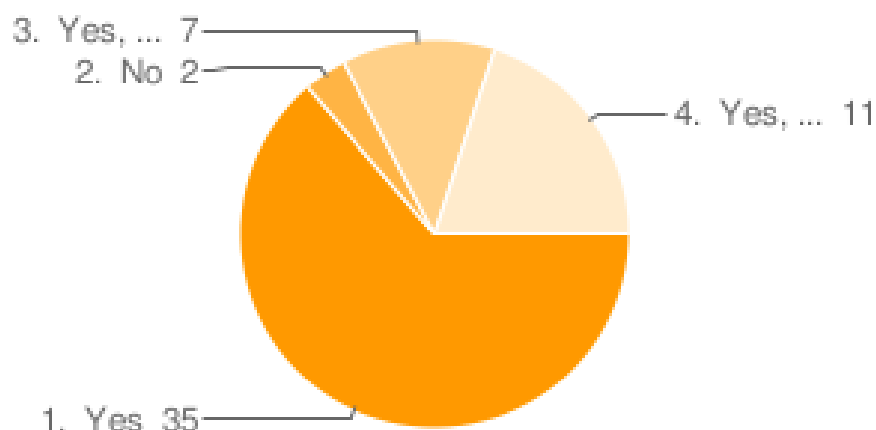


It was allowed optionally to specify Name, e-mail address and Organisation.

4.2. Do you know what Grid Infrastructures are?

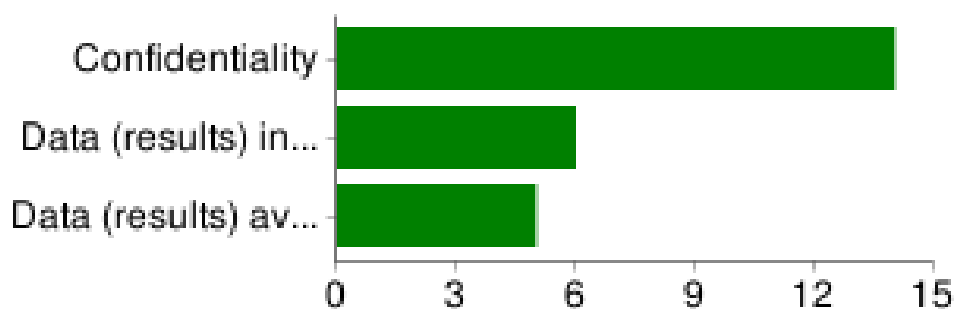


4.3. Do the policies of your organization allow you to send your data over the internet to be used as input for calculations by web servers?



1. Yes	35	64%
2. No	2	4%
3. Yes, but only non-sensitive data	7	13%
4. Yes, but only if there are stringent, documented measures to guarantee data / information security	11	20%

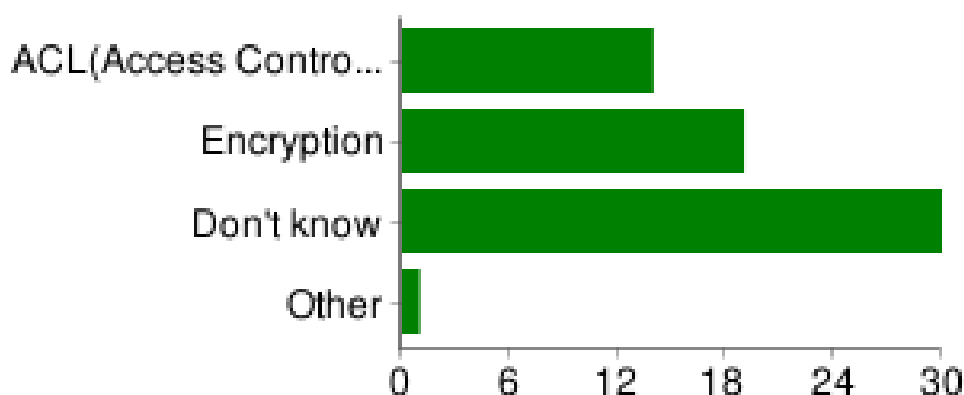
4.4. If your answer to the preceding question was 4, what is your major concern?



Confidentiality	14	88%
Data (results) integrity	6	38%
Data (results) availability	5	31%

People may select more than one checkbox, so percentages may add up to more than 100%.

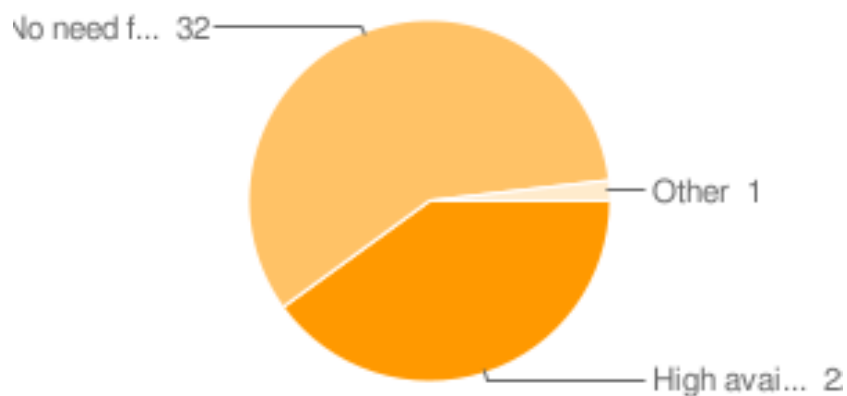
4.5. What kind of measures would you need to satisfy your confidentiality requirements?



ACL (Access Control List)-based access	14	25%
Encryption	19	35%
Don't know	30	55%
Other	1	2%

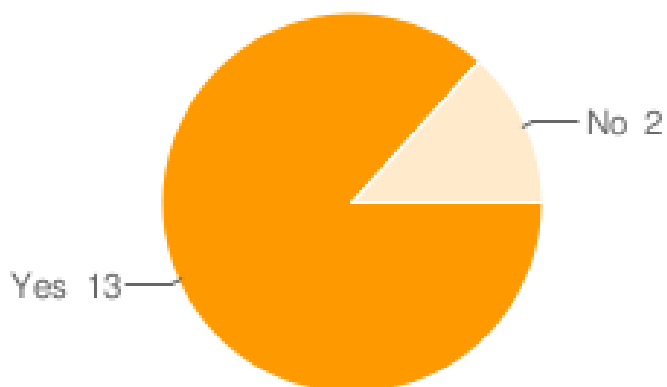
People may select more than one checkbox, so percentages may add up to more than 100%.

4.6. What kind of measures would you need to satisfy your data availability requirements?



High availability (data available via the web anytime)	22	40%
No need for high availability but for predictable, dependable availability	32	58%
Other	1	2%

4.7. If your answer to 4.3 was no, would you be interested in obtaining a version of the e-NMR platform for local installation?



Yes	13	87%
No	2	13%

5. Summary

A lot of effort has been dedicated in the past, since the beginning of the first phase of EGEE, in providing technical input to grid middleware developers from the analysis of applications requirements and use-cases. It is strongly believed in fact that grid technology should evolve in response to the requirements of its user communities.

In particular, being the Life Sciences sector one of the most important consumers of EGEE grid services, after High Energy Physics, the gLite middleware has been carefully analysed with respect to specific requirements coming from applications of the bioinformatics domain, which study genes, proteins, and all components of living organisms. These include enabling system biology on grid, oncology study at the molecular level, genome wide association studies of human complex diseases, binding of protein and DNA in the cell nucleus, complete genome comparison, as well as portals or web services that enable grid access for users in areas such as protein sequence or genome level analysis.

Such domain is clearly very similar to the bio-NMR one, so we actually did not expect to have any particular additional requirement for gLite other the ones already well detailed in many documents produced in the past by the EGEE Life Science cluster.

The most recent lists of requirements from various user communities are listed here:

- a) http://knowledge.eu-egi.eu/knowledge/index.php/User_Requirements
- b) <https://twiki.cern.ch/twiki/bin/view/EGEE/EGEEIIIPriorityList>
- c) <https://savannah.cern.ch/support/?group=egceptf>

The link a) in particular has been set up in the context of EGI_DS project, with the goal of collecting requirements from a variety of user communities including the Life Science one. It was intended as a first input to the Operation and User Requirements Working Group created in order to prepare the UMD definition. UMD stands for Unified Middleware Distribution [R4], the proposed approach of handling middleware maintenance, integration, testing, and deployment within the EGI and NGI infrastructure. It defines components, processes, involved parties etc. in order to guarantee the infrastructure to get reliable middleware in terms of both functionality and quality. The above working group collects preliminary wishes from developers, operations and users on missing functionalities, defines a preliminary roadmap for the future UMD evolution and comes up with concrete ideas for future UMD developments

It is interesting to notice the five requirements which have the highest priority for any middleware, as shown at the link of bullet a):

- | | |
|--|-----|
| 1. Reliable grid middleware; Quality of Service, especially when submitting a huge number of jobs. | All |
| 2. Easy access to data and databases; easy to use, standardized API and fine grained access policies possible. | All |
| 3. Advance reservation or the information when your job will be scheduled: monitoring of jobs, estimate of queue delay, near- real time reservation. | All |

4. Common ways of authentication and authorization: Standardized Authentication and authorization mechanisms for usage in portals as well as in direct Grid access. A globally accepted, trustworthy Grid user identity or 'Passport' with an infrastructure providing it is part of this.	All
5. A support to international VOs which are in many NGIs each with its own preference for grid services.	All

Requirements 1 and 3 arose clearly from both surveys (items 3.5.9 and 4.6), while requirement 4 is also relevant for e-NMR platform which make heavily use of web portals to hide the grid complexity to their end-users.

At the same link a) is listed another important requirement, pointed out by Life Science community, which partially emerged also from our survey (items 4.3 - 4.5):

• Encryption / protection of data on grid storage elements	LS
--	----

Actually EGEE already provides some solutions to cope with this requirement, such as the Encrypted Data Storage system based on Hydra keystores. However, the current applications forming the e-NMR platform do not require yet such functionality (items 3.2.5 and 3.5.4), so it was not possible to evaluate it in this context. Furthermore, it should be noticed that the large majority of our sample surveyed about data security issues is not limited by local policies in sending around their data over internet to be used as input for calculations via web servers.

EGEE is also providing solutions (GENIUS, P-GRADE) for running grid jobs via web portals using personal X509 certificates and VOMS for authentication and authorization. These solutions were evaluated by the e-NMR developers during the first year of the project. However, due to the existence of pre-existing applications portals already used by the community on the local resource, it was preferred to develop an in-house solution based initially on the use of the X509 personal certificates for accessing the web portal, and of robot certificates for grid job submission. These latter are fully supported and compliant with the rules defined by the international Joint Security Policy Group (JSPG) [R5]. The e-NMR developers' team is currently working on an alternate solution based on the use of personal certificates also for grid job submission.