EUROPEAN MIDDLEWARE INITIATIVE

DNA3.1.1 - COLLABORATIONS, EXPLOITATION AND SUSTAINABILITY PLAN

EU DELIVERABLE: D6.1.1

Document identifier:	**EMI-DNA3.1.1-1450889-Collaborations_Exploitation_and_Sustainability_Plan_M24_Public-v1.1**
Date:	**30/04/2012**
Activity:	**NA3 – Sustainability and Long-Term Strategies**
Lead Partner:	**NIIF**
Document status:	**Final**
Document link:	**https://cdsweb.cern.ch/record/1450889?ln=en**

Abstract:
This document is a 12-month follow-up on the initial Exploitation and Sustainability Plan, DNA2.4.2, and a 16-month follow-up of the Initial Collaboration Programs DNA2.1.1. In this document we described the current definition of the sustainability and exploitation plans, summarize the progress made in its implementation and give an outline of the work planned for the final year of the project.

I. DELIVERY SLIP

	Name	Partner/Activity	Date
From	Peter Stefan	NIIF/NA3	24 May 2012
Reviewed by	Alberto Di Meglio Florida Estrella Ludek Matyska	CERN/NA1 CERN/NA1 CESNET/CB	25 May 2012
Approved by	PEB		26/05/2012

II. DOCUMENT LOG

Issue	Date	Comment	Author/Partner
0.7	18/05/2012	First draft	Peter Stefan/NIIFI
0.8-0.16	22-25/05/2012	Iterative revisions	Peter Stefan/NIIFI Alberto Di Meglio/CERN Florida Estrella/CERN
1.0	26/05/2012	Final version for release	PEB
1.1	01/08/2012	New minor release	Alberto Di Meglio

III. DOCUMENT CHANGE RECORD

Issue	Item	Reason for Change
1.1	Document title and properties updated	The deliverable number was incorrect

IV. DOCUMENT AMENDMENT PROCEDURE

This document can be amended by the NA3 Exploitation, Sustainability and Collaborations task leader (task NA3.2), or people appointed by him/her to this task further to any feedback from other teams or people. Minor changes, such as spelling corrections, content formatting or minor text re-organization not affecting the content and meaning of the document can be applied by the NA3.2 task leader without peer review. Other changes must be submitted to peer review and to the EMI PEB and TCB for approval.

When the document is modified for any reason, its version number shall be incremented accordingly. The document version number shall follow the standard EMI conventions for document versioning. The document shall be maintained in the CERN CDS repository and be made accessible through the OpenAIRE portal.

V. GLOSSARY

Acronym	Long name or definition
AMGA	Metadata Catalogue Service
APEL Client	APEL is an accounting tool that collects accounting data from sites participating in the EGI and WLCG infrastructures as well as from sites belonging to other Grid organizations that are collaborating with EGI
API	Application programming interface
AREX	ARC Resource-coupled Execution Service
Argus EES	Execution Environment Service
BDII	Berkeley Database Information Index
BES	Basic Execution Service
BLAH	BLAH grid service
CEMON	The CEMON service provides a common interface for publishing information about a computing element in your network
CREAM	Computing Resource Execution And Management
CREAM CE	Computing Resource Execution And Management Computing Element
dCache	System for storing and retrieving huge amounts of data, distributed among a large number of heterogeneous server nodes, under a single virtual file system tree with a variety of standard access methods (http://www.dcache.org/)
DCI	Distributed Computing Infrastructure
DECIDE	Diagnostic Enhancement of Confidence by an International Distributed Environment (https://www.eu-decide.eu)
DelC	Danish eInfrastructure Collaboration
Debian	Debian is an operating system and a distribution of Free Software
DGAS	Distributed Grid Accounting System
DPM	Debian Package Manager
EDGI	European Desktop Grid Infrastructure
EGeE	Enabling Grids for E-sciencE (http://www.eu-egee.org/)
EGI	European Grid Infrastructure (http://www.egi.eu)
EGI-InSPIRE	Integrated Sustainable Pan-European Infrastructure for Researchers in Europe (http://www.egi.eu/projects/egi-inspire/)
EMI	European Middleware Initiative (http://www.eu-emi.eu)
EMI GGUS	European Middleware Initiative bug reporting system (http://www.ggus.org)
EPEL	Extra Packages for Enterprise Linux (Fedora)
ERINA+	ERINA+ is EU funded project developing and promoting a self assessment methodology for a systematic socio-economic analysis of research projects

	(http://www.erinaplus.eu/)
ESFRI	European Strategy Forum on Research and Infrastructures
ETICS	eInfrastructure for Testing, integration and Configuration of Software
Fedora	The Fedora project is a Red Hat sponsored and community-supported open source collaboration project with volunteers from around the world
FGI	Finnish Grid Infrastructure
gCube	gCube is a framework dedicated to scientists. It enables the declarative and interactive creation of transient Virtual Research Environments that aggregate and deploy on-demand content resources and application services by exploiting computational and storage resources of a grid infrastructure
GÉANT	Gigabit European Advanced Network Technology
gLite	The next generation middleware for grid computing born from the collaborative efforts of more than 80 people in 12 different academic and industrial research centers as part of the EGEE Project (http://glite.web.cern.ch/glite/)
Globus	The Globus Alliance is an international collaboration that conducts research and development to create fundamental Grid technologies (http://www.globus.org)
GLUE2	Registry Information Model and Data Model (http://confluence.globus.org/display/IIS/GLUE2+Design)
gridFTP	GridFTP is an extension of the standard File Transfer Protocol (FTP) use with Grid computing.
HEP	High Energy Physics
HPC-BP	HPC Basic Profile
Hydra	An EMI service for storage encryption/decryption
IaaS	Infrastructure as a Service
IGE	Initiative for Globus in Europe (http://www.ige-project.eu/)
ISCSI	Internet Small Computer System Interface
iSGTW	International Science Grid The Week
JDSL	Java Data Structures Library
KnowARC	Grid-enabled Know-how Sharing Technology Based on ARC Services and Open Standards (http://www.knowarc.eu)
LB	Logging and Bookkeeping Service
LCG CE	Large Hardon Collider Grid Computing Element
LFC	LGC File Catalogue
LHC	Large Hadron Collider
MPI	Message Passing Interface
MoU	Memorandum of Understanding
NDGF	Nordic Data Grid Facility
NelC	Nordic eInfrastructure Collaboration

NGI	Next Generation Internet
NREN	National Research and Education Network
OLA	Operational Level Agreement
OpenNebula	OpenNebula cloud management system (www.opennebula.org)
OS	Operating System
OSG	Open Science Grid
OSIRIS	Optimized Strategies for Risk Assessment of Industrial Chemicals through Integration of Non-Test and Test Information (http://www.osiris.ufz.de/)
PGI	Production Grid Infrastructure work group in Open Grid Forum
PRACE	Partnership for Advanced Computing in Europe (http://www.prace-project.eu/)
RedHat	RedHat Linux Distribution
SAML	Security Assertion Markup Language
SLA	Service Level Agreement
SME	Small and medium enterprise
StoRM	Storage Resource Manager (http://storm.forge.cnaf.infn.it/)
Ubuntu	A super-fast, easy to use and free operating system
UMD	Universal Middleware Distribution by EGI
UNICORE	UNiform Interface to COmputing REsources (http://www.unicore.eu/)
UNICORE CE	UNiform Interface to COmputing Resources Computing Element
UNICORE TSI	UNiform Interface to COmputing Resources Target System Interface
UNICORE WS	UNiform Interface to COmputing Resources
UNICORE XUUDB	The UNICORE XUUDB is used to map user credentials (such as an X.509 certificate or X.500 distinguished name) to a set of attributes
VDT	Virtual Data Toolkit
VOMS	Virtual Organization Membership Service
WLCG	Worldwide LHC Computing Grid (http://cern.ch/lcg)
WMS	Workload Management System
WP	Work Package
XSEDE	Extreme Science and Engineering Environment

VI. COPYRIGHT NOTICE

TABLE OF CONTENTS

1 INTRODUCTION 9
1.1 EXECUTIVE SUMMARY 9
1.2 PURPOSE AND SCOPE 10
1.3 DOCUMENT ORGANISATION 10
2 EMI EXPLOITATION AND SUSTAINABILITY STRATEGY 12
2.1 OVERVIEW 12
2.2 MARKET ANALYSIS MODEL 12
2.3 WORKS WITH EMI 14
2.4 THE EXPLOITATION PLAN AND ACTIONS 15
2.5 THE SUSTAINABILITY PLAN AND ACTIONS 16
3 THE RESEARCH GRID AND HPC MARKET 19
3.1 MARKET DRIVERS 19
3.2 TECHNOLOGY 20
3.2.1 Collaborations 20
3.3 CHANNEL 22
3.3.1 Collaborations 22
3.4 INFRASTRUCTURE 27
3.4.1 Collaborations 28
3.5 END-USERS COMMUNITIES 30
3.5.1 Collaborations 30
3.6 EXPLOITATION 32
3.6.1 Exploitable Items 32
3.6.2 Licensing 34
3.6.3 Exploitation results in Research Grids 34
3.6.4 Exploitation results in HPC 36
3.7 SUSTAINABILITY 38
3.7.1 Interoperability and standardization 39
3.7.2 Competitors analysis 40
3.7.3 Long-Term Support Plans 42
4 THE COMMERCIAL DISTRIBUTED SERVICES MARKET 48
4.1 MARKET DRIVERS 48
4.2 TECHNOLOGY 50
4.2.1 Collaborations 50
4.3 CHANNEL 50
4.4 INFRASTRUCTURE 50
4.5 COMMUNITIES 52
5 THE OPEN SCIENCE MARKET 53
5.1 MARKET DRIVERS 53
5.2 TECHNOLOGY 55
5.3 CHANNEL 55
5.4 INFRASTRUCTURE 56
5.5 COMMUNITIES 57
6 OTHER COLLABORATIONS 58
7 ACTION ITEMS FOR THE NEXT PERIOD 59

7.1 EXPLOITATION ACTIONS 59
7.2 SUSTAINABILITY ACTIONS 59
8 CONCLUSIONS 62
9 REFERENCES 63
APPENDIX A: THE SCIENCESOFT OVERVIEW DOCUMENT 65

1 INTRODUCTION

1.1 EXECUTIVE SUMMARY

The activities of Exploitation and Sustainability (E&S) of the EMI software are of paramount importance for the success of the project. Exploitation represents the measure of how successfully the software products are used not only by EMI users, grid sites administrators, application developers, NGIs managers, but also by the project partners themselves in their future activities. Sustainability gives the measure of how well the EMI project has managed to create, support and develop existing and new markets through a proper development of its traditional channels and the exploration of new channels. The creation of value and the continuous innovation of the EMI ideas and products are the ultimate goals that drive the plans.

After the initial explorations in year 1 of the concepts of exploitation and sustainability and the definition of a generic set of drivers, it became clear that a number of changes in the project structure and in the focus of some of its activities were needed. Similar recommendations were the result of the first year review. The creation of a dedicated Workpackage (NA3 – Sustainability and Long-Term Strategies), the definition of a new project-level Strategic Director and the redefinition of core deliverables were the result of this new understanding.

On the wave of the changes, the E&S team in the second year has actively worked on the definition of an ambitious, but realistic Exploitation and Sustainability Plan, drawing ideas and support from existing and new collaborations with other projects, user communities, commercial companies and from existing open source community models.

The EMI E&S strategy finds its roots in the core business of providing the best possible support for European research grids. The relationship with EGI, PRACE, WLCG, application developers and user projects in this domain have been expanded and strengthened by establishing formal collaboration agreements and technical cooperation. Important parameters as the cost of supporting and developing the EMI products, the core business plans of the EMI partners, the available and expected funding sources have been analysed and discussed. Long-terms support plans have been defined and shared with major stakeholders.

The core market represented by the European infrastructures and the researchers using them must be preserved and kept efficient and operational. These activities require as expected most of the EMI effort, energy and funds.

At the same time, the fact that the research grid market is a niche, mature market with a relatively moderate growth must be acknowledged. The EMI E&S activities have therefore been also focused in this second year in understanding how innovation can be better sustained and what new markets could provide the highest chances of returns on investing the available limited effort. Based on the ideas already explored in the first year, the EMI Sustainability team in NA1 and NA3 have decided to focus on two distinct areas, namely the implementation of at least one good example of commercial exploitation of some of the most promising EMI services and the creation of wider-scope open source initiative for science with the potential of serving the longer-term goals of the European Commission Open Science strategies.

The first objective, the commercial exploitation, has resulted in the collaboration with dCore Systems, a Luxembourg-based holding of a number of high-tech SMEs. The company (previously known as TheSyrrus) has an ambitious business plan based on the concept of commercial distributed services that fits very well the distributed nature of the EMI products.

The second objective, the creation of wider-scope open source initiatives, has resulted in the creation of ScienceSoft, a vertical open source community dedicated to software for scientific application. The

ideas behind ScienceSoft have been widely discussed and presented to as large an audience as possible at this stage. A plan for its further implementation has been agreed and it's being followed.

This document describes the EMI Exploitation and Sustainability activities that have led to the current state of the art and how the different parts fit together in a broad-scoped strategy that will guide EMI for its third year and the EMI partners for the next two to three years.

1.2 PURPOSE AND SCOPE

The purpose of this document is to provide a detailed overview of the current status of the EMI Exploitation and Sustainability plans and of the collaborations that support its implementation. It applies to the activities performed by the NA3 Workpackage after its definition in May 2011. Some of the actions were already defined in deliverable DNA2.4.2 submitted this time last year. Although using a revised format, this document should be considered as the ideal continuation of the previous sustainability and exploitation deliverable, taking into account the structural changes in the project. In addition to the E&S activities, also the critical areas of interoperability and standardization were moved into the NA3 Workpackage, since they are recognised to serve an important role in the implementation of the E&S strategy and are therefore described in this document. This document can be therefore considered an evolution of the following documents:

DNA2.4.2 - Exploitation and Sustainability Plan
(http://cdsweb.cern.ch/record/1277605/files/EMI-D2.4.2-1277605-Exploitation_and_Sustainability_Plan_M12-v1.0.pdf)

DJRA1.6.2 - Integration Work Plan and Status Report
(http://cdsweb.cern.ch/record/1277594/files/EMI-DJRA1.6.2-1277594-Integration_Work_Plan_M12-v1.0.pdf)

DJRA1.5.1 – Standardization Work Plan and Status Report
(http://cdsweb.cern.ch/record/1277526/files/EMI-DJRA1.5.1-1277526-Standardization_Workplan-v1.0.pdf)

DNA2.1.1 - Collaboration Programs
(http://cdsweb.cern.ch/record/1277553/files/EMI-DNA2.1.1-1277553-Collaboration_Programs-v1.0.pdf)

1.3 DOCUMENT ORGANISATION

The document follows slightly different structure than the previous versions of E&S studies. Besides the standard E&S follow-up a new specific sections describing the EMI collaborations with other project and initiatives and their impact on the implementation of the plan have been added.

Section 2 describes the generic E&S strategy as it was originally defined in deliverables DNA2.4.1, and DNA2.4.2, revised according to the progress made in year 2. The former action items, as well as their correspondence to the year 2 activities are also highlighted here in tabular form.

In section 3 the relationship between the EMI R&D activities and other projects are described. Their roles and added value in the defined market operation models are analysed. This section and also the next two sections are structured in a similar way guiding the reader through the main market drivers, the possible user community outreach channels, the infrastructure enabling the progression from the pure technological level to the end-user communities. This section gives a thorough description of the formal collaboration relationships that are particularly relevant for EMI, the main achievements, the exploitation results and the specific sustainability plans.

Section 4 is dedicated to industry-related efforts showing the most important industrial collaborations that have been established in the last year, such as Technicolor, DCore, and SysFera joint efforts.

While reaching out for new industrial user communities is important, it is of great strategic importance to grow and reinforce the relationships with new scientific communities beyond the traditional grid user base. Section 5 introduces a new initiative called ScienceSoft that is meant to emphasize the most common communication, exploitation, and sustainability activities of different scientific R&D projects, trying to bring the shared efforts under a common umbrella, and facilitating them with shared knowledge and infrastructure as a possible building block of the future EC en Science initiatives.

Section 6 contains those collaboration initiatives that have been created to support EMI in its exploitation and sustainability activities, such as the collaboration with e-ScienceTalk, ERINA+ or OSIRIS.

Section 7 defines action items for the next period as a concrete set to be used to guide the work of the NA3.2 team in the third year.

Section 8 then concludes and summarizes the document.

2 EMI EXPLOITATION AND SUSTAINABILITY STRATEGY

2.1 OVERVIEW

The EMI Exploitation and Sustainability strategy consist in understanding the existing and potential markets where the EMI products can be successfully used to produce innovation and enable scientific research.

The strategy is implemented through a series of steps:

1) Understanding of the markets (market analysis)
2) Competitors analysis
3) Definition of the most profitable collaborations within the market
4) Dissemination of the EMI values to the market
5) Monitoring of results and corrective actions

This document describes how the 5 steps have been applied in practice during the second year, which results have been obtained and what actions are proposed for the third and final year of the project.

2.2 MARKET ANALYSIS MODEL

The EMI Exploitation and Sustainability strategy requires an understanding of the EMI markets and users. As part of the implementation of the strategy, a market analysis has been performed using a model composed of as a set of stacked layers, each one with associated responsibility and scope:

1) **Technology**: This layer represents EMI and its products and services and other technology providers. Within the scope of the EMI plan this is essentially the set of middleware services that provide distributed compute and data functionality
2) **Channel**: this is the means that EMI uses to distribute its product to its users. It can be considered a distribution channel made of entities providing entry points to the infrastructures where the services are deployed
3) **Infrastructure**: this is where the EMI services are deployed or directly used. This layer is made by the projects, initiatives and people deploying, managing and monitoring the services and the resources where the services are deployed. It includes also the communities of grid application developers directly using the EMI APIs to produce the high-level services used by end-users. This layer is the actual reference market segment for EMI. The people in this layer are what we define **EMI users** in this document
4) **End-users communities**: this is the layer made of the professionals, researchers, scientists doing work using the infrastructure services. This layer is composed of several different types of users, from high-level domain-specific application developers, to the scientists having (and wanting to have) little or no knowledge of how things work under the hood
5) **Market domain**: this is a specific market defined in terms of its scope, composition, mission, requirements, etc. and structured by the above-described layers

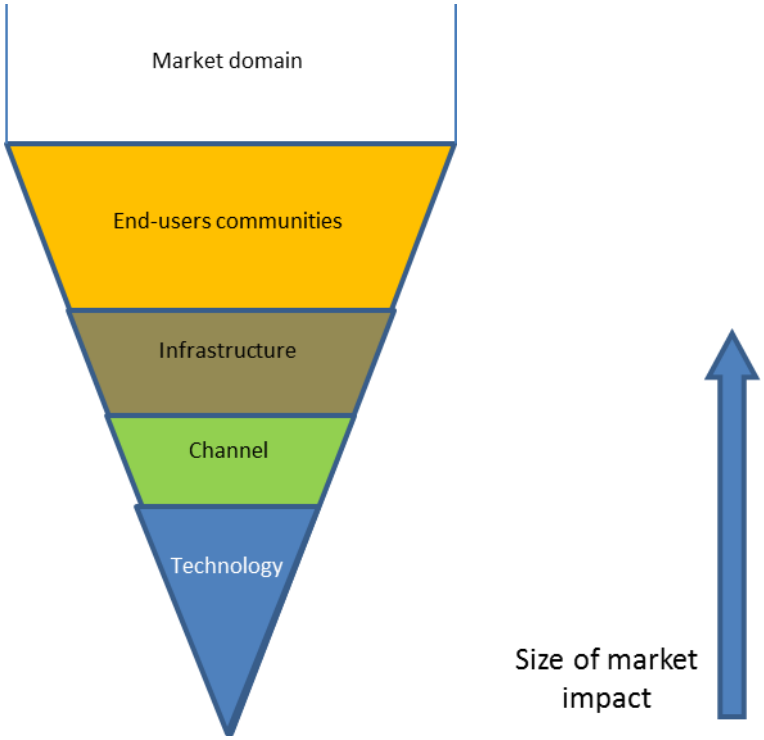

Figure 1: The EMI E&S layered stack concept

An important concept in the proposed EMI layered stack concept is the amplification factor (the size of market impact) that each layer provides for the layer above. The actual entities in each layer must be selected to provide an added value so that a relatively small improvement in the technology layer can be turned in a large improvement in the domain market size. This is something that requires a realistic verification of the assumptions and continuous collaboration among the different entities.

The current and potential EMI markets and the work done in terms of creating, maintaining and extending collaborations is described in the rest of this document. Three market domains have been identified and exploited to different extents. The market domains are:

1) Research Grids and High-Performance Computing
2) Commercial distributed services
3) Open Science

Given the above-mentioned definition of a Market domain, the concepts of Exploitation and Sustainability can be specified as:

Exploitation is defined as the use of the EMI products and services by the layers above the Technology layers and by the EMI partners themselves as part of their core business. Each layer has different exploitation characteristics and different expected and actual results based of EMI Exploitation plan.

Sustainability is defined as the capacity of producing continuous innovation, enabling the layers above to provide increasingly better services and to be willing to pay directly for that or to support and

require the EMI partners to seek additional funding in the future. Also in this case the values that each layer perceive as critical are different and so are the expected and actual results of the EMI Exploitation plan.

In the rest of this document, the role of each market in the overall strategy and the work done in implementing it are described together with the main collaboration, exploitation and sustainability activities. All actions and milestones agreed with collaborating programs are recorded and monitored using a dedicated tracker at:

https://savannah.cern.ch/task/?group=emi-mou

The results obtained so far for each one of the market layers are described. The majority of the entities that EMI is collaborating with, projects, initiatives, Institutes, companies, etc., fit in one of the above-defined layers. However, in a number of cases a clear one-to-one match is not possible since an entity may perform activities in more than one layer and a sharp distinction of responsibilities across layers is not possible. In this case the core activity of that entity is used to map it to one of the layers.

2.3 WORKS WITH EMI

The "Works with EMI" program started in the first year has been successfully continued in the second yea. WWE is a managed collaboration program dedicated to software providers who develop applications using the EMI middleware services, or who develop additional middleware or distributed computing services extending or complementing the functionality of the EMI services.

Figure 2. The Works with EMI logo

The collaboration is based on the establishment of a Memorandum of Understanding between EMI and the technical partner joining the program. The MoU describes the mutual benefits for EMI and the partner in entering the collaboration and the content, scope and duration of such collaboration. In the specific case of a Works with EMI collaboration, the MoU will include specific clauses regarding

1. Access to the ETICS Production Service
2. Preview of EMI Releases
3. EMI Release exploitation
4. Partner feedback

During the second year the program has been applied to the collaborations with EDGI, iMarine and IGE. In the case of EDGI, it has resulted in the release of the EDGI Desktop Grid Bridges as part of the EMI 2 release (see also section 3.2.1).

2.4 THE EXPLOITATION PLAN AND ACTIONS

This section summarizes the results of the action items on Exploitation defined in DNA2.4.2 for the second project period and how they have evolved in the past 12 months under the new NA3.2 activity:

#	ACTION	RESULT
E1	*Improving of the currently existing team and ensuring that sufficient amount of time is spent on increasing exploitation are the highest priority action items in the next period. The expected outcome is to have a significant amount of influential people involved in the group and to boost exploitation efforts further, making them continuous during the entire lifetime of the project, instead of slow irregular fractions of workload. It is also a key outcome here to incorporate significant role-players even from outside the project. A reasonable deadline schedule for reaching all of these is September 2011.*	A revised Exploitation and Sustainability team has been established including partner representatives. Regular bi-weekly meetings have been organized to coordinate the activities of the group and follow up on the action items defined within the group and during the project face-to-face meeting
E2	*It is planned to establish a working group developing a proposal for a common coordinating body after the EMI end. The expected outcome of this action is a brief plan, which is expected to be released in September 2011, on how to assemble this group, as well as the setting up of this team, which should be in place by December 2011*	A working group within the NA3.2 task has been created with the overall responsibility to manage E&S related activities in the project. The task of developing a proposal for an EMI collaboration after the end of the EMI project has been directly undertaken by the Collaboration Board with the mediation of the Project Director. A number of discussions have been held during the All-Hands meetings and CB meetings in Padova in October 2011 and Hamburg in May 2012.
E3	*Either a group will be formulated or a person will be appointed being in charge of reaching out to new user communities beyond the audience identified in this document. The expected outcome of this action is the existence of a responsible entity that currently is not the case, as no such assignments are present in the project. This should be completed by September 2011. This group or person is expected to study the needs of newly discovered user groups and to work out a user base enlargement strategy by*	The task has been undertaken by the Project Director and the Strategic Director. The outcome of the work, the industrial collaborations established is described in section 4 of this document.

	December 2011.	
E4	*It has been decided to provide a schedule of dedicated meetings between NA2 and involved parties at regular EMI All-hands meetings (twice a year) to follow-up on the progress of this plan. It is a continuous task until the end of the project.*	The action item has been completed as decided. E&S activities have been moved into NA3 from NA2 as a separate task after re-structuring the project.
E5	*The expected outcomes here are brief analyses or case studies on these products revealing reasons why they can be appealing to the users. The deadline is basically the end of the next period, but as these components are rolled out gradually in the next year, they will be inspected in a month of time following the release.*	Various activities have been performed to understand which products have the best chances of receiving continuous support after the end of EMI. The major actions have been the cloud survey and a number of use cases discussions (described in section 3.7.2) and the technical investigation with DCore (described in section 4.1 and following sections). However, the difficulties encountered in getting direct user feedback on the EMI services and their actual usage have contributed to the idea of the ScienceSoft initiative (described in chapter 5).
E6	*Based on the recently concluded EMI All Hands meeting in June 2011, the following items have been identified as having potential to be exploitable in the next period. Identification of additional exploitable items is an on-going task within this activity.* *The EMI client-side component supporting multiple middleware, multiple operating systems and multiple job descriptions;* *The common MPI execution interface and the possibility to exploit MPI executions on graphical processors;* *The high availability ARC execution service;* *The uniform security service;* *The transfer from GSI to TLS.*	This action has been cancelled. One reason is that the identification of exploitable item is also the object of the previous action. The other reason is that the initial collection of exploitable items was found to be rather arbitrary and not supported by a real market analysis.

2.5 THE SUSTAINABILITY PLAN AND ACTIONS

This section summarizes the results of the action items on sustainability defined in DNA2.4.2 for the second project period and how they have evolved in the past 12 months under the new NA3.2 activity:

#	ACTION	RESULT
S1	*The current EMI 1 release contains 54 products. The project will attempt to determine the 10 most relevant services being vitally necessary to operate grid infrastructure. The expected*	This action as formulated could not be performed. The task of identifying "the 10 most relevant" service proved to be impossible, because the "critical set"

	outcome is a list of mandatory components for any reasonable HTC/HPC infrastructure services. The suggested deadline is July 2011. The action supports sustainability driver "decrease of development and maintenance costs".	varies depending on the size of communities, their expertise and their usage requirements (data-oriented, computing-oriented, both). The action has been replaced with the definition of a set of Profiles matching specific usage patterns actually used in production. The action will be completed in year 3
S2	*For each product enlisted in Section 3 of this document[1] an update on the sustainability and maintenance model will be given. The expected outcome of this action is the refined version of the table. The work will be performed continuously during the next period; the periodic deadline is April 2012. The action supports sustainability driver "decrease of development and maintenance costs".*	This has been done as part of the Product Team Effort Analysis and the Product Support Survey described in sections **Error! Reference source not found.** and 3.7.3 respectively.
S3	*The sustainability objectives need to be refined as more experience is gathered. The expected outcome of this action is the refined strategy having all amendments needed to reach the refined targets. The deadline of this is September 2011. The action supports all sustainability drivers.*	The refined strategy and its implementation are presented in this document.
S4	*As licensing is quite heterogeneous in the current EMI distribution a convergence plan should be constructed in order to leverage the various licensing schemes. This year some of the component licenses have been collected, but it is far obvious from that poll that some has to be altered by the end of project. It is a realistic estimate that by April 2012 a step forward can be made in the form of a suggestion being formulated for component license amendment. The expected outcome of this action is a study on how to make the different licensing schemes more convergent. The action supports all sustainability drivers.*	All license used by EMI products are valid OSI licences as of the release of EMI 2. The work done is described in section 3.6.2.
S5	*As EMI 1 release has come out just slightly after the first reporting period terminated, and it is expected that EMI 1 will be put into production after the release over numerous sites, it is planned for the coming period to collect usability evaluation data from among the users, to process these data and to provide a feedback on the positive and also on the negative*	Feedback has been continuously collected during the year from EGI and other users. Section 2.3 gives a nutshell summary on some mainstream user concerns. Major issues have been discussed at the EGI TCB and at the WLCG MB and GDB. However, the difficulties encountered in getting direct user feedback on the EMI

[1] It refers to deliverable DNA2.4.2.

	reactions. The expected outcome here is the list of positive and negative reactions after installing the new release. This action item can theoretically be completed by December 2011. It supports sustainability driver "extension of the user base".	services and their actual usage have contributed to the idea of the ScienceSoft initiative (described in chapter 5).
S6	*Even though that there is a cloud task force in EMI, the further elaborating of the cloud strategy laid down in DJRA1.6.2 and DJRA1.4.2. is scheduled for the next period. Staff being responsible for sustainability issues must thoroughly follow the work that will be conducted within this working group watching for new potential. This is supposed to be a continuous task in the next period, having a periodic deadline of April 2012. It supports sustainability driver "extension of the user base".*	The cloud strategy has been proposed during the second year. A Cloud Survey has been conducted by the Strategic Director through interviews with users and operators. The EMI cloud strategy and the current status of the implementation are described in section 3.7.2. A detailed description about cloud and trust federation related activities are found in deliverable DNA3.2.1.
S7	*It is also planned to submit a second questionnaire in the next period that incorporates lessons learnt from the first poll attached as appendices to this document. A suitable schedule for this would be October 2011. It supports sustainability driver "extension of the user base" by understanding the "side-effect-features" of the products more.*	This action has not been completed yet. A draft questionnaire has been designed, but its launch has been postponed several times due to several constraints, including a plan to extend the survey to include the ScienceSoft initiative, which has delayed the outcome of the task. The action will be completed in the coming months.

3 THE RESEARCH GRID AND HPC MARKET

3.1 MARKET DRIVERS

The Research Grid and HPC market is the core EMI market domain. EMI was designed since the beginning as a successor of the EGEE and KnowArc projects and as an integral part of existing UNICORE and dCache activities.

The EMI middleware products are mainly designed to work on infrastructures within this market. They all follow the traditional grid hourglass paradigm introduced in the late '90s of the last century by Ian Foster and Karl Kesselman [R1].

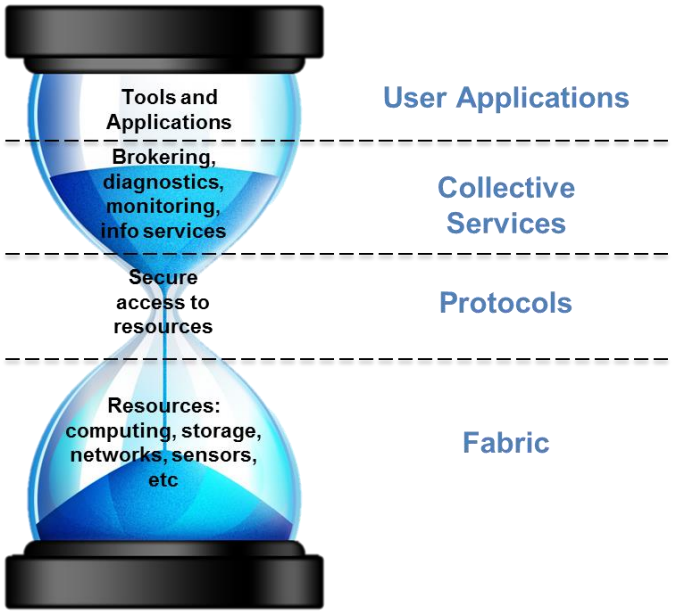

Figure 2: The Grid Hourglass model

An important aspect to be considered is that in the original grid blueprint, most of the emphasis was put on protocols and not on APIs and SDKs. This was an explicit design decision that has influenced most of the grid middleware until today. This design choice is showing today its limitations and it is in a way responsible for much of the attention on cloud, which has more emphasis on APIs than protocols. Work is being done within EMI to rationalize the APIs, but the original choice has consequences on the exploitation and sustainability of existing grid services.

Most of the EMI objectives as stated in its Description of Work are related to the maintenance, support and evolution of the existing European grid infrastructures and their interoperation with High-Performance Computing centres currently coordinated by EGI and PRACE respectively. As described in more details in the next section, these activities require no less than 80% of the EMI project budget and effort, although some of the work done in this market, especially in the area of standardization, is very important for the extension in other markets.

The exploitation and sustainability activities in this market have received most of the attention, since the preservation and continuation of the research grids operations is the main goal of the EMI project.

The composition of the market stack is shown in Figure 3.

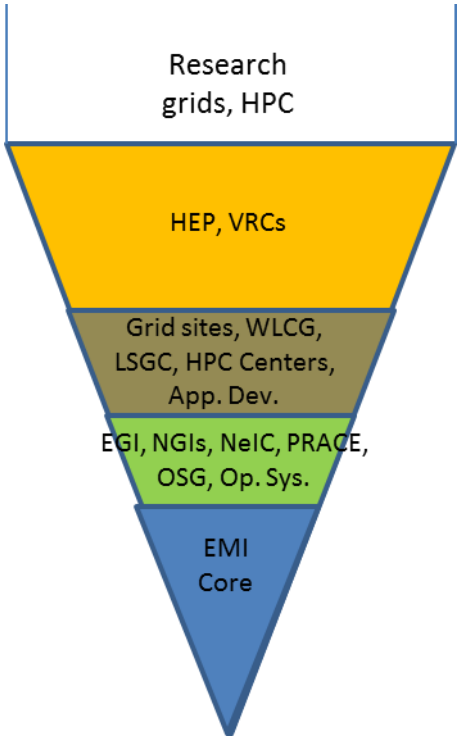

Figure 3: The Research Grid market

3.2 TECHNOLOGY

The Technology layer is composed of all the middleware services developed, maintained and supported by EMI. As of April 2012, the EMI 2 Matterhorn release candidate includes 52 different services in the four Technology Area of Compute, Data, Security and Infrastructure. Detailed information and the full list of products included in the EMI 2 Matterhorn distribution can be found on the EMI web site at:

http://www.eu-emi.eu/emi-2-matterhorn

This release is the result of the work of almost about 80 people in 29 Product Teams with global tasks such as Release Management, User Support coordination and Quality Assurance and Control provided as Work Package-level tasks. It is the main exploitable item of the second year of activity of EMI.

3.2.1 Collaborations

The main collaborations in the Technology layer are established with other middleware providers. The two main collaboration mechanisms are international coordination activities and the Work with EMI technical program. The most important collaborations continued or established during the second year have been (all actions are tracked at https://savannah.cern.ch/task/?group=emi-mou):

EDGI

The European Desktop-Grid Initiative provides software services and infrastructure coordination services. EMI and EDGI have signed an MoU [R3] in September 2011 and an OLA [R4] in May 2012 as part of the Work with EMI program. The main objectives of the collaboration are:

EMI release preview and assessment: The EDGI infrastructure is a federation of service grids, desktop grids and clouds. For the service grids it integrates software components from EMI, among others. Within the framework of the 'Works with EMI' program, EDGI has access and preview to early releases of EMI products via the EMI Testing Infrastructure. Users subscribe to the testers.eu-emi.eu VO to access the resources of this infrastructure. There is a period of testing and assessment whereby EDGI evaluates the usability and reliability of EMI products through the execution of EDGI use cases on the infrastructure, e.g. single or multiple job execution on desktop grid, status/logging information retrieval from desktop grid, job result downloading from desktop grid, etc.

Support the development of EDGI bridge middleware: EDGI develops "computing element bridge service" plugins for AREX, CREAM and UNICORE CEs that have been integrated by EMI in its EMI 2 distribution. The previous implementation of the CREAM CE plugin inherited from the EDGeS project had scalability and functionality limitations; however the current EDGI-CREAM CE plugin implementation resolves this problem. The EDGI-ARC CE plugin was already released in EMI 1 as a standard update. The EDGI-UNICORE CE plugin is expected to follow immediately after the release of EMI 2. In addition, scientific communities and application developers harnessing EDGI services may require new or improved capabilities that in turn translate to feature requests for EMI products.

EMI release exploitation: Subject to its acceptance criteria, EDGI upgrades its infrastructure to use the latest releases of the relevant EMI products from the EMI repository. EDGI also intends to use new and improved offerings from other EMI products that may include features EDGI development team expresses interest in. EMI includes the EDGI plugins into its releases, and support for the EDGI plugins distributed together with the EMI release will be provided by EDGI through the relevant EMI GGUS Third Level Support Unit(s). The support relationships, including process and time-frame for delivery of services, between the projects are described in an OLA and are used in support of the EMI SLAs. The EMI-EDGI OLA complies with the EMI service levels.

Outcome: the main outcome of this collaboration is the integration of the desktop grid bridges with the EMI Computing Elements. The bridges are now released as part of the EMI 2 distribution. An Operational Level Agreement has been signed in May 2012 for the support of the EDGI Bridges in a way compatible with the EMI SLA.

IGE

The Initiative for Globus in Europe provides software services to European Globus users. EMI and EGI have signed an MoU [R5] in March 2012 as part of the Work with EMI program. The main objectives of the collaboration are:

Globus support and maintenance: EMI relies on IGE, as provider of official Globus packages, for dedicated support from Globus developers and maintainers. EMI has moved away from the previous usage of Globus packaged by the US-based VDT program to a more standard use of the Globus packages maintained by IGE and distributed as part of operating systems distributions like Fedora/EPEL and Debian/Ubuntu. EMI provides feedback, recommendations and lessons learned to validate and improve IGE offerings.

Standardization, and Interoperability: EMI and IGE work together to make their data and compute products that provide similar functionalities interoperable through the adoption of agreed standards, e.g. PGI, SRM and GLUE2, and API harmonization where appropriate. This activity focuses on making implementations, and ensuring they remain interoperable in practice based on existing standards and those emerging from PGI, including the EMI-proposed solution to PGI on common job submission and management methods (EMI-ES). As both projects provide software to grid infrastructures through EGI, better-integrated EMI and IGE components can ease their deployment and operation.

Globus Requirements: EMI provides requirements on Globus to IGE. As consumer of a number of Globus components, EMI submits enhancements or feature requests to IGE to better support its middleware distribution and address some of its own requirements, if needed. The Standardization and Interoperability activity may also generate requirements to facilitate EMI and IGE components working together. EMI and IGE monitor the implementation of these requirements in the Globus or IGE release.

Dissemination: EMI and IGE have a coordinated dissemination strategy, raising the awareness of communities relying on the projects' technologies about new and improved offerings.

Outcome: The main outcome of this collaboration is the common work done on Job Management interface interoperability (within the OGF PGI group), on integration of Globus services with the EMI Argus authorization framework (released by IGE in the form LCMAPS callouts) and on common software packaging and releases policies for EPEL and Debian.

StratusLab and Venus-C

Collaboration with the two Cloud-related projects, StratusLab and Venus-C, was not formalized via MoUs. However, the projects kept a close relationship during the past two years to discuss how cloud may impact the existing research infrastructures and what changes if any the EMI project should introduce. Since the concept of cloud is of very limited interest for EMI and currently has little impact on production grids, no concrete work plan was defined. The only item of interest is the possible use of Argus-EES with OpenNebula to extend the ARGUS authorization framework to the management of virtual machines, which has an operational interest for cloud site managers.

Outcome: the main outcome in year 2 has been the agreement to provide virtual appliances of EMI services for the StratusLab marketplace, which is now part of the EMI plan in year 3. However, since the StratusLab project ends in June 2012, EMI needs to understand whether the StratusLab Marketplace will be supported in the future. The possibility of integrating it with the ScienceSoft Marketplace has been discussed (see chapter 5 section 5.3 for more information on ScienceSoft)

3.3 CHANNEL

The main distribution channel of the EMI 2 release and the related release, support and training activities is represented by EGI and the NGIs (via the EGI-InSPIRE project) and PRACE. EGI and PRACE provide the coordination activities that allow software products from EMI to be deployed by NGI grid site managers and HPC site on the infrastructure. They provide the main distribution channel for EMI products. Similar collaborations have been established with smaller, national or regional infrastructures, like NeIC, and international coordination projects, like OSG, CHAIN or EUIndia-Grid 2.

3.3.1 Collaborations

The main collaboration mechanisms are MoUs and SLAs. The most important collaborations continued or established during the second year have been (all actions are tracked at https://savannah.cern.ch/task/?group=emi-mou):

EGI

EMI's most prominent collaboration is with EGI. EMI signed an MoU with EGI-InSPIRE in December 2012 and an SLA in April 2011, whose application is regularly monitored and content revised. The main objectives of the collaboration are:

User and operation-centric Middleware: Evolve middleware components based on requirements gathered through EGI-InSPIRE, and EMI within the various scientific communities within EGI and expressed through the EGI Technology and UMD Roadmaps. In particular focus was put on finding a common strategy in the setup and deployment of tools for service monitoring and management. Work on this topic is part of EMI continuous service and product provisioning activities. This includes acceptance criteria verification, release transition and security assessments.

Standards: Contribute to steer, and advance high-priority standards as identified in the EGI Standards Roadmap and in the EMI Standardization Work Plan. Work on this topic is done as part of the effort both EGI and EMI put in collaborating with standardization bodies like OGF. Important standardization and convergence items like the new common EMIR service, the StAR storage accounting records, and the common EMI-ES interface were amply discussed with EGI.

Operational Support: Agree and monitor SLA for third-level user support for incidents and requests. The SLA was designed and negotiated with EGI and signed in April 2011. Both EMI and EGI have monitored its application in the provision of software products and shown that EMI is applying consistently and with very little deviations.

Dissemination: Disseminate the progress and final results of the collaboration within EGI, and EMI communities. The EMI collaboration with EGI on disseminating the values of the research grid infrastructures has been very active and it's well represented by the joint organization of the yearly EGI User Forum / EMI Technical Conference event. The first one was organized in Vilnius in March 2011 and the second one in Munich in March 2012.

Sustainability Models: In all work items supported by EGI-InSPIRE explore how EGI.eu through its participants (NGIs) can help sustain the middleware provided by EMI that it relies on. This may include developing a joint strategy for engaging business partners, having a better understanding as to who uses the current middleware technology. Sustainability of both the software and the infrastructure was one of the most discussed topics during the past 12 months. Since EMI and EGI are the first two layers of the research grid market, ensuring their sustainability is paramount to preserving the end-user communities. EMI has put great effort into understanding its sustainability options and how the partners will keep supporting their products after the end of the project. This will be described in more details in the sustainability section below.

Training: Develop complementary training strategy focusing on middleware installation, configuration and usage by end users. The collaboration results are documented via roughly 20 tracked milestones distributed over the lifetime of the EMI project. Training events have been organized at the joint EGI/EMI conferences for system administrators and application developers. Recently EGI has started collecting training requirements for their reference communities of site administrators and an targeted training days will be organized together with EMI during year 3.

EGI is the most prominent partner for EMI as the coordinator of the largest community of user of the EMI products (the NGIs). The EGI infrastructure relies heavily on the EMI provided middleware as shown in the exploitation section 3.6. EGI serves also as the primary collector of requirements on the EMI middleware. Although WLCG is the largest EGI user, EGI is working hard on attracting a broad spectrum of scientific communities and supporting their use of the distributed grid infrastructure. The collaboration with EGI runs through the Technical Coordination Board (TCB) as the focal point of these efforts, to discuss and evaluate requirements and to transfer them to technology providers.

As a part of the MoU between EGI and EMI a specific SLA for the third level support from EMI to EGI users has been put signed. This SLA covers incidents and requests handling and serves as a model for future agreements between users and resource providers and the middleware developers' community (see also the more specific deliverable on the SLA definition and implementation [R2]).

Outcome: the outcome of the collaboration with EGI represents most of the work done by EMI as part of the support for Research Grids. Among the main items we want to mention the coordination work as part of the EGI TCB, the 3-level support activities in GGUS, the application and monitoring of the SLA, the exploitation of EMI products in the UMD distribution and common work on sustainability as part of the ScienceSoft initiative.

NGIs

The NGIs are the basic building and governance blocks of EGI. They are represented in the EGI Council and are also direct formal members of the EGI InSPIRE project. There is a large overlap in the membership between EGI InSPIRE and EMI projects, as CERN, INFN, CESNET, SWITCH, CSIC, GRNET, STFC, TCD, as well as non-European partners ASGC and KISTI are participating in both projects. This overlap serves as a dual advantage for the EMI project:

1. Direct participation of NGIs in the EMI projects provides additional channel between the operational requirements and the EMI work. EMI does not rely exclusively on the information provided through the EGI-InSPIRE TCB and reflected formally in the EGI technology and UMD Roadmaps, but can use direct contacts through the NGIs both to the resource centres (where the EMI middleware is actually run and managed) and the use communities (that rely on the middleware availability, functionality and quality).
2. The NGIs are considered as stable and expectedly self-sustainable components of the EGI ecosystem. This means the experience developed at individual NGIs through the execution of the EMI project should not be lost when the EMI project finishes. For components of general interest, the NGIs are in a good position to take over the development and support of at least some of these activities even after the EMI project funding is finished (more info is given in section 3.7). Also, this approach may serve as a model for other NGIs to also participate in the middleware support, getting into closer collaboration with remaining EMI project members.

PRACE and Other HPC Projects

The collaboration with PRACE has been negotiated for several months. A draft MoU has been agreed between EMI and PRACE representatives and is currently being reviwed within PRACE by the Management Board and the PRACE Association Board. The process will take up at least six to eight more weeks due to the current meeting schedule of the two Boards. When signed, the duration of this agreement will overlap the partial duration of two consecutive PRACE projects, PRACE-1IP and PRACE-2IP. The main activities foreseen in this document are as follows:

Exploitation of EMI Components: Allow PRACE access to EMI releases so to evaluate and test its components – in particular UNICORE Gateway, UNICORE Registry, UNICORE Services Environment, UNICORE TSI, UNICORE WS, UNICORE XUUDB. Expected outcome: If no incompatibilities with PRACE systems exist (i.e. different operating systems are present within the PRACE network), PRACE will upgrade its infrastructure to the latest versions of mentioned components included in EMI 2 and EMI 3 releases through the EMI repository. Feedback, recommendations and lessons learned are provided back so to support the improvement of EMI services/products.

Evolution of UNICORE components to foster the interoperability with other e-infrastructures (e.g. XSEDE): Allow PRACE to establish and maintain interoperability with XSEDE (https://www.xsede.org/) through UNICORE open standards BES, JSDL, HPC-BP, SAML in order to prepare a common application calls among the two projects that could take advantage of interoperability. Expected outcome: PRACE supports the evolution of UNICORE components and exploits them in establishing the interoperability with XSEDE. The EMI standardization activity is reported more fully EMI's standardization work plan [R3].

Evolution of the EMI Security Token Service (STS): PRACE supports the development of the EMI Security Token Service (STS) and exploit development results. Expected outcome: PRACE evaluates and adopts the EMI Security Token Service (STS).

Evolution of EMI Components: Evolve PRACE services on the base of user requirements that are gathered through PRACE-1IP user requirements activity (WP6 Task 6.3). Further requirements collected after the end of the PRACE-1IP project will be streamlined towards EMI by the PRACE-2IP WP10. Expected outcome: PRACE supports the evolution of EMI components and contribute to their testing.

Operational Support: Agree with PRACE on a support model for deployed components. Expected outcome: provide PRACE with third-level user support for incidents and problem management. This activity entails the signing and monitoring of an SLA for third-level support on incidents and general requests from users (expected date October 2012).

Dissemination and Training: EMI will participate in PRACE's major events, such as the PRACE Day: http://www.prace-ri.eu/PRACE-Day-2012. EMI and PRACE will jointly disseminate the progress and final results of the collaboration within the PRACE and EMI communities. EMI and PRACE will develop a training strategy for services that are deployed in PRACE.

Outcome: a detailed and tracked list of common actions is not yet available since the MoU is still under review by the PRACE MB and AISBL.

NDGF and NeIC

NDGF has been a key customer for several EMI components, most notably ARC services and clients, dCache storage, as well as VOMS and FTS services. Originally driven by the goal of providing computing support for few large user communities (LHC experiments, bioinformatics, CO2 sequestration, computational chemistry), in 2012 NDGF has been transformed to a generic Nordic e-Science Collaboration (NeIC), striving to provide computing and storage support for all e-science activities in the Nordic region. Although NeIC has no formal agreement with EMI, it contributed notably to evaluating products developed in EMI, like those related to data handling (ARC data management utilities and dCache) and operations support (Nagios probes for ARC components). NeIC sites are key in early testing of new ARC product releases. Four EMI partners (LU, UCPH, UiO and UU) contribute computing and hardware resources to NeIC and affiliated Nordic e-Science infrastructures (FGI in Finland, Swegrid in Sweden, DeIC in Denmark, Norgrid in Norway). These are long-term initiatives that rely on ARC and dCache middleware components currently produced in EMI. The spectrum of user communities supported by these infrastructures is already quite wide, varying from country to country, and future extension of NeIC to new e-Science areas will impose new requirements on middleware functionality and support.

One can identify two broad communities around NeIC that are directly interested in continuous long-term support of ARC and dCache EMI products: resource providers (academic computing centres) and research groups. All of them have a long history of contributing to ARC and dCache development, operation and support, pre-dating EMI, and it is foreseen that this will continue in future as well, in

close cooperation between the infrastructures and the NorduGrid collaboration that coordinates ARC development within EMI.

OSG

The US Open Science Grid signed a letter of interest and collaboration with EMI already during the project preparation phase. The collaboration has continued during the past two years with regular meetings and discussions on topics like interoperability (via the Interoperability Working Group) and sustainability (via collaboration on the concept of Scientific Software Innovation Institutes or S2I2). OSG software release process has been changed in the past year to be closer to standard open source guidelines, especially in the area of packaging. This decision was triggered also by the collaboration with EMI that adopted explicitly this strategy in his original proposal. OSG uses grid middleware services coming from many different technology providers, projects and university. It uses also a number of important EMI services, namely CREAM and VOMS. Work on extending the EMI VOMS Admin interface to become a replacement for the original VOMRS portal was done as part of this collaboration to converge to a single VOMS portal for both European and US Virtual Organizations.

More recently at the WLCG Workshop held in New York in May 2012, WLCG pledged support for the convergence of software engineering practices that EMI and OSG are working on to make the deployment and use of grid services increasingly more user-friendly.

EU-IndiaGrid2

The EU-IndiaGrid2 project [R4] capitalized on the achievements of EU-IndiaGrid by acting as a bridge across European and Indian e-infrastructures to foster evolution in these regions and to ensure sustainable scientific, educational and technological collaboration.

The MoU with EU-IndiaGrid2 was signed on 05/09/2011 and lasted until 31/12/2011, when the project ended. This agreement formalized the on-going collaboration between EU-IndiaGrid and the gLite partners, and testifies to the usage of EMI in India.

EU-IndiaGrid worked with gLite middleware in the context of the EGEE project series and continues to maintain ties with gLite-related policies.

Outcome: the collaboration action “Transfer of know-how on best practices in configuration and deployment of production services (this may cover release and change management, quality assurance guidelines, operational procedures, certification guidelines, etc.)” was fulfilled by the sharing of the SA2 QA policies [R5] and the evidence of EU-IndiaGrid2’s own policies as published at http://www.euindiagrid.eu/index.php/grid-elements-deployment.

The EMI exploitation action “Upgrade of EU-IndiaGrid gLite sites to the latest versions included in EMI 1 Kebnekaise release” was also fulfilled, as all project sites migrated to EMI 1 by 20/01/2012.

CHAIN

The CHAIN project [R36] aims to coordinate, harmonise and optimised the interaction among e-Infrastructure and specifically the grid interfaces between Europe and the rest of the world. The EMI harmonization goal finds fertile ground within the CHAIN vision. CHAIN and EMI have collaborated in occasion of several events in the past two years to define common activities especially in the field of interoperability. EMI members have provided input to several iteration of the chain interoperability plan. The main objective for EMI is to external the deployment and use of the EMI-ES interface and collaboration on adding direct support for it in the CHAIN portal. An MoU is currently being discussed.

FutureGrid

The FutureGrid Project [R37] provides a testbed capability that makes it possible for researchers to tackle complex research challenges in computer science related to the use and security of grids and clouds. These include topics ranging from authentication, authorization, scheduling, virtualization, middleware design, interface design and cybersecurity, to the optimization of grid-enabled and cloud-enabled computational schemes for researchers in astronomy, chemistry, biology, engineering, atmospheric science and epidemiology. The test-bed includes a geographically distributed set of heterogeneous computing systems, a data management system that will hold both metadata and a growing library of software images, and a dedicated network allowing isolatable, secure experiments. FutureGrid will form part of NSF's high-performance cyberinfrastructure. It will increase the capability of the XSEDE to support innovative computer science research requiring access to lower levels of the grid software stack, the networking software stack, and to virtualization and workflow orchestration tools.

A collaboration between EMI and FutureGrid was started in November 2011 to deploy a subset of EMI service images on the testbed to be used as demonstrative or training platform in US communities. Given the specific focus on XSEDE the first set of services deployed was taken from the UNICORE stack [R38]. More services will be added in the third year.

EPEL/Debian

The most efficient distribution channels for open source software are the official repositories of the Fedora/EPEL [R39] and Debian [R40] distributions. One of the major goals of the EMI project is to be able to release the major services and most of the clients via standard repositories by the end of the project. This objective is being pursued by explicit software engineering policies add an increasingly stricter set of release criteria at every yearly EMI distribution. At the time of releasing EMI 2 in May 2012 the following products are directly available from EPEL or Debian: ARC (various services), BDII, DPM, Gridsite, lcg-utils, LFC, VOMS and a number of other security libraries and components.

3.4 INFRASTRUCTURE

The infrastructure layer is the main market for EMI. The infrastructure operators and managers are the actual users of the EMI services and products. The EMI products are currently deployed on grid sites coordinated by EGI and managed by site administrators in the collaborating NGIs and by HPC sites coordinated by PRACE. In addition, several communities have coordination or development projects that perform regular assessments of the status of the infrastructures, plan their evolution in function of the expected needs and analyse the technological requirements. Most of the dissemination and training activities are targeted at these categories of users.

This layer is also where most of the exploitation results are achieved. A measure of the success of EMI is the continued deployment of its products in European and international grid and HPC sites. The main exploitation results are described in the Exploitation section 3.6.

This is a niche market and is currently saturated by EMI and other technology providers like Globus (through the IGE project), EDGI, Condor and a number of experimental cloud projects and initiatives (StratusLab, OpenNebula, Eucaliptus, Nimbus, OpenStack, Globus Online among the most relevant). The market of site and infrastructure managers and has a size of around 2000 users (estimated from the number of registered users in the infrastructure management VOs) and is quite stable in terms of new operators. The number of grid sites has increased by about 3% in the past 12 months [R6],

although the overall computing and storage capacity has grown more than 30% in the same period. Most site managers are experts in their field and the changes in the user base are essentially due to generational and professional turnover in the existing sites more than the increase in the number of site managers. This is consistent with the number of visitors of the EMI web site and the people attending conferences and training events.

3.4.1 Collaborations

The main collaboration mechanisms are MoUs and SLAs. The most important collaborations continued or established during the second year have been:

WLCG

The Worldwide LHC Computing Grid is the body coordinating the computing activities of the High-Energy Physics (HEP) community. HEP is the largest community making use of the grid today having more than 90% of the grid resources allocated to them by the grid sites. WLCG monitors the use of the infrastructures by HEP users, coordinates the computing and storage resource pledges of grid sites supporting the HEP VOs and decides on technology matters. WLCG has two main coordination bodies. The Management Board (MB) is composed of the HEP experiment computing coordinators, representatives of the main HEP centres and technology providers. The Grid Deployment Board (GDB) is composed of representatives of the grid sites.

EMI is a member of both the MB and GDB and participate to WLCG activities by collecting feedback, reporting on software products releases and participating to technology evaluations.

At the end of 2011, WLCG started a revision of the computing and data management technologies used by HEP. A number of Technology Evaluation Groups (TEGs) have been formed to evaluate different technological areas (Computing, Data Storage and Transfer, Databases, Operations and Software Configuration, Security) and provide requirements for future development. EMI members have been members of all WLCG TEGs (except the Databases one, which is out of scope). All requirements have been discussed and will be incorporated as needed in the third year technical plan.

Given the direct membership of EMI in the main coordination bodies of WLCG, no need was found to have a formal MoU.

LSGC

The Life Sciences Grid Community is a group of 5 VOs for a total of about 500 users served by a small support and system admin team. The LSGC VOs have very specific requirements in terms of data management and have provided the largest number of requirements to EGI and EMI. During the first year of EMI, some difficulty was encountered in discussing the requirements using the EGI-led Technology Coordination Board, since the users were not directly represented. No technical discussion could take place. In order to avoid this issue, EMI has proposed in January 2012 during the Sustainability meeting organized by EGI in Amsterdam that direct technical contacts must be established between the technical experts of the grid user communities and the technology providers. This proposal has been met with the strong support by LSGC and discussions were organized in March at the EMI Technical Conference in Munich and again in April. Regular technical discussion to follow up the requirements are to be organized in the coming months.

D4ScienceII and iMarine

The D4Science-II (Data Infrastructure Ecosystem for Science) project aimed to provide a mechanism for facilitating interoperation of the D4Science e-Infrastructure with other diverse data e-Infrastructures that are running autonomously to create the core of an e-Infrastructure Ecosystem.

The iMarine (Data e-Infrastructure Initiative for Fisheries Management and Conservation of Marine Living Resources) project is chartered with establishing an open data infrastructure that will

effectively support the implementation of the ecosystem-based approach to fisheries management and conservation of marine living resources.

These consecutive projects also maintain and evolve the gCube software [R7] that enables the creation of Virtual Research Environments (VREs). gCube depends on EMI products to work and uses the EMI-managed ETICS system for its builds.

MoUs were signed between EMI and D4Science II in November 2012 and between EMI and iMarine in February 2012. The main objectives of this collaboration are:

EMI Build infrastructure: iMarine requires access to EMI build services to build their software. EMI provides access as part of the "Work with EMI" program.

Adoption of EMI releases: to facilitate the adoption of EMI releases on the iMarine infrastructure, iMarine requires early access to new EMI technology. After the EMI 2 release, iMarine will start deploying the new EMI-ES enabled CEs and build operational experience with it.

Feedback: EMI requires timely feedback on its pre-production and production services, which iMarine can provide as it assesses the EMI software.

EUDAT

The Pan-European EUDAT project aims at establishing a collaborative data infrastructure (CDI) to serve the needs of a wide variety of pan-European research communities including those being listed in the different scientific themes of the roadmap released by the European Strategy Forum on Research Infrastructures (ESFRI). The EUDAT CDI consists of policy-based data management services, persistent identifier services (PIDS), and monitoring services that all support the major goal to improve the data preservation, maintenance, and accessibility of scientific data arising within research communities.

Another crucial service identified as part of the EUDAT CDI ecosystem is a service registry where several solutions are currently reviewed for its suitability including the GOCDB [R8] system known from the EGI infrastructure.

At the EGI User Community Forum 2012 in Munich [R9], several discussions took place with EUDAT representatives that lead to the investigation of the EMI Registry (EMIR) product in the next months.

If the evaluation is successful, the EMIR product will represent the EUDAT CDI registry listing available data healing, and PID services across Europe [R10] [R11] [R12].

An important infrastructure component is also provided by collaborations with projects developing high-level services of general use, like workflow management tools and portals. Collaboration with application developers in this context helps ensuring that compatibility between EMI and the high-level tools used by end-users is kept and that EMI products can be deployed on infrastructures used by those communities. The following collaborations have been established:

SHIWA

The SHIWA project (Sharing Interoperable Workflows for large scale scientific simulations on Available DCIs [R13]) is developing interoperability solutions for a number of workflow systems, including technologies from ARC, gLite and UNICORE.

SHIWA and EMI are currently in discussions to formalize their collaboration via an MoU that is currently envisioned to contain the following actions:

- EMI release preview and assessment;
- EMI release exploitation;
- SHIWA requirements;
- Access to SHIWA Simulation Platform.

The collaboration with SHIWA would be based on the "Works with EMI" program.

3.5 END-USERS COMMUNITIES

The end-users communities are the actual users of the cumulative set of added-value services of the layered market stack. End-users use top-level services like portal, user interfaces, domain-specific applications to perform their work as scientists, engineers and researchers. End-users have very limited knowledge of the underlying technology at the level of middleware and are not concerned with it. They are only concerned with the availability and efficiency of the top-level services provided by the infrastructure, which in turn relies on the stability and efficiency of the middleware. If the value-added chain is correctly implemented a relative small change at the middleware layer has the potential of bringing many additional end-users. However, it is necessary to have compelling services at the infrastructure level to attract the users in the first place.

The major end-user community in the research grid market is the High-Energy Physics (HEP) community. The community size is currently estimated at around 8500 users out of the about 20000 users registered in all EGI VOs. These users use 90% of the grid resources. The second largest community in terms of registered users, if one excludes the community of infrastructure managers (about 2000 users), is Life Sciences with 1100 users.

EMI has spent considerable effort in establishing direct contacts with other important Virtual Research Communities. However, so far this has been possible only with projects that directly manage an infrastructure or include technical activities close to the infrastructure layer, for example to develop portals or user interfaces that need to interact directly with the EMI APIs.

3.5.1 Collaborations

The main collaboration mechanisms are MoUs and Letters of Interest. The most important collaborations continued or established during the second year have been:

HEP

The formal collaboration with the HEP community mainly goes through the WLCG project as described earlier. However, EMI interacts directly with members of the community in occasion of conferences like ISCG 2012 (February 2012), the EMI Technical Conference (March 2012) and most importantly the Computing for High-Energy Physics (CHEP) 2012 in May 2012. CHEP 2012 has been sponsored by EMI and several presentations on the technology and the software engineering methodology have been presented. In addition, EMI directly promotes its values and technology by presenting EMI in occasion of workshops of specific HEP experiments, like the ATLAS Computing Week in July 2011.

WeNMR

The Worldwide e-Infrastructure for NMR and Structural Biology (WeNMR) project builds on the established e-Infrastructure for Nuclear Magnetic Resonance (eNMR) and its steadily growing virtual organization, which is currently the second largest VO in the area of life sciences. WeNMR will serve all relevant INSTRUCT communities in line with the European Strategy Forum on Research Infrastructures (ESFRI) roadmap.

EMI signed an MoU with WeNMR in April 2011. The collaboration actions include:

Preview of EMI release: The WeNMR infrastructure uses the CREAM CE, LCG CE, WMS, LB, BDII, Hydra, LFC, VOMS, DPM and StoRM services. Within the framework of the 'Works with EMI' program, WeNMR has access and preview to early releases of CREAM CE, WMS, VOMS and other EMI provided services via the EMI Testing Infrastructure. Users subscribe to the testers.eu-emi.eu VO to access the resources of this infrastructure. WeNMR may also investigate early releases of other compute flavors to increase the size of its infrastructure and validate the robustness of WeNMR to plug-into other compute services. This action is part of the *Works-with-EMI* program. The project joined this activity informally before the signing of the MoU.

EMI release exploitation: WeNMR upgrades the gLite nodes of its production infrastructure to the latest versions of CREAM CE, WMS, VOMS and other relevant services included in EMI releases. In addition, WeNMR also intends to use new and improved offerings from the EMI releases that may include several features that the WeNMR community expresses interest in. As of January 2012, 6 CEs supporting WeNMR have deployed EMI 1; others have deployed VOMS from EMI 1 and generally gLite 3.2.

WeNMR feedback and requirements: As WeNMR gets early access to services in the development phase and exploits the EMI services, feedback, recommendations and lessons learned are provided to validate and improve EMI services. Additional specific requirements may also be given to EMI to increase the usability and reliability of its services. WeNMR's requirements for 2011 are documented in the savannah task 22854 [R14]; the project is providing regular feedback.

Outcome: As part of this collaboration, many service instances in the WeNMR infrastructure have been migrated to EMI 1 services. Specific WeNMR requirements have been collected and merged with the global requirements trackers managed by EGI and EMI. The feasibility of a pilot project amgin WeNMR, Sara and INFN to test the use of WNoDeS is being discussed (see also section 3.7.2.3)

DECIDE

The DECIDE (Diagnostic Enhancement of Confidence by an International Distributed Environment) project aims to design, implement, and validate a GRID-based e-Infrastructure building upon neuGRID and relying on the Pan-European backbone GEANT and the NRENs. Over this e-Infrastructure, a service will be provided for the computer-aided extraction of diagnostic disease markers for Alzheimer's disease and schizophrenia from medical images.

EMI signed on MoU with DECIDE in September 2011. The collaboration actions are as follows:

Testing and exploitation of EMI release: DECIDE has access and preview to releases of the following EMI services via the EMI Testing Infrastructure: (a) Compute Area: gLite Job Management (includes BLAH, CEMon, CREAM, WMS/LB), gLite MPI; (b) Data Area: StoRM, DPM, AMGA; (c) Infrastructure Area: APEL Client, DGAS Client and services from gLite Information System; (d) Security Area: VOMS and services from gLite Security. Users subscribe to the testers.eu-emi.eu VO to access the resources of this infrastructure. DECIDE upgrades the gLite nodes of its infrastructure to the latest versions of gLite Job and Data management services included in EMI 1 and EMI 2 releases.

Other EMI services will be integrated in the DECIDE infrastructure on the basis of evaluation and testing results. Feedback, recommendations and lessons learned are provided to improve EMI services. This action is part of the *Works-with-EMI* program.

DECIDE requirements: DECIDE submits requirements to EMI to better support the development of its diagnostic services relying on EMI services. The evaluation of these requirements and specification of implementation plan, if approved, are coordinated by EMI PTB.

Outcome: at the time of writing, the DECIDE project has not yet submitted any requirement to EMI

3.6 EXPLOITATION

The exploitation of EMI results in the Research Grid market is primarily aimed at the grid and HPC sites deploying the EMI products and using the documentation and support services. Additional exploitation results are represented by the use of the EMI software engineering services and tools by other projects.

3.6.1 Exploitable Items

3.6.1.1 The EMI Software Products

The main exploitable item is the EMI software. As of end of April 2012 the EMI 2 release contains 56 products [R15] composed of a total of more than 450 packages on three different operating systems (SL5, SL6 and Debian 6). All products are available on the EMI Repository at:

http://emisoft.web.cern.ch/emisoft/index.html

A number of important products like VOMS, BDII, gridsite, LFC, all ARC services, UNICORE services and others are also distributed via standard operating system repositories from Fedora/EPEL and Debian or from Maven Central, the main public Java repository.

As of the EMI 2 release, EMI doesn't not provide anymore a custom release mechanism for EGI UMD, since the main goal is to transition to a standard open source model. EGI UMD can compose the distribution selecting which products out of the EMI releases are most suitable for its user communities, but the recommendation we have provided is to allow users to deploy from Fedora/EPEL with the EMI repository as official upstream repository. The transition will have to be completed by the end of the third year of EMI by when most EMI Products are expected to be available from operating systems repositories.

3.6.1.2 The Products Technical Documentation and the Factsheets

Technical documentation for all products is available from the EMI web site on each product page. The documentation is produced in standard formats (mostly PDF or HTML) and is regularly updated in case of change. Most of the documentation is targeted at EMI users in the Infrastructure layer that is grid and HPC sites administrators and application developers. Generic documentation for the Clients is also provided, although the actual specific documentation on how to use the Clients for particular VOs or sites is left to the VO and infrastructure managers.

In addition to the technical documentation, an important new type of document has been introduced in the second year, the Product Factsheet. An EMI Product Factsheets Suite is available on the EMI website at

http://www.eu-emi.eu/product-factsheets

The suite currently consists of 16 factsheets describing at a glance in a condensed way all critical information about individual EMI products. The aim of these factsheets is to attract interest in the products and not to provide the sort of technical information that can be looked up in the EMI product documentations if needed. Apart from being available at the EMI Website, the factsheets are used as a tool at various events (e.g. Supercomputing 2011, ISGC 2012, 2nd EMI Technical Forum) in discussions with people stopping at the booth EMI usually installs at those events. The purpose is also to highlight key features that make it easier to allow technical users to assess whether the EMI products can be used in their specific context of interest. The number of factsheet is expected to grow during the course of the third EMI year and provide also information about specific 'use cases' that take advantage of the described EMI product.

The Factsheets download count is also being monitor to gauge the relative interest in each product.

Title	Downloads
CREAM-CE.pdf	214
WMS.pdf	203
StoRM-SE.pdf	185
ARGUS.pdf	171
dCache-SE.pdf	168
UVOS.pdf	146
DPM.pdf	124
VOMS.pdf	115
AMGA.pdf	93
EMIR.pdf	91
LFC.pdf	87
HiLA.pdf	73
UNICORE-Gateway.pdf	70
EMI-ES.pdf	67
MPI.pdf	67
ARC-CE.pdf	13

Table 1: Factsheet download count

3.6.1.3 The Software Engineering Process and Tools

The software engineering process is fully documented on the EMI TWiki web site at:

https://twiki.cern.ch/twiki/bin/view/EMI/SA2

The EMI SE policies have been shared with other projects and initiatives and have influenced or are being directly used by EGI, IGE, iMarine, EUIndiaGrid 2 and OSG. Most of the tools used in the EMI Software Engineering process are standard open source tools like mock and pbuilder. The ETICS system used to coordinate the process and the releases is currently maintained at CERN and is

provided to other partner projects as part of the "Work with EMI" program (iMarine) to manage their own software.

3.6.1.4 Training course and online tutorials

Training material and online tutorials are registered in the EGI Training Marketplace for further use by the EGI trainers. The main training strategy is to train the trainers from NGIs and other projects. This concept will be further extended in year 3 in collaboration with EGI with the production of online tutorials (webinars) on topics selected by the site administrators especially those from newer sites (mainly in Eastern Europe).

3.6.1.5 Expertise

An important exploitable item provided by EMI is the expertise built by its members during the implementation of the project objective. As the project progresses, the Product Teams are visibly becoming more competent not only on domain-specific technologies such as data management or security, which is their main strength, but also on several software engineering areas, like software packaging, release management and software testing and quality control. As reported in the yearly deliverable on QA by SA2, the number of components rejections by EGI due to non-compliance with the agreed acceptance criteria has steadily decreased during the year. This expertise will remain after the project ends and will become an integral part of the activities of the EMI developers.

3.6.2 Licensing

The use of proper licenses for both software and documentation plays an important role in the exploitation implementation. The EMI project partners have pledged to use only OSI open source licenses for software and OpenAccess licenses for documentation. The use of the correct licenses is monitored by NA1 and any deviation is discussed and corrected.

Until now, only one problematic case was identified, namely the use of a proprietary license for the dCache software. Before 1 January 2012, the license under which the dCache software was published allowed open and unlimited access to the product and the source code, but was not compatible with any license commonly known as "Open Source License", as defined by the "Open Source Initiative" (OSI) [R16]. DESY, the provider of dCache, has actively addressed this issue. With effect from 1 January 2012 [R17] the dCache code is available under the GNU Affero General Public License (AGPL) Version 3 [R18]. Subcomponents might be offered under other Open Source Licenses. However, those will always be less restrictive than AGPL (e.g. BSD, LGPL). The dCache collaboration reserves the right to make different branches of the code available for paying customers under a proprietary license. However, this does not apply to code distributed by EMI. Although the AGPL license is in effect, some dCache.org web pages or source code header files might not be updated yet. This will be fixed soon.

3.6.3 Exploitation results in Research Grids

The main exploitation result in the Research Grid part of market is the actual deployment of EMI service in the European and international grid infrastructures.

As of May 2012 [R19] the EGI infrastructure is composed on 352 sites, of which 299 from 42 European Countries and CERN, 27 from Asia-Pacific and 26 from Canada and Latin America (Figure 4).

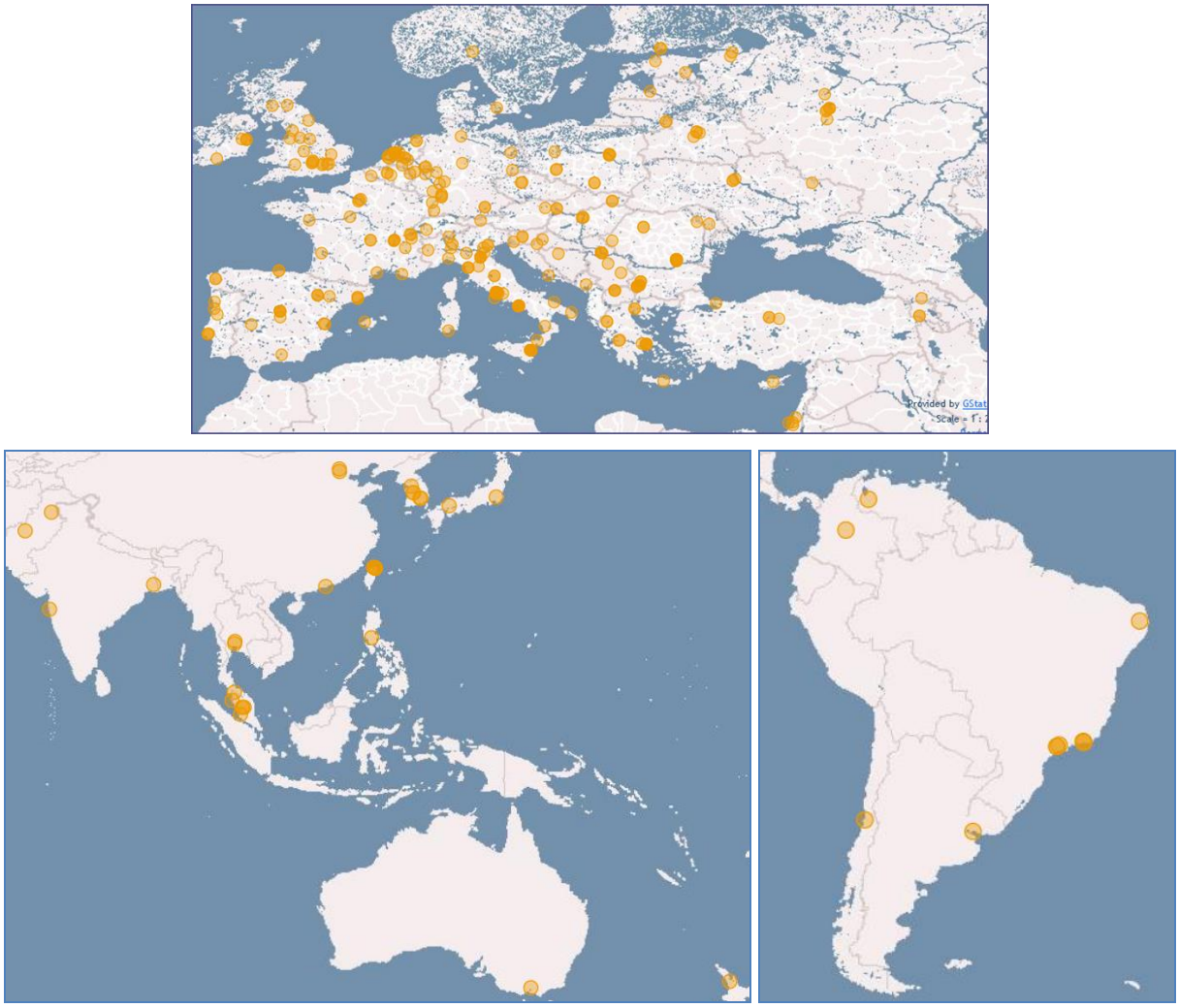

Figure 4: Maps of EGI sites using EMI services (includes gLite 3.x)

All sites have deployed one or more EMI-supported services (including services from pre-EMI gLite releases still supported by EMI). All grid sites have installed a cumulative number of 359 core services (64 VOMS instances, 150 WMS, 41 LFC and 104 Top-BDII), 408 compute elements (272 CREAM, 31 ARC instances, 5 UNICORE Gateway) and 328 storage elements (206 DPM, 67 dCache, 51 StoRM and 4 UNICORE StorageManagement).

Since the release of EMI 1 in May 2011, about 24% of service instances have been upgraded from older releases [R25] with peaks of 40% to 50% for the Storage Elements. This represents a very fast uptake considering the traditionally very slow upgrade rate of the grid infrastructures (a number of sites are still running services released more than 4 years ago and superseded by newer services, like the LCG-CE service superseded by CREAM-CE already in EGEE II). The rate is expected to increase even faster with EMI 2 thanks to the support for SL6.

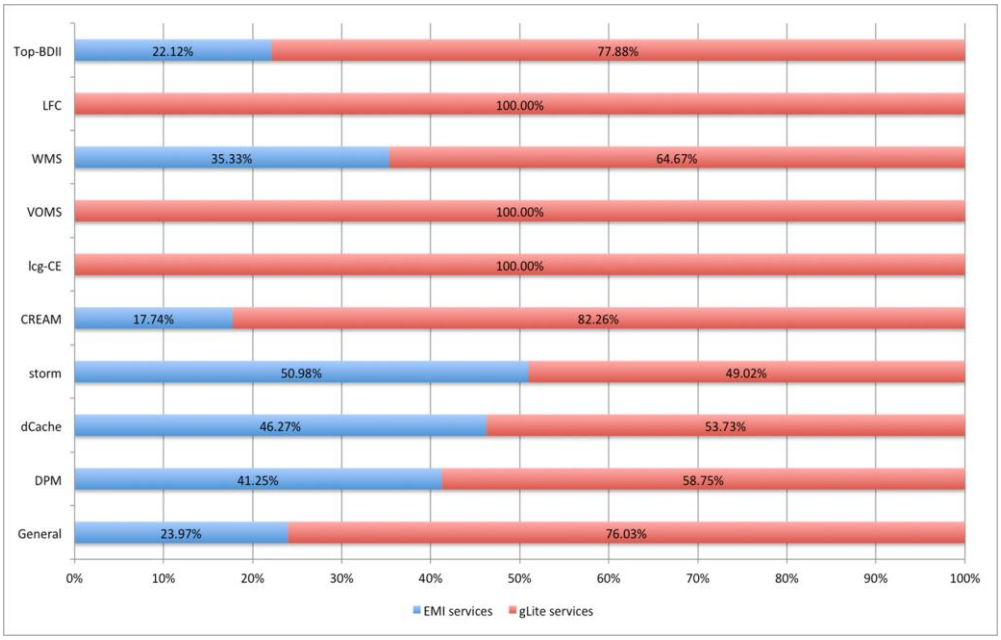

Figure 5: Overview of upgrade ratios from gLite 3.x to EMI 1

An important result of the collaboration of EMI with EU-IndiaGrid 2 (as mentioned in section 3.3.1) was the migration of all gLite instances installed at sites in India to the equivalent EMI 1 versions.

3.6.4 Exploitation results in HPC

PRACE

The PRACE user community use mostly UNICORE services. The heart of the PRACE environment is composed of the PRACE Common Production Environment, the Accounting Information service, and the UNICORE middleware. The following services are used:

- GridFTP is used for data transfers.
- The PRACE Common Production Environment (PCPE) is a set of software tools and libraries that are available on all PRACE execution sites. It defines a set of environment variables that try to make compilation on all sites as homogeneous and simple as possible.
- UNICORE Rich Client (URC) is a utility to create, submit and monitor PRACE production jobs in a simple way. The user doesn't have to know the details of the target platform's batch system.
- MyProxy Service helps in storing the proxy certificates for later usage when, for instance, your private key is not at hand. It is best to avoid distributing the private key across several machines, but to keep it safe on your workstation instead.
- PRACE Batch Systems are currently deployed at each PRACE platform. They offer basic commands of each system, and then list some site-specific commands, which may be of interest.
- Interactive access is provided to all systems for which access is granted as part of a PRACE allocation of resources. Within PRACE employing gsissh is recommended for interactive access to your Execution Site.

- Accounting exercises are provided by the Distributed Accounting Report Tool a client application that gathers accounting information provided by every PRACE site in a standardized manner. It generates reports of the usage of distributed computing resources within PRACE. These reports can be specific for project supervisors, site administrators and users.

Actual deployment figures have been asked to PRACE, but are not yet available at the time of writing.

HP-SEE

HPSEE is a project that targets to mitigate the distributed computing and data storage infrastructural differences in the South Eastern European region. One of its primary objectives is to collect, and integrate resources, and make them accessible uniformly for those users who cannot afford such facilities in their native environment (typically to the researchers of the Mid-East countries, Albania, etc.).

The project has a dedicated work package that addresses interoperability and scalability issues, and also technology watch reports. Currently there are 4 major resource providers in the project: Bulgaria, Greece, Hungary, and Serbia. The project also puts efforts into harmonizing software stacks over the different resources in order to leverage the user experience on all sites.

Currently the main technical development direction follows similar patterns as in PRACE, but that level of maturity has not been achieved yet. There are some metrics defined over the different pieces of software that measures the software distance among the sites. Nevertheless most of the sites do not even contain grid middleware, and there is no strategy defined for installing common middleware components so far. Based on the findings, this community is straightforward space for EMI components to be distributed and installed.

Some sites, e.g. the Hungarian sites already feature some of the ARC components taken from EMI repositories. An action plan will be formulated to promote EMI software towards the infrastructure operators of this community (HP-SEE WP5) [R21].

ARC components installation on Hungarian supercomputer sites

The Hungarian NIIF Supercomputing network is a heterogeneous system. The four geographically distributed, and WAN-connected sites are operating different architectures with different Linux operating systems.

On two of the supercomputer sites (located in Szeged and Budapest) Red Hat 5.6 Linux is installed. There have been suitable EMI-ARC packages available in the corresponding repositories for this operating system platform and the EPEL repository contains every missing dependency. The components were installed using YUM package manager. These nodes are actually running ARC 1.1.0.

On the two other sites (located in Debrecen and Pécs), the platform is Suse Linux for that there were no packages available in the official channels. For these nodes the components were manually compiled, and packaged. The middleware was patched to enable remote grid management components. The following services were installed and configured:

- nordugrid-arc-arex,
- nordugrid-arc-arcmgmt,
- nordugrid-arc-client, and

- nordugrid-arc-aris.

The set of installed services is quite similar on the different sites; there are only minor differences. On every location there is a dedicated head-node where the services are running. The common setup is that A-REX execution service (with SGE 6.2u5 LRMS backend) and ARIS (LDAP) information system services are located on every machine. The sites in Szeged and Budapest are running GridFTP, while sites in Debrecen and Pécs are using management services too.

For the information level integration we had to register them into a central, country scoped EGIIS service that is located at [R22]. On the supercomputer sites the ARIS services are communicating with the index server that stores, and then forwards the service registrations to the global NorduGrid index servers. The country-scoped service is hosted in the NIIF cloud. The registered services also can be queried from the global NorduGrid database and for example the computing services can be visualized on the following homepage: [R23].

The sites are available for not only the national, but also for the international user communities too, as a user as well as a project registration facility has been set up.

User authentication is purely performed by supported ARC methods like the grid-mapfile, identity.map, or allow.pdp technologies. After the successful authentications the certificate user DNs are mapped to local usernames, so the jobs can be executed separated from each other.

3.7 SUSTAINABILITY

As mentioned at the beginning of this document, the EMI sustainability plan consists in supporting continuous innovation and identifying the possible financial mechanisms that allow the partners to fund it.

The Research Grid and HPC market described so far can be considered a niche market providing services to slightly more than 20883 users [R24], an estimated 1% of the potential scientific research market in Europe. As outlined already, in order to create sustainability for the EMI services it is necessary to expand into additional markets. However, the Grid market cannot be lost in the process and can actually serve very well as a mature base from which wider innovation can be created. The first step is to understand how EMI can maintain and increase its current market share. The second step is to understand how the support and innovation can continue after the end of the EMI project itself.

The current market in Europe is served by technologies from EMI and from a number of other technology providers, including Globus (supported by the IGE project), Desktop Grids (supported by the EDGI projects), various cloud IaaS implementations (OpenNebula and the StratusLab project, Microsoft Azure and the Venus-C project, the US-born OpenStack), QosCosGrid from the Poznan Supercomputing and Networking Center and a few others. At the moment the EMI services are by far the dominant technology within EGI and PRACE. However, ignoring the concrete risk of losing market share in favour other technologies, especially cloud, would be shortsighted. In order to maintain the EMI services competitive a number of activities are explicitly part of the EMI program of work and have been implemented during the second year.

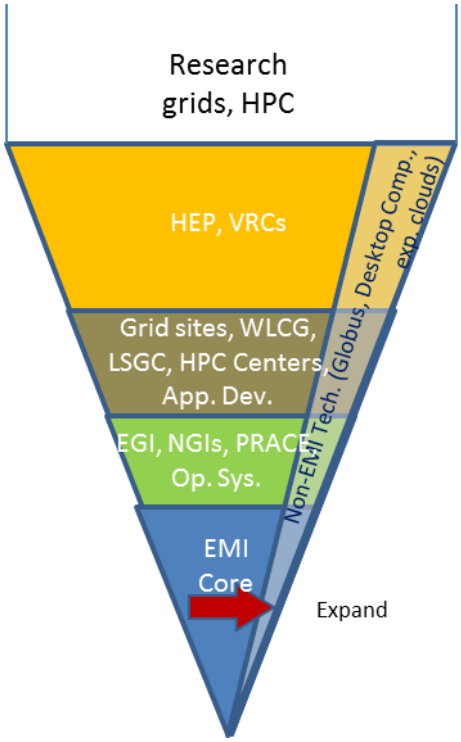

Figure 6: Market expansion

3.7.1 Interoperability and standardization

Interoperability and standardization are two key drivers of the sustainability plan. A detailed description of the interoperability and standardization activities and results is provided in the NA3 deliverable DNA3.2.1 [R34]. In this section it is important to highlight their impact on the sustainability of EMI products.

The importance of interoperability and standardization lies in two main factors:

1) Interoperability among EMI services and between EMI services and other technologies allows different technologies to be installed in the same sites or sites based on different technology to work together. Interoperability among different infrastructures in different geographical regions is a fundamental requirement for all infrastructure managers, since it is neither possible nor desirable to have a single technology everywhere, yet resource and data sharing is necessary. Non-interoperable middleware service would not be used for long time. In addition, sites deploying the EMI services can get access to non-EMI-based resources using the standard EMI services and can therefore increase the number of end-users and attract new communities.
2) Standardization has similar advantages as interoperability with additional important values. Services adopting standard interfaces are easier and less expensive to maintain within open source collaborations, since more people are available to provide tests and fixes. In addition, standard services can in principle be replaced with other standard services. This is very important for smaller infrastructure that may want to try new services, but cannot afford the effort needed to climb new learning curves.

It is of course understood that the advantages of interoperability and standardization are bi-directional and can be used by other technology providers to gain market shares against EMI. However, in the presence of interoperable and standard solutions, the users can decide based on other factors like the performance, efficiency, security and quality of the solutions, for which EMI has an advantage compared to other less mature or less complete technologies.

3.7.2 Competitors analysis

During the course of the second year, the relative importance and potential impact of other technology providers has been monitored and a number of actions have been taken.

3.7.2.1 Desktop Grids

Desktop Grids are not heavily used within the standard EGI infrastructure with only a minority of the countries providing resources [R35]. However, Desktop Computing gives potentially access to a very large number of resources to run large numbers of small highly parallelized jobs. The EDGI project and EMI have signed an MoU and an OLA to jointly work on the integration of EMI-based service grids via the EDGI Bridges for gLite CREAM, ARC and UNICORE. The bridges are distributed as part of the EMI 2 release and allow plugging desktop grid resource into the EMI Compute Elements to be used as standard batch queues. This allows the integration of Desktop Grids within the EGI infrastructure provides access to additional low-cost resources and generates the potential to increase the EMI and EGI user base without requiring a switch to Desktop Grid clients.

3.7.2.2 Globus

Globus-based infrastructures are a minority in Europe and EGI with the largest presence in Germany and a few instances of GridFTP in Finland and Spain. In addition, support for traditional Globus is constantly decreasing also in the US, while the new cloud-like Globus Online service becomes more popular. Some work on common standardization has been done with OGF/PGI, but standard Globus services are not seen as a major threat.

3.7.2.3 Cloud

The opinions on the impact of cloud on the current grid market vary wildly. However, it is also true that the word “cloud” is often used with many different meanings. In order to understand how EMI should position itself with respect to cloud and cloud projects, it is necessary to separate the concept of “cloud as a technology” from “cloud as a business model”.

From a technological point of view, cloud is essentially a resource provisioning method-making use of virtualization technology. In this respect cloud can easily be seen as a natural evolution of grid enhanced by more efficient provisioning methods. Indeed, the use of virtualization, both static and on-demand, is now routinely used by grid sites[2].

From a business model point of view, the situation is more complicated. While grid is essentially a way of sharing resources across different domains, cloud is a way of getting resources for a single domain or even a single user. In addition, the current economic model of the grid is of CAPEX type, the NGIs are the owners of the resources and have a large investment in hardware and infrastructure, which must be amortized and currently prevents buying large quantities of resource access from cloud providers. Cloud will become more appealing when the current hardware is decommissioned. In that

[2] EMI itself is managing its main testbed at CERN using this method. When users submit build or test jobs, a simple script analyses the request to extract platform information and then uses a HyperV-based hypervisor to instantiate the proper image, which in turn becomes a Worker Node in the system grid scheduler.

case instead of replacing the hardware, many Institutes may decide, depending on price and other strategic factors, to buy resource access, thus turning the model from CAPEX to OPEX. New or small communities not supported by existing NGIs resources can of course be already in an OPEX scenario, In this case the use of cloud could already make sense.

The question therefore arises whether today the cloud model is suitable for the type of scientific research supported by the grid for existing large infrastructures like WLCG or LSGC. The CAPEX vs. OPEX problem can be addressed by turning the grid sites into private clouds. As a consequence, most of the existing investigation in EGI, StratusLab and other projects is about understanding how to federate multiple private and commercial clouds, so they retain the federated aspects of grid while exploiting the easier provisioning model of cloud at the user level.

However, the simple transformation of grid resources into private cloud resources has two consequences to be taken into account. On one hand end-users today are not used to be given nodes to be used as they want. The typical end-users want to run an application and get results. Therefore an intermediate layer of cloud-enabled services has to be created to exploit the new private clouds. This layer of services of course already exists today in the grid and is provided by technology providers like EMI and application developers having developed grid-enabled resources.

The second issue is that federating private clouds requires addressing a number of issues in terms of security, accounting, monitoring and other operational aspects, which have been already addressed within the grid. The risk exists of making very complicated the simple model that makes cloud interesting. Indeed the sharing aspects of scientific research lies more in data sharing than computing resource sharing and an altogether different approach to the problem may be necessary.

The current market position of EMI in the context described so far is that EMI must not invest effort in providing cloud technologies or trying to address specific cloud requirements. There are, however, strategic advantages in understanding which EMI technologies can be used within the current cloud and grid interoperability investigations. In addition, EMI needs to understand more carefully what users really want when they say they want cloud. These two aspects have been so far addressed as follows:

WNoDeS

The Worker Nodes on-Demand Service is a new service in the EMI 2 distribution. The service has been developed by INFN, an EMI partner, outside the EMI project. In occasion of a "vision" meeting organized at CERN in December 2011, it was decided to include the service in the distribution to address the need of integrating grid resources with cloud provisioning methods. WNoDeS allows users to request access to grid resources using standard cloud APIs. The advantage is that the same resources can be seen as grid or cloud resource depending on the user needs. In addition WNoDeS allows integrating external cloud resources in the standard grid CEs (currently CREAM) so they are seen as batch queues. This has the advantage of expanding the available resources for traditional grid users from private or commercial clouds without any change in existing UIs and applications. The service has been assessed by EGI as part of the on-going federated cloud activities and its deployment in the infrastructure will be assessed in the coming months.

ARGUS-EES

This service allows applying the Argus authorization framework at the core of the EMI AA implementation to virtualization managers. The functionality is targeted to system administrator that can control who can request and control virtual machines in their private cloud using a central set of policies. The service is being adapted to work with OpenNebula, other virtualization managers can be added depending on user requirements. The advantage of this service is the extension of existing grid

security solutions to the federated private cloud concept retaining the technological leadership of EMI in this area.

Cloud survey

An important aspect of the cloud competition analysis is the understanding of why users are asking for clouds. The rationale is that with a better analysis of requirements, a number of perceived or real user issues can be solved by the proper use of existing EMI services. The EMI Strategic Director has conducted a number of interviews with representatives of user communities and projects within HEP and other VRCs asking why they want cloud (in case they do want it). The main motivations behind the request of cloud are the expectation of higher efficiency and the possibility of customizing the "execution environment" of their jobs. One of the most representative cases is that of WeNMR. In a presentation delivered at the EGI User Forum in Munich in March 2012 [R25], Alexandre Bonvin, the WeNMR Director, presented two use cases for using cloud in WeNMR. The first use case is about getting access to machines with 32-cores of memory and more. The second use case is about being able to have specific domain applications installed on the Worker Nodes. Neither use case requires specifically cloud to be solved. In the first case, access to machines with high numbers of cores is of course possible with grid, provided that the nodes exist. Of course cloud would allow users to buy access to such resources from commercial providers, but there is no technological limitation and more and more sites are deploying large multi-core hypervisors. The second case can also be addressed with existing EMI services either using traditional software installation methods used for example by HEP VOs or using the new WNoDeS service with custom images. A pilot project among INFN, Sara (the main site in the Netherlands supporting WeNMR jobs) and WeNMR is being currently discussed to validate these assumptions. In addition, in the same presentation Dr. Bonvin clearly stated that the WeNMR community is currently happy with grid and have over the years set up a well-managed portfolio of grid-enabled portals that allows them to carry out efficiently their work and that they do not want to be disrupted.

3.7.3 Long-Term Support Plans

The long-term support plans for the EMI products after the end of the EMI project have the object of many discussions among the partners. Opinions on how to support and develop further the software differ among partners depending on the specific business objectives of their Institutes. However, there is general consensus that most of the software products must be foreseeably supported for at least the next 12 to 18 months as they currently are. The support period is chosen to match at least the duration of the EGI-InSPIRE project.

In March 2012 a plan was agreed with the mediation of the Project Director to implement a support plan in a sequence of steps of increasing level of commitment:

Step #	Description	Complete by
Step 0	Collect information about which EMI partners will support the existing products after the EMI project and the scope of the support (which communities or categories of users). The reference period is 12 to 18 months. Identify any critical gap and discuss it with the major stakeholders (EGI, WLCG, PRACE).	**May 2012**
Step 1	Make a list of the global tasks currently performed by EMI (release management, support coordination, quality assurance, technical coordination, etc.) and analyse which tasks, if any,	**July 2012**

	can still be performed at a global level after the EMI project with the available funds	
Step 2	Investigate which form the future collaboration should take in order to provide the software and services defined in step 0 and 1	**September 2012**
Step 3	Implement the agreed structure, if agreed, and prepare the transition by the end of the project	**April 2013**

Table 2: Long-term support plan decision process

Step 0 was performed and completed in May 2012. The complete list of EMI products with the supporting Institutes after April 2013 and the scope of the support has been produced and presented internally to the EMI CB and circulated to EGI and WLCG for further discussion. The list is shown below in Table 3. The identified gaps are reported in Table 4. A formal discussion between EMI, EGI and WLCG is planned for June 2012.

Product	Institute	Scope
AMGA	KISTI	Support for existing users in WISDOM, GISELA, gLibrary, DKRZ, INDICATE, Belle II
APEL Parsers	STFC	Existing batch systems in EGI
APEL Publishers	STFC	Existing batch systems in EGI + SSM
ARC CE	UIO, LU, UCPH, UU, NIIFI, UPJS and other NorduGrid partners	worldwide support (anybody who installs it from any repository)
ARC Clients	UIO, LU, UCPH, UU, NIIFI, UPJS and other NorduGrid partners	worldwide support (anybody who installs it from any repository)
ARC Core	UIO, LU, UCPH, UU, NIIFI, UPJS and other NorduGrid partners	worldwide support (anybody who installs it from any repository)
ARC GridFtp server	UIO, LU, UCPH, UU, NIIFI, UPJS and other NorduGrid partners	worldwide support (anybody who installs it from any repository)
ARC InfoSys	UIO, LU, UCPH, UU, NIIFI, UPJS and other NorduGrid partners	worldwide support (anybody who installs it from any repository)
ARGUS	INFN	Support of components required by the current Italian User Communities and their international collaborative efforts (IGI). For future major developments needs EC funding
	SWITCH	As long as the products' developers remain in SWITCH, and the tools are in active use, SWITCH can commit to "best-effort" support from these developers. That support would not be restricted to specific users

ARGUS-EES	FOM	Development and support provided to communities of interest to NL-NGI or EGI. Critical security fixes provided to any user of the EPEL/Debian repositories
BDII core	CERN	WLCG (but functionality will be reduced wrt current)
BDII site	CERN	WLCG (but functionality will be reduced wrt current)
BDII top	CERN	WLCG (but functionality will be reduced wrt current)
BLAH	CESGA	
	INFN	Support of components required by the current Italian User Communities and their international collaborative efforts (IGI). For future major developments needs EC funding
CEMon	INFN	Support of components required by the current Italian User Communities and their international collaborative efforts (IGI). For future major developments needs EC funding
CREAM LSF module	INFN	Support of components required by the current Italian User Communities and their international collaborative efforts (IGI). For future major developments needs EC funding
CREAM SGE module	CESGA	
	INFN	Support of components required by the current Italian User Communities and their international collaborative efforts (IGI). For future major developments needs EC funding
CREAM Torque module	INFN	Support of components required by the current Italian User Communities and their international collaborative efforts (IGI). For future major developments needs EC funding
CREAM SLURM module	INFN	Support of components required by the current Italian User Communities and their international collaborative efforts (IGI). For future major developments needs EC funding
CREAM	INFN	Support of components required by the current Italian User Communities and their international collaborative efforts (IGI). For future major developments needs EC funding
dCache	DESY	Development and support guaranteed for communities associated with either DESY, FERMIlab or NorduNet (e.g. WLCG, CFEL, XFEL, SNIC, EMBL, New Frontiers …). For other communities 'best effort' can be provided.
Delegation Java	Unsupported	
DPM	CERN	WLCG (expected to be phased out in the future)
	AS	Largest DPM site in the world, discussing about CERN plans to phase it out
	UU	packaging support for Debian in case it is not provided by CERN
EMI-UI	UCPH, NIIFI, LU, UIO, UU, UPJS	joint support with other contributors to ensure compatibility with ARC services
EMI-NAGIOS	CERN	WLCG

	UCPH, UU	joint support to provide probes for ARC services
EMI-WN	INFN	As a joint effort with other Institutes
EMI CANL	CESNET	Support for C API, support for multiple authentication mechanisms
	FOM	Development and support provided to communities of interest to NL-NGI or EGI. Critical security fixes provided to any user of the EPEL/Debian repositories
	UIO, UCPH, UU	Support for C++ API
FTS	CERN	WLCG
GFAL/lcg_util	CERN	WLCG
EMI datalib	UIO	joint support with other contributors to ensure compatibility with ARC services
gLExec-wn	FOM	Development and support provided to communities of interest to NL-NGI or EGI. Critical security fixes provided to any user of the EPEL/Debian repositories
gLite CLUSTER	INFN	Support of components required by the current Italian User Communities and their international collaborative efforts (IGI). For future major developments needs EC funding
gLite-gsoap/gss	CESNET	To be fully integrated within L&B
gLite-proxyrenewal	CESNET	Stable component, expected best effort support
gLite-MPI	CESGA, CSIC	Depends on actual EGI needs
gLite-yaim-core	Unsupported	
gridsite	CESNET	
Hydra	UH	Hydra will live on as part of an “encryption-to-cloud” company. Support to EGI can be negotiated
L&B	CESNET	Funding available from local NGI for own needs
lcg-info-clients	CERN	WLCG
LFC	CERN	WLCG (expected to be phased out in 2013)
	UU	packaging support for Debian in case it is not provided by CERN
RAL-SAGA-SD	Unsupported	
StoRM SE	INFN	Support of components required by the current Italian User Communities and their international collaborative efforts (IGI). For future major developments needs EC funding
TORQUE server config	Unsupported	
TORQUE WN config	Unsupported	
Trustmanager	n/a	Planned to be phased out by the end of EMI in favour of EMI CANL
UNICORE Client6	JUELICH/CINECA	best effort free of charge for academic, scientific and research users, fees and scope for commercial users based on mutual agreements

UNICORE HILA	JUELICH	best effort free of charge for academic, scientific and research users, fees and scope for commercial users based on mutual agreements
UNICORE Gateway6	JUELICH	best effort free of charge for academic, scientific and research users, fees and scope for commercial users based on mutual agreements
UNICORE XUUDB	JUELICH	best effort free of charge for academic, scientific and research users, fees and scope for commercial users based on mutual agreements
UNICORE Registry6	JUELICH	best effort free of charge for academic, scientific and research users, fees and scope for commercial users based on mutual agreements
UNICORE TSI6	JUELICH/CINECA	best effort free of charge for academic, scientific and research users, fees and scope for commercial users based on mutual agreements
UNICORE UVOS	UWAR	best effort free of charge for academic, scientific and research users, fees and scope for commercial users based on mutual agreements
UNICORE/X6	JUELICH/CINECA/TUD	best effort free of charge for academic, scientific and research users, fees and scope for commercial users based on mutual agreements
VOMS	INFN	Support of components required by the current Italian User Communities and their international collaborative efforts (IGI). For future major developments needs EC funding
VOMS-Admin	INFN	Support of components required by the current Italian User Communities and their international collaborative efforts (IGI). For future major developments needs EC funding
WMS	INFN	Support of components required by the current Italian User Communities and their international collaborative efforts (IGI). For future major developments needs EC funding
STS	SWITCH	As long as the products' developers remain in SWITCH, and the tools are in active use, SWITCH can commit to "best-effort" support from these developers. That support would not be restricted to specific users
	UH	Hydra will live on as part of our encryption-to-cloud company. Support to EGI can be negotiated
Pseudonymity	UH	Hydra will live on as part of our encryption-to-cloud company. Support to EGI can be negotiated
EMIR	NIIFI, LU	joint support with other contributors to ensure compatibility with ARC services
	JUELICH	EMIR Server, best effort free of charge for academic, scientific and research users
WNoDES	INFN	Support of components required by the current Italian User Communities and their international collaborative efforts (IGI). For future major developments needs EC funding
EMI messaging layer	CERN	WLCG

Table 3: Long-term support plans at Step 0 (unsupported components are marked)

After the preliminary collection of information, the following gaps have been identified:

Product	Current Institute	Currently proposed mitigation
Delegation Java	UH	This component is currently needed by FTS (CERN), CREAM (INFN) and dCache (DESY). Discussion is on-going among the three partners on how to provide support after the end of EMI
gLite-yaim-core	INFN	The long-term plan is to phase out the proprietary YAIM configuration system in favour of a standard open source method (probably puppet). However, the process will not be completed in time for end of EMI. Discussion will take place in the coming months
RAL-SAGA-SD	STFC	STFC is discussing with the SAGA project about ownership of this component
TORQUE server config	FOM	See glite-yaim-core mitigation action. This component is used to configure the EPEL Torque server with YAIM
TORQUE WN config	FOM	See glite-yaim-core mitigation action. This component is used to configure the EPEL Torque client with YAIM

Table 4: Identified support gaps at Step 0 and proposed mitigations

At the end of the Step 0 phase, all EMI products with the exception of five have an associated owner and a support plan spanning the 12 to 18 months after the end of the EMI project. For the five remaining components mitigation actions are being discussed. In the next step the EMI partners will work with EGI and other stakeholders on identifying any gap between user communities needs and pledged support, which global tasks will be necessary after April 2013 and how to perform them. This activity will be reported during the course of the next year.

4 THE COMMERCIAL DISTRIBUTED SERVICES MARKET

4.1 MARKET DRIVERS

As highlighted in the sustainability plan overview in section 2.5, in order to bring additional innovation into the existing EMI products and expand their usage, an expansion in new markets is necessary. One of the extensions investigated by NA1 and NA3 as part of the sustainability plan is towards the commercialization of services using a number of EMI products.

Since the beginning of the project several discussions were had with commercial companies of varying size and business model. The initial collaboration with major companies like Google and Red Hat did not bring useful results. A more detailed description of the individual collaborations and their outcome is given below. However, some general lessons can be drawn.

EMI products are essentially designed to fit into a distributed service model using a specific infrastructure. The main characteristics of this infrastructure have been presented in the previous chapter. The direct offer of EMI services to end users cannot be considered, since the services cannot be used without an infrastructure where to run them. The only possible market targets for EMI services are entities with a distribution channel and infrastructure that end-users can access.

Major companies like Google and Red Hat have of course the right combination of channel and infrastructure. In addition the EMI products fit with the technological areas where they live. However companies like Google do not easily rely on software written by other parties and have enough resources to develop what they need or buy entire companies to acquire it. In the case of Red Hat, the business model is mainly to sell support licenses. In terms of technology Red Hat already includes in its products a grid, virtualization and messaging using products that are in direct competition with some EMI products.

The strategy has been revised in the second year of EMI. Rather than trying to attract the interest of big companies, the focused was moved on smaller companies with a service provision business model, limited interest in developing their own software and relying on open source technologies. Discussions have been started with the following companies, although concrete results have been obtained only in one case so far (DCore):

SYSFERA

One of the most promising industry contacts at the time of writing is SysFera [R26], a company located in France, offering software as well as providing consulting services. Most notably, it develops a system called SysFera-DS, which is a software product for managing and optimizing heterogeneous and distributed computing infrastructures. The system thus is targeted at similar infrastructure setups as the EMI products currently used in e-Infrastructures or more generally research infrastructures, but SysFera focuses more on the commercial domain.

Discussions at several events such as the ISC2011 and SC2011 lead to closer interactions that aim to explore the complementary use of several EMI products with SysFera-DS in commercial environments. A closer collaboration in the future is also planned by inviting the SysFera members to be strong commercial drivers of the ScienceSoft post EMI activities (see later in chapter 5) complementing the e-Science-oriented EMI consortium with e-Business experience.

Buerhoop GmbH & CoKG

Buerhoop [R27] is a software engineering company providing tailor-made software for mission critical systems in the aviation and logistics sector. An initial contact has been established at ISC 2011 and some interest has been shown to know more about the EMI product suite and its potential commercial

uptake. The Strategic Director is actively trying to organize a technical meeting to discuss the matter further.

Technicolor

Technicolor [R28] is a leading technology company in the media and entertainment sector. First contact was established at ISC 2011. There is potential interest in grid technology as computing platform for video rendering and delivery. The Strategic Director is exploring this further.

DCore Systems SA

DCore Systems SA [R29] is a Luxemburg based holding company, the aim of which is to coordinate the business and technology operations of a network of high-tech companies and academic institutes active in the field of open source software for scientific applications. Its objective is to add value to the original open source software by integrating it into a portfolio of distributed services for health care & medical, financial markets, eGovernment, security, energy, media and education. DCore is a recently created company using a start-up business model. This company is the extension of a company called TheSyrrus, which was already briefly introduced at the end of the first year of EMI.

The business model of DCore Systems matches very well the EMI portfolio of products and its technological vision. The collaboration with DCore has been extended during the second year while the company was finalizing its business model and designing the first set of applications. The overall business model can be easily mapped on the layered market structured used by EMI. Figure 7 shows the market layers for the Commercial Distributed Services market compared to the Research Grid and HPC market. The roles of the entities in the layers are explained below. The collaboration and DCore itself are at the beginning and have a lot to prove, but they show an interesting potential that EMI considers worth exploring.

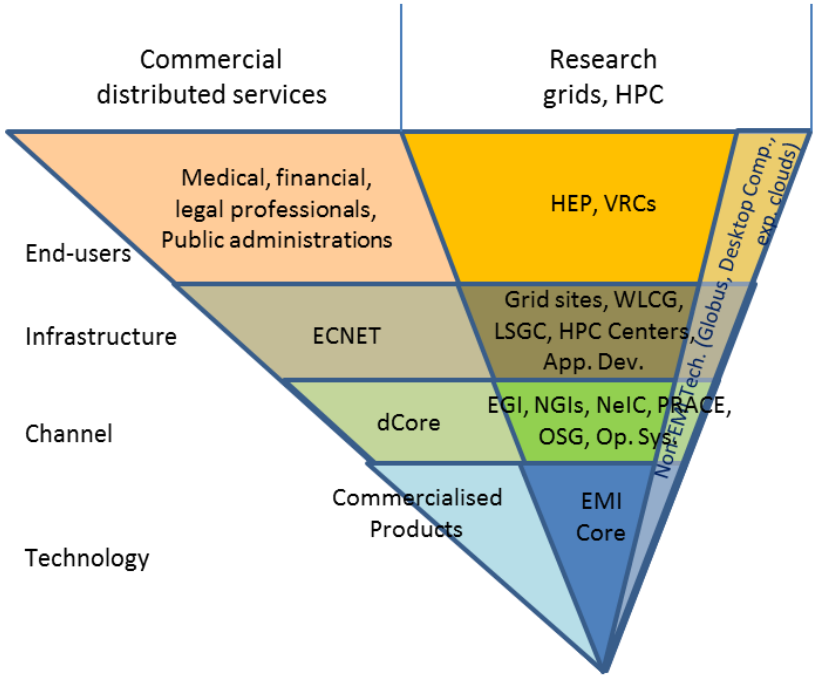

Figure 7: Commercial Distributed Services market

4.2 TECHNOLOGY

The Technology layer is composed of a number of EMI products that have been identified as building blocks of the planned DCore service portfolio. Specifically the products that match with the general architecture are data management components (dCache), encryption (Hydra) and security (VOMS and Argus). In addition, a number of other products from other DCore partners are part of the technology layer.

4.2.1 Collaborations

The business model proposed by DCore is based on creating local partnerships between an academic institutes and commercial companies. The focus of the partnership is a specific piece of technology or an application. The role of the academic institute is to think and develop the software and release it as open source. The role of the company is to take care of the user-facing and commercial aspects, like user-friendly interfaces, marketing and first-level support. The set of institute/company pairs (there can be more than two partners in one partnership, though) provide all the building block of the overall architecture. The advantage is that this minimizes the risks of relying on start-up companies. Redundancy can be introduced and the cost of every pair can be contained. In case of failure, the overall system can still be profitable. The model introduces additional sustainability by sharing the revenues between the institute and the company with locally negotiated agreements.

At the time of writing, two collaborations have been established with EMI partners. A first collaboration was established with DESY to use dCache and a second one with UPJS on virtualization technology. Proposals have been made by DCore to establish similar partnerships with CERN, NIKHEF, INFN and UH. The feasibility and scope will be discussed in the coming months.

4.3 CHANNEL

The Channel layer is provided by DCore itself as holding company. DCore Systems SA is incorporated in Luxembourg and has its main headquarter in Switzerland, close to Geneva. Its technical manager is a former CERN employee with previous experience in grid systems, cloud, and media applications. DCore retains the ownership of the IP of the part of the distributed services developed by the SMEs that it controls and share their revenues. It provides overall technical and commercial coordination. DCore is itself a start-up backed by a private bank and the SCS Group a Swiss-based business consulting firm.

The collaboration with EMI has so far been managed by DCore. Together the EMI and DCore management have discussed the DCore plans and identified potential matches with EMI technologies. Since EMI is not a legal entity, any further collaboration after the initial matching is not established by EMI management, but directly between DCore and interested EMI partners.

4.4 INFRASTRUCTURE

The infrastructure layer is composed by the network of companies and laboratories developing and providing the user-facing services and engaging in technical and marketing coordination activities. The possible form of this network is still being discussed. One possibility being envisaged is for it to take the form of a collaboration program among DCore, the EMI collaboration (in the form it may

assume after the end of the EMI project) and other participating companies and Institutes. Currently the network has been given the codename of ECNET or European Commercial Network.

The goal of ECNET is to manage the collaboration among the participating academic Institutes and commercial companies. This concept has some aspects in common with the open source initiative proposed by EMI, ScienceSoft, although the commercial focus is much stronger. DCore has so far participated to the discussions on ScienceSoft to investigate whether it can be used as channel for ECNET collaboration.

The current visible outcome of ECNET is the design and development of prototypes of a portfolio of distributed applications, accessible from web portals and targeted to professional in the medical, financial and public administration sectors. The following is a list of the high-level specialized applications where EMI middleware products are or are planned to be used (please note that most of this is in the prototype stage, some of the web sites are restricted, and the marketing information may be still fictitious)

Brand Name	Description	Web site
GALA Flexible. Secure. Efficient	GALA is a pioneering archiving system, providing clients worldwide a new and efficient way to archive their documents. GALA offers the flexibility to manage and control the distribution of information from a centralized database. Highly secure end-to-end service that utilizes advanced encryption management techniques to ensure that client information is virtually impenetrable.	http://www.gala-flex.com (requires username and password)
SAFEH	HDM is a Medical data management systems and e-health solutions to provide a safer, more secure and efficient patient care. HDM include new solutions to improve quality of life for senior citizen with low cost and easy to use health care monitoring devices that ensure maximum security and confidentiality while operating smoothly in practice.	www.Safe-h.com (requires username and password)
QPAD	The CubePad system is a highly secure, online project management platform that enables users to work remotely in a secure virtual collaborative space without capacity limitations. Cubepad offers a highly secure range of products and services and is focused on offering the most secure cloud computing platform on the web.	www.cubepad.com
COURTICE	Court-ice is an electronic data management systems and solutions to enable the transfer and streamlining of documentation in an ultra-secure environment for the legal and financial sectors.	www.court-ice.com (requires username and password)

Sinefo	Sinefo bAround is a one stop solution which enables to securely and privately surf the web from mobile devices and tablets. bAround provides a complete mobile solution for companies, business travelers and expats	www.sinefoapplications.com (requires username and password)

Table 5: A list of DCore/ECNET application prototypes

4.5 COMMUNITIES

The expected market for the DCore/ECNET application is composed of professionals in sectors ranging from healthcare/medical, financial, legal, public administrations, education, media, etc. All sectors and the proposed applications are developed around the common theme of secure data management of sensitive information. Essentially the core made of a scalable storage element (dCache), an encryption system (Hydra) and a multi-level authorization manager (Argus, VOMS) is common to most ECNET applications. Domain-specific user interfaces are then added to be efficient and appealing to end-users. The CubePad system is targeted to a more general type of user and can be proposed as a consumer or enterprise-level secure distributed file system (like Dropbox or SkyDrive with added security).

At the time of writing, tests of the prototype systems for Gala and CubePad are being performed. In particular Gala is being proposed as a secure digital archiving system for public administrations. Tools to scan and store paper documents are provided and different level of encryption and access authorization can be attached to each document. Authorization is granted based on VOMS certificates issued by the public administration to its employees. According to information provided by DCore, the system is being tested with an East-European government.

The potential behind the DCore/ECNET system is clearly visible and has direct impact on the sustainability of the EMI software or other software that participating EMI partners may develop as part of the collaboration. The business model allows DCore to exploit the large base of open source software developed by academic institutes and focus on the commercial aspects. It potentially allows the academic institutes to share the generated revenues to fund their research.

However, both the concept and its proposer are in their early stage of implementation and have to confront the challenges of the real market. During the third year EMI will continue the collaboration and where possible will assist in the establishment of fruitful partnerships between EMI institutes and DCore start-ups. In addition EMI will try to understand how the model can be proposed to other similar companies as an example of commercial use of academic open source software.

5 THE OPEN SCIENCE MARKET

5.1 MARKET DRIVERS

During the discussions about what can generate innovation and produce sustainability, EMI contacted many projects, infrastructure, communities and commercial companies' representatives to share problems and ideas. Over the past 12 months it became obvious that most of the people interviewed on the subject share a common set of problems. Some of these problems are more felt by software developers, others are ore felt by users. However, they all have common elements of lack of communication, lack of information, lack of interaction across different projects, different communities, and different scientific disciplines.

In most cases people from different contexts meet and talk in occasion of conferences and other similar events. However the level of interaction is often limited to a presentation followed by a few question. Many issues are briefly discussed, but are not efficiently followed up, recorded, solved.

As far as software development and usage is concerned, there are limited ways of understanding what software exists, who is using it, what they think about it. Despite the fact that all software produced is released under a valid open source license, the structure that animate a community around that software are currently missing in most of the production of scientific software. The cases were software is released via email from a developer to an end-user to be soon forgotten are not rare.

Good quality software has limited chances to be publicly praised by satisfied users and bad software often survives because users have limited ways of clearly saying that it should be stopped.

Many scientific publications describe fundamental results obtained in many different fields and cite past related papers with fully searchable identifiers that increase the academic rating of their authors. However, the software used to generate those results is often mentioned in the text of paper as a note or little more.

The successful open source initiatives do organize conferences and events, but have mechanisms in place by which community members can interact on a daily base to exchange information, discuss problems, find help from other users, and organize ad hoc technical interest groups and so on. Due the traditional structure of limited duration projects typical of the research infrastructures, such mechanisms are very difficult to set up and maintain after the end of the project.

In turn the end-user communities find it difficult to fully commit to rely on the software produced by projects without long-term operational lifetimes.

In order to change what is seen by many as a status-quo, a different approach has to be taken. EMI has started the process of adopting standard open source methodologies and will release by its end date most of its products in standard operating systems repositories. However, the missing element of an active open source community must also be established. Interactions within the community are stronger if supported by motivated individuals within the more general interests of institutes or companies. Of course both developers and users must see advantages in being part of the community, in terms of more visibility, better support, career enhancements, personal technical interests, etc. Institutes in turn have to implement policies to encourage this behaviour and take advantage of the possibility of having themselves more visibility and better tools to prove their value in the research community and therefore increase their chances to be funded to continue providing that value.

In September 2011 EMI started discussing the feasibility of creating an open source community for science with other projects like EGI, StratusLab, OpenAIRE, iMarine, and IGE, SMEs like DCore, Maat, SixSq, SharedObjects, communities like WLCG and LSGC and funding bodies in Europe and the US (NSF).

The general idea of establishing an open source community dedicated to software for scientific applications was understood and appreciated by most people. However, the lack of a precise definition of goals and scope is a limiting factor that has also made many people sceptical of the initiative.

In order to understand more precisely what such an open source initiative should do and how, EMI has started a more formal feasibility study around a concept called *ScienceSoft – Open Software for Open Science*. A group of people from interested parties was created in December 2011 to be the ScienceSoft Steering Committee with the short-term mandate to formalize the discussions about the initiative and produce a document with an initial high-level description of the motivations, issues and possible solutions and a general plan to make it happen. The conclusions of the initial investigation were presented at CERN in February 2012 at a ScienceSoft Workshop organized by NA1 and NA3. The final document is published as an EMI Technical document [R30]. Presentations of ScienceSoft have been made in various occasions, in Amsterdam in January 2012 at the EGI Workshop on Sustainability, in Taipei in February at the ISGC 2012 conference, in Munich in March at the EGI/EMI Conference and at OGF 34 in March.

The main business model is based on the creation of a bottom-up open source community to be established after a period of incubation as legal entity in the form of a not-for-profit organization. Such an entity must be funded as a short-term project or it would defeat its purpose before even starting. However, the initial feasibility study is performed as part of the sustainability activities of EMI, EGI and other projects. The main objective is to define whether the initiative is useful and provide a prototype by the end of the EMI project using contributions in volunteered effort from the general community of developers and users.

The business model can be once again be mapped on the EMI layered market structure for ease of discussion. The full model including the relationships with the other identified markets of Research Grids and HPC and Commercial Distributed Services is shown in

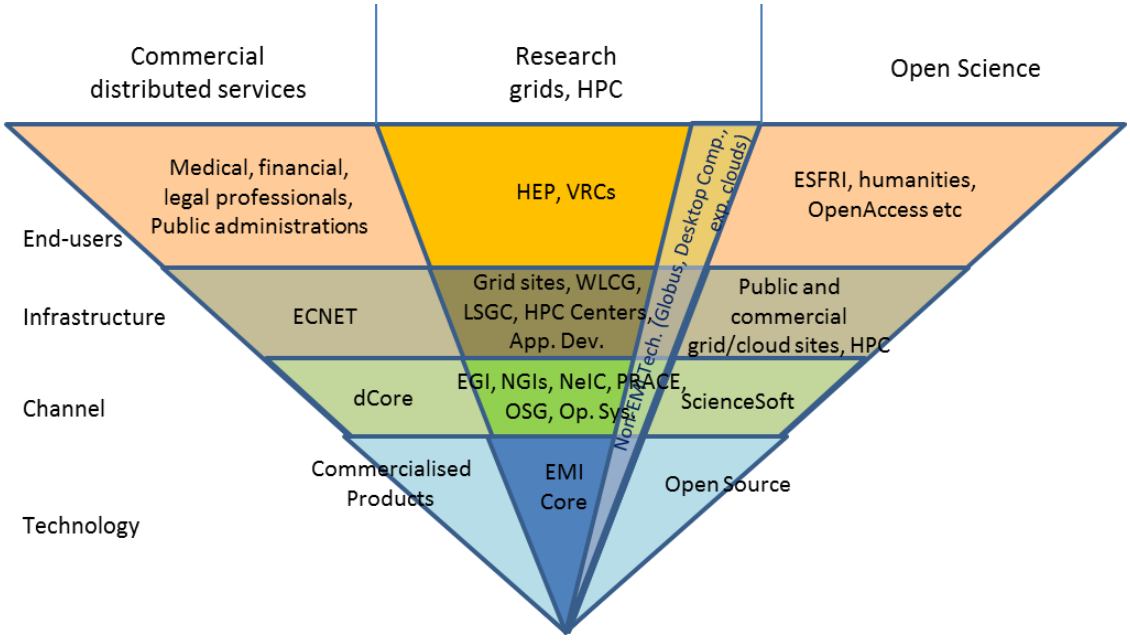

Figure 8: The Open Science market (and its relations with the other EMI markets)

5.2 TECHNOLOGY

The Technology layer is composed of open source software products, including the EMI products, but also any other application, service, tool supporting or enabling scientific research. The common characteristic of all technological products in this layer is their correct open source nature in terms of licensing, source code availability in accessible repositories, openness to contributions, and compliance with standard open source guidelines. It must be noted that this doesn't limit the choice to open source operating systems. Open source software can and is available also for Windows and other proprietary systems.

The EMI products will of course be among the first to participate to this initiative, but it is very important to make clear that ScienceSoft is not in any way a continuation of EMI. This issue was raised several times during the presentations of the Initiative. ScienceSoft must be broader that EMI to work, although the EMI products must play a concrete leading and inspirational role.

5.3 CHANNEL

The Channel in broad terms is provided by the ScienceSoft initiative itself, its portal and the set of collaborative services provided. At the time of writing ScienceSoft is still in the discussion phase and its functionality and scope are not yet fixed. However, a placeholder portal has been created and it's accessible at:

http://sciencesoft.org

The role of ScienceSoft and in particular of its Marketplace is to provide the tools for the community to take shape and interact around the topic of interest. It acts as a real channel between the technology providers offering solutions and the users using them. It has also the role of information collector and publisher about any object that defines the software products, how they are used, where and by whom. It is proposed also as the initial incubator for a more general discussion about using identifiers to uniquely tag the software products and releases along the lines of the DOIs used for papers. A specific pilot initiative is being discussed with the OpenAIRE project.

Figure 9: The ScienceSoft logo and motto

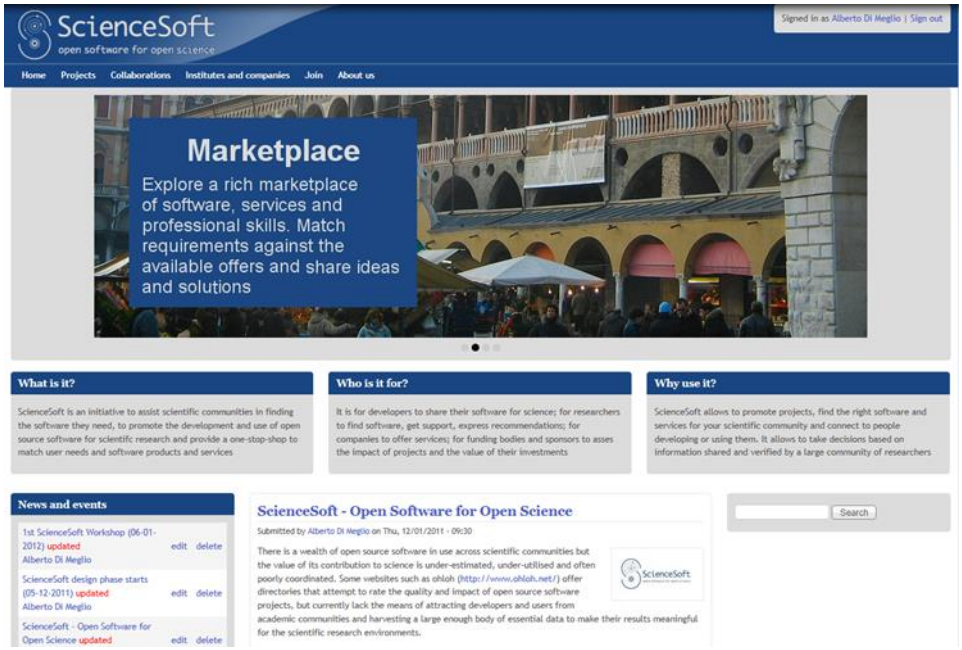

Figure 10: The ScienceSoft portal

Figure 10: Graphical concept of the ScienceSoft Marketplace form the promotional video (http://www.youtube.com/watch?v=M1GJnvEXLlI)

5.4 INFRASTRUCTURE

The Infrastructure layer is composed of any infrastructure used to enable scientific research. In general it encompasses the existing EGI and PRACE infrastructures and includes any existing or foreseeable infrastructures including service grids, desktop grids, private clouds, public clouds, commercial

clouds, etc. Different users community will choose the infrastructure most suitable to their needs and the choice will depend and in many cases already depends on the size, scope, funding availability and level of expertise of the community. Larger, well-organized communities may find more efficient to use public grid or cloud infrastructures and exploit their sharing properties, while smaller, ad-hoc communities may find more efficient to buy computing and storage space for a commercial cloud provider. However the need for finding and running the right software services and applications, getting support and contributing to it is common.

5.5 COMMUNITIES

The ultimate goal of the proposed initiative is to enable the scientific communities at large to efficiently perform their work. The main players in this layer are the major scientific initiatives that are being established in Europe and internationally, the ESFRI projects, communities that so far have not found a perfect match in the grid infrastructures, humanities, social sciences and so on.

The overarching vision is to enable the Horizon 2020 objective of Excellent Science by laying the foundations of a sustainable open software marketplace that future scientific research can rely on without having to ask the usual question: “when does your project end?”

It’s a truly collaborative initiative that can be implemented with minimal, but fair contributions from everyone and that EMI is strongly pushing for.

6 OTHER COLLABORATIONS

A special category of collaborations that fit into the market model in a supporting role is represented by a set of important that are assisting EMI with dissemination, impact evaluation, sustainability planning, etc. The most important collaborations are described below.

e-ScienceTalk

The e-ScienceTalk [R31] project works with EGI-InSPIRE and other collaborating projects to expand the scope of the existing outputs of the GridTalk project, and to report on the interactions of grids with e-Infrastructures such as cloud computing and supercomputing.

EMI signed an MoU with eScienceTalk to formalize joint activities in the following areas:

- Disseminating new applications of relevance through the e-ScienceTalk dissemination channels, such as the GridCafé, GridCast, iSGTW and GridGuide;
- Disseminating the outputs ofEMI in e-ScienceTalk publications;
- Raising the awareness of other potential users and decision makers, through the presentation of the results of EMI, about the usefulness of e-Infrastructures for e-Science.
- Featuring information about the projects on each other's websites and dissemination materials where appropriate;
- Announcing each other's events on websites and in publications where appropriate;
- The e-ScienceTalk Gridcast media team will extend its services to the EMI dissemination team, whenever feasible;
- EMI will provide content and information about the project and applications of relevance to their community to e-ScienceTalk publications such as iSGTW and the policy documents, the GridBriefings;
- EMI will promote e-ScienceTalk's products to its communities and contacts, for example by offering registrants to events the opportunity to subscribe to iSGTW;
- EMI will offer e-ScienceTalk the opportunity to act as media sponsors for events that it organises, including adding the e-ScienceTalk logos and links to e-ScienceTalk products to the event website, in return for promotional activities for the event.

These activities are mostly being fulfilled, in particular by NA2 on the EMI side. For details, please see the NA2 deliverables.

ERINA+ and OSIRIS

The ERINA+ and OSIRIS projects focus on assessment of e-Infrastructures in Europe. ERINA+ aims to evaluate the impact of e-Infrastructure funded projects through the deployment of an effective socio-economic methodology as well as proposing, by the end of the project, a pro-active self-assessment methodology. OSIRIS will pave the way for a platform to enable continuous analysis and recommendations on existing and future European ICT RIs.

EMI signed a MoU with ERINA+ on 08/09/2011. The document calls for an exchange of information whereby EMI provides answers to structured questionnaires submitted by ERINA+, among other things.

OSIRIS has also submitted a questionnaire to EMI. The questionnaire has been filled out, and the EMI project is looking forward to further work with OSIRIS, independently of whether this is formalized by a MoU.

7 ACTION ITEMS FOR THE NEXT PERIOD

7.1 EXPLOITATION ACTIONS

This section summarizes the exploitation-oriented action items targeted for the next project period. Those action items that have neither been completed, nor cancelled in the last period are highlighted with the same numbering as they appear in section 2.4.

#	ACTION	EXPECTED IMPACT
E5	This action item is taken over from year 2. *The expected outcomes here are brief analyses or case studies on these products revealing reasons why they can be appealing to the users. The deadline is basically the end of the next period, but as these components are rolled out gradually in the next year, they will be inspected in a month of time following the release.* This is a continuous task as new product features have appeared in EMI release 2. The expected deadline of this task is April 2013.	Broader usage of EMI component in non-traditional communities
E7	The product team efforts will be monitored further in terms of the factsheet evolution, and also in terms of the commitment of the partners. The expected deadline of this task is March 2013.	Better estimation of development and support costs
E8	The industrial user communities will be engaged through the already established collaborations, with special respect to DCore and SysFera. EMI products will be analysed from industrial utilization perspective. New industrial collaborations will be sought. The expected final deadline of this task is April 2013; nevertheless it is a continuous task that sets milestones for all the all-hands meetings in the next year.	Increased and improved EMI visibility for industrial user communities. Enhanced industrial interest for EMI products.

7.2 SUSTAINABILITY ACTIONS

This section summarizes the sustainability-oriented action items targeted for the next project period. Those action items that have neither been completed, nor cancelled in the last period are highlighted with the same numbering as they appear in section 2.5.

#	ACTION	EXPECTED IMPACT
S6	This action item is taken over from year 2. *… Staff being responsible for sustainability issues must thoroughly follow the work that will be conducted within … the cloud and the trust federations working groups … watching for new potential. This is supposed to be a continuous task. It supports sustainability driver "extension of the user base".* The year 3 periodic deadline is April 2013.	The technical synergies are expected to strengthen especially in the fields of trust federations within the next project period with the first release of the STS service. The presence of strong synergies between EMI products and global AAI activities is expected.
S7	*It is also planned to submit a second questionnaire in the next period that incorporates lessons learnt from the first poll attached as appendices to this document. A suitable schedule for this would be October 2011. It supports sustainability driver "extension of the user base" by understanding the "side-effect-features" of the products more.* A new expected deadline for this action item is July 2013.	The new questionnaire will collect product-related information with respect to E&S, and also product-related commitments of the partners.
S8	This is a successor action item for the former S1 as it has been declared inappropriate. A definition of a set of profiles matching different production usage scenarios will be collected and analysed, thus improving EMI's vision on the demand side. Successful case studies will also be collected and reported with special respect to large "customers", such as PRACE or EGI. This action item is supposed to support sustainability driver *"extension of the user base"*, and is expected to finish in January 2013.	Product portfolios about the most typical EMI usage scenarios will be collected. The most successful installation and configuration use cases will be summarized. EMI release 2 success stories will also be collected.
S9	The project E&S team will offer strong support and commitment for the ScienceSoft initiative to expose E&S success stories, knowledge, know-how, and infrastructure for this inter-project, trans-European collaboration. This action item is supposed to support all sustainability drivers, and is expected to finish in April 2013.	ScienceSoft is expected to become a single point of entry for European scientific projects regarding E&S matters, or any corresponding activities.
S10	The collaborations with the different user communities representing the user demand side for EMI will be strengthened further. A closer collaboration with NRENs has been suggested. This action item is supposed to support sustainability driver *"extension of the user*	Strengthened collaboration, increased, and improved user base, and possible embedding into the NRENs service profiles. The motivation is explained in the following section. This item is being currently discussed within the sustainability working group

	base", and is expected to finish in April 2013.	

NRENs

National Research and Education Network service providers (NRENs) formulate a strong organizational network throughout Europe. Their primary objective is to provide data network services, as well as corresponding facilities to the higher education, and to research organizations. The service portfolio of NRENs shows a broad variety: some of them offer HPC services, but most of them do not. Nevertheless most of the NRENs are particularly interested in operating cloud and distributed storage services to promote those customers who cannot afford large self-owned, and self-operated storage devices. The storage service portfolio is also broad; it ranges from remote block devices to high-level backup/archive applications.

There is a European organization, called TERENA, that incorporates all the European NRENs, and who runs thematic work groups to aid information dissemination within the European service providers. One of the affected workgroups is the task force storage, TF-Storage that is founded in order that all storage infrastructure development and operations related information items, best practices, and success stories are distributed, and shared among the interested parties. These workgroups are particularly interested in not only low-level storage operating platforms/protocols, such as iSCSI, or shared file systems, but also in more application-oriented services, such as distributed storage management.

In the new action items it is highly recommended to contact this task force and conduct discussions about the exact storage-related demands [R32], [R33].

8 CONCLUSIONS

In this deliverable the second year's progress on the exploitation, sustainability and collaboration activities are summarized. This document uses a revised structure compared to the version created last year, due to the structural, and the task-assignment changes introduced at the beginning of the year.

The exploitation and sustainability action items set out in the previous DNA2.4.2 are summarized and evaluated against the corresponding activities of this year. The focus of the content has been shifted towards a more market-oriented approach, where the market structure, the EMI position within it and the market exploitation possibilities of the EMI software stack is examined.

The main market for EMI is by definition provided by the Research Grid infrastructures using either HTC or HPC technologies. Most of the effort of EMI in supporting and evolving the middleware services goes into this market. However, in order to push for further innovation and fulfil the future needs of Horizon 2020 in both Excellent Science and Competitive Industry, ERMI is actively exploring commercial distributed services and open science markets.

On the one hand dCore initiative helps assessing the value of EMI components in more commercial activities. On the other hand ScienceSoft opens the path for new, joint communication, exploitation and sustainability efforts shared among a larger base of European scientific initiatives. The main objective is to provide the model for a uniform policies and communication channels primary targeted at the scientific research user base.

The long-term partner commitment in the EMI products and services are been evaluated and formalized. The current status is highlighted in this deliverable and the results of the on-going discussions with major stakeholders like EGI, WLCG and PRACE will be reported in the coming months.

The deliverable also gives a summary on the process achievements, such as the establishment of a stronger E&S team within the project, the appointment of a Strategic Director to lead the E&S activities, a better organization of the meetings and discussions including regular dedicated face-to-face sessions at the project All-Hands meetings.

On the solid foundation of the significant progress made in year 2, the E&S activities are expected to intensify even further in year 3, the last year of the project. The methodologies and tools worked out so far will be applied to the three identified markets with a clear focus on production use. A set of procedural actions have been defined and will provide the background for the work to be done in the coming months.

9 REFERENCES

R1	http://www.globus.org
R2	Specific deliverable on the SLA definition and implementation http://cdsweb.cern.ch/record/1277519/files/EMI-DNA1.2.2-1277519-SLA_Template-Rev.1-v1.0.pdf
R3	EMI Standardization Workplan http://cdsweb.cern.ch/record/1277526/files/EMI-DJRA1.5.1-1277526-Standardization_Workplan-v1.0.pdf
R4	EU-IndiaGrid http://www.euindiagrid.eu
R5	SA2 QA policies: https://twiki.cern.ch/twiki/bin/view/EMI/SA2
R6	EGI Document 1059-v3 https://documents.egi.eu/document/1059
R7	gCube software http://www.gcube-system.org
R8	GOCDB http://goc.egi.eu
R9	EMI User Community Forum http://cf2012.egi.eu
R10	EUDAT http://www.eudat.eu
R11	GOCDB https://goc.egi.eu/portal
R12	EGI CF 2012 http://cf2012.egi.eu
R13	SHIWA http://www.shiwa-workflow.eu
R14	Savannah Task 22854 https://savannah.cern.ch/task/?22854
R15	EMI release http://www.eu-emi.eu/emi-2-matterhorn-products
R16	OSI http://opensource.org
R17	https://cdsweb.cern.ch/record/1450862/files/EMI-DOC-WP03-dCacheLicenseAgreement-v1.0.pdf?subformat=pdfa
R18	GNU Affero General Public License (AGPL) Version 3 http://www.gnu.org/licenses/agpl-3.0.html
R19	EGI Infrastructure https://documents.egi.eu/document/1059
R20	https://www.egi.eu/indico/getFile.py/access?contribId=6&resId=0&materialId=2&confId=719
R21	HP-SEE Project http://www.hp-see.eu
R22	EGIIS www.nordugrid.org
R23	http://www.nordugrid.org/monitor/loadmon.php?display=vo=Hungary
R24	https://documents.egi.eu/document/1059
R25	EGI User Forum in Munich in March 2012 https://indico.egi.eu/indico/contributionDisplay.py?sessionId=57&contribId=69&confId=679
R26	SysFera http://www.sysfera.fr
R27	Buerhoop http://www.buerhoop.de
R28	Technicolor http://www.technicolor.com
R29	DCore Systems SA http://www.dcore-sa.com
R30	ScienceSoft Overview, http://cdsweb.cern.ch/record/1451518/files/sciencesoft_v1.0.pdf?subformat=pdfa
R31	e-ScienceTalk www.e-sciencetalk.eu

R32	TERENA http://www.terena.org
R33	TERENA http://www.terena.org/activities/tf-storage
R34	https://twiki.cern.ch/twiki/bin/view/EMI/DeliverableDNA321
R35	http://desktopgridfederation.org/infrastructures1
R36	CHAIN, http://www.chain-project.eu
R37	FutureGrid, https://portal.futuregrid.org/
R38	UNICORE on FutureGrid, https://portal.futuregrid.org/manual/unicore
R39	Fedora/EPEL, http://fedoraproject.org/wiki/EPEL
R40	Debian, http://www.debian.org

APPENDIX A: THE SCIENCESOFT OVERVIEW DOCUMENT

ScienceSoft: Open Software for Open Science

An Open Community of People, Software and Services for Scientific Research

v. 1.0

Authors

The ScienceSoft Steering Committee

Geneva, 29 February 2012

Table of Contents

Executive summary 2
History 3
Market Analysis 3
Identified Problems 5
Possible solutions and benefits 7
Functionality 7
Benefits 9
State-of-the-Art 9
ScienceSoft Organization, Structure, Operations 11
Current Activities and Timeline 12

Executive summary

There is a wealth of open source software in use across scientific communities but the value of its contribution to science is under-estimated, under-utilised and often poorly coordinated. Some websites such as ohloh (http://www.ohloh.net/) offer directories that attempt to rate the quality and impact of open source software projects, but currently lack the means of attracting developers and users from academic communities and harvesting a large enough body of essential data to make their results meaningful for the scientific research environments. Being able to aggregate the power of cataloguing services, trends and statistics would provide a sound basis for judging the popularity of specific software, enable social-networking amongst users and developers, create active communities and promote citizen science. Rating software and providing a means by which it can be cited in a similar manner to publications and datasets would enable the authors to gain merit and career advancement for their work and accelerate the open source software movement in scientific communities. Being able to quantify the impact of open source software would allow funding agencies, companies and venture capitalists to better target their investments leading to a more vibrant and sustainable open source market for open science.

This document outlines the state-of-the-art of open source activities in the scientific research environments and describes a number of problems and potential solutions identified by discussing with user and developers of existing projects and communities. The potential benefits of the proposed solutions are described and how scientists, developer, administrators, managers and funding bodies could exploit them to take decisions and plan their activities.

History

During September 2011 the EMI project started a discussion on a general proposal to investigate, design and establish an open source software initiative as part of the EMI long-term sustainability plans. The main objective of this proposal was initially to create the conditions for the continuing development, support and use of the EMI software products after the end of the EMI project by establishing a broad open source community of developers and users.

However, it became rapidly clear from discussing with users and developers from scientific communities and research projects that similar concerns about long-term sustainability were shared by many other projects developing or using middleware, applications, tools and related services of critical interest for the European scientific research communities.

The discussion on establishing a broad open source community around software for scientific research was therefore extended to a number of interested parties. The goal is to define and setup a truly open community of software developers, administrators and users in the context of global scientific research by including scientists, developers, service providers, research institutes and commercial companies in the process since its inception.

In December 2011, the ScienceSoft Steering Committee (SC) was formed with participants from a number of projects, communities and SMEs. The mandate of this group was to bring into the discussion their particular perspective and refine the initial concept and ideas. On February 8th, 2012 most SC members and a number of additional interested participants have attended a workshop at CERN (https://indico.cern.ch/conferenceDisplay.py?ovw=True&confId=160503) where the initial definitions of problems, solutions and possible next steps have been discussed.

This document summarizes the work done since September 2011 with particular reference to the outcome of the CERN Workshop.

Market Analysis

As a starting point for the discussion, a brief market analysis was performed. This analysis is probably biased towards current research grids more than other types of scientific computing and data environments. However it is general enough to be easily adapted and extended as necessary.

Five types of computing and data environments have been considered:

1) Grid (High-Throughput) computing
2) High-Performance Computing
3) Cloud computing
4) Desktop or volunteer computing
5) Stand-alone computing

Combination of these basic types can be considered as well, like for example the provision of grid services using dynamically provisioned IaaS clouds or the provision of cloud-like access to grid resources.

The software layers involved in the use of the selected types of environments include:

1) Operating system software
2) Middleware software
3) Service management software
4) Applications and community specific services
5) User interfaces and portals
6) Tests, tools, benchmarks and other supporting software

The roles involved in the use of the software can further be classified as:

1) Developers
2) Users
3) Service providers
4) System engineers, platform integrators
5) Institutes and Companies
6) Collaborations (projects, initiatives, communities, virtual organizations, etc.)
7) Funding bodies

The roles are not mutually exclusive. On the contrary the complex relationship among people in various roles is indeed one of the fundamental issues that ScienceSoft proposes to investigate. For example the developer of a software product can be at the same time the user of another product (a

dependency) or a service (a testing service or a cloud IaaS service) and a provider of applications for a group of scientific researchers.

In this classification, the different roles have different problems and are looking for different specific solutions. A consistent proposal must be able to take into account and express the diversity across the market (vertical dimension), while at the same time recognising the commonalities across software layers (horizontal dimension). The distinction between vertical and horizontal dimensions will be further explored later on in this document.

Figure 1 illustrates the previously defined categories and their relationships.

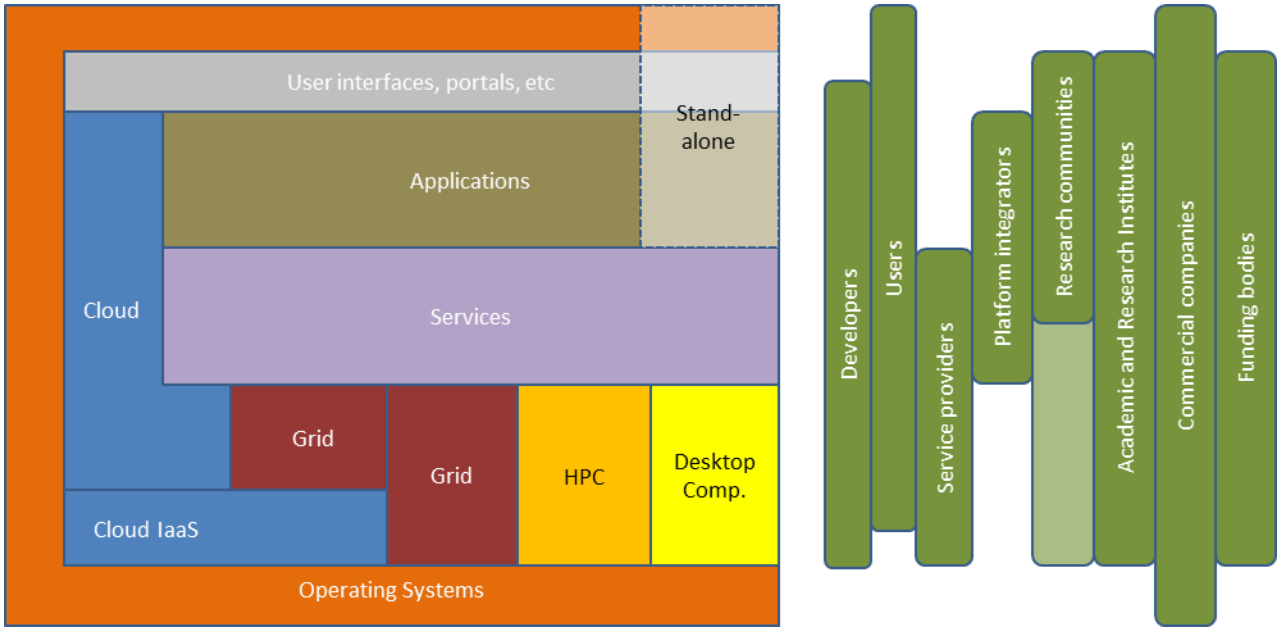

Figure 1: The software and services market classification

Identified Problems

Users and developers of software for scientific research projects have been contacted during conferences and presentations and asked about what problems they are confronted with in their normal working activities. Among the identified problems the following are considered the most critical:

- Lack of continuity in support, development, coordination of software: this issue is mostly felt by users and developers of software developed by short-term projects. Users face difficulties in adopting or relying on software without longer term support commitments. At the same time developers, who may be willing to continue developing and supporting the software, find themselves without a repository or release channels and have no easy way of keeping contact with their users.

- Non-optimal communication between users and developers: most projects tend to be rather horizontal and focus on specific, limited aspects of the scientific research software stack. Although this is a quite natural approach to setting up projects, it has the drawback of creating boundaries between different categories of users. The interaction happens often in occasion of conferences or other public events, but many users have expressed the need for a more direct and continuous relation with the developers in order to discuss requirements, features, issues, etc. It is felt that there is excessive formalization and loss of information due to being in different projects.

- Lack of consistent real usage information: although it is relatively easy to count downloads from a web site, information on actual usage is much more difficult to collect. In a number of cases, information systems are available to provide this data, but not for all potential services and applications. The lack of this information prevents understanding the actual user base and evaluating the relative importance and impact of the produced software

- Limited access to other users' experience: before using an application or a software service and investing time and money in them, users would like to know what other people think of those applications or services, what problems have been found, if and how they have been solved. Problems solved in one case can be applied in other cases and save time for both users and support people

- Limited or complex ways of finding what exists already: a lot of duplication is often found in the software production. In many cases the duplication is due to real needs of producing something that works better or fits better a community need, but quite often it is due to the lack of knowledge of what exist already. Having information about existing software and services, who develops and supports them and who uses them may help taking informed decisions and quickly bootstrap new projects without waste of time and resources

- No way of influencing the production of software: many users feel that they have little saying on what functionality software should have or what software should be supported and which should be stopped. Although functionality and quality are always a matter of negotiation and priorities, the lack of the possibility of expressing opinions often drives users away or prompt them to develop their own solutions in the hope of having more control

- Lack of visibility of the software activities: many software developers working in research institutes produce fundamental tools used by scientists to make important discoveries. Academic recognition is usually measured in terms of publications and the contributions of programs and tools cannot be formally acknowledged with a reliable citation system. Having formal ways of citing and listing software used in conjunction with published scientific results would increase motivation and help career recognition and development

- No way of assessing the user "market": the main of foremost barrier to the involvement of commercial ventures into the development and support of software and services for scientific research is the difficulty of sizing the market and the potential revenue streams. Having ways of performing realistic market analysis and offering targeted services to the scientific communities may help SMEs if not larger companies to define achievable business plans

Most of the software developed today by research institutes, university, research projects, etc. is typically stored in local source and binary repositories and readily available for the duration of the project lifetime only. Finding software based on given functional characteristics or field of application is very difficult especially for new projects or young researchers. Binaries to be run on the most used operating systems are available from many different places ranging from local university repositories to mainstream community repositories like EPEL or Debian. Cases of conflicts are often found between different versions distributed by different people from different places. Source code is even more difficult to locate and access and contributing with comments, patches and fixes, which is a very common activity in the open source world, is traditionally very difficult to do in the research communities. This has been for years a primary complaint from users.

Similar requirements have been expressed by application developers, infrastructure managers and users. Within the HPC community the organization of common repositories of application code is a known concern. Although such code is usually highly dependent on the hardware on which they are run, the lack of code sharing and availability is considered one of the reasons why the European HPC efforts are falling behind similar activities in the US and Asia (see the recent IDC report presented at the EGI Technical Forum in Lyon in September 2011).

Most of the reported problems can be categorized as a lack of consistent and transparent information about the software being used in scientific research. The problem is not necessarily a lack of technical information (such as documentation or user guides, although this has also been described as a problem in many cases), but rather a lack of metadata.

Information about who develops, contributes and uses a given program is very difficult to find out and yet the widespread availability of such information would give more visibility and credibility to the software products. In addition, the EC invests considerable amounts of money into funding projects that directly or indirectly need to develop software. A single repository of metadata information about software products would allow projects to avoid re-developing existing solutions and would provide valuable statistics about software usage. Such information could be used also by the EC to monitor the outcome and impact of funded projects, the extension of adoption of open source software and the compliance with OSI licenses and possible as input to future EC calls objectives and framework programmes. In the same way, the information could give more strength and credibility to project proposals, which could be backed by realistic information about usage, impact and exploitation of the software.

Possible solutions and benefits

Functionality

From the discussions with users and developers and the outcome of the CERN Workshop in February, a number of desirable functionality has been defined:

- Software, services and people catalogues: the first and foremost desired functionality is the provision of catalogues of information about software products, software-related services and people. The catalogues should provide information based on tags or taxonomies that allow to group together sets of related products, services and people based on flexible search criteria. The provision of technical metrics about software and services is also desired, for example license, programming languages, compatibility, supported systems, etc.
- Generation of statistics: the information collected and processed should not only be used to search about software, but also to general relevant and useful statistics. The most requested statistics are for example actual usage information, as opposed to downloads of packages from a web site, geographical distribution of usage and production, involvement of Institutes and Companies in projects based on scientific discipline, etc.
- Honour system: community users should be able to rate the registered software and services based on their experience. Ratings can be provided based on predefined categories such as reliability, support quality, documentation, ease-of-use, standards support, etc. Ratings should be supported by comments describing the user experience with the software or service
- Citation system to allow software to be referenced in papers: registered software should receive some sort of unique identifier like the DOIs used for papers, so they can be reliably cited in scientific publications.
- Marketplace for products, services, and people: this is one of the most interesting features and one that may well define a community. Matching demand and offer of software products, service and people skills should be enabled based on the catalogues maintained by the community tools. Some form of advertising could be considered both for non-profit and for profit activities, although different models should be envisaged in this case
- Links to technical services: one of the marketplace-related activities is the provision of technical services. A range of services could be designed and provided by community members and

provided to other members for free or for a fee depending on conditions. Such services could go from straight access to IaaS resources, to software testing services, to consultancy and support, etc. Services can be generic or community-specific. The common requirement is to provide a way to advertise, promote and access the services

- Platform integration support: using the collected information and the product catalogues, it should be possible to define community-specific software stacks to be supported by platform integrators. These community-specific profiles or stacks can then be pre-packaged and easily deployed using the more and more standard virtualization and cloud technologies

- Support for creation of ad-hoc communities and groups: this requirement is what may actually define ScienceSoft as a community-enabler. ScienceSoft per se should act a super-community provide a framework and tools to enable more specific communities to interact without being isolated from other communities. Cross-community groups, for example for common activities like security, configuration, etc. could also be created and populated with the relevant products and people

- Coordination, collaboration and discussion tools: collaboration tools typical of distributed online infrastructures should be available through the ScienceSoft community portal. It should be possible for each hosted community to have independent channels, but a general social network for science should be supported allowing direct interaction among users in addition to standard forums

- Support for organization of technical events: it should be possible to advertise and manage technical events related to the hosted communities or projects. Support can range from dedicated event pages, to agendas, advertising, collection of material, etc.

Benefits

The creation of links or relationships not only among pieces of software, but equally among the people interacting with the software, would foster a more active community and create the conditions for sharing ideas and skills and a more rapid improvements of the software quality. The use of modern social networking techniques would greatly help the establishment of active open source communities and focused sub-communities around specific scientific and technical interests. For example, sub-communities could be established for people interested in testing or in writing documentation, communities of packagers, or experts of specific standards or security and so on.

The possibility of sizing and profiling the usage of software by user communities would potentially allow commercial companies to offer added-value services based on concrete needs.

The availability of value-added information about software and its usage would bring several benefits. It would allow Institutes to perform more realistic assessment of costs and optimize resources by focusing on unique propositions rather than duplicating existing software. The possibility of concretely sizing the user base of an application or a service would provide tangible supporting evidence for funding requests with concrete impact analysis. Open source software licenses adoption and compliance could be verified in order to enforce legal requirements.

The establishment of a software rating system based on both technical criteria (Is a given platform supported? Is a certain package format available?) and usability criteria (What do existing users think of it? Is documentation up-to-date and well written?) would allow filtering mature products from less mature products and would increase the developers' motivation to improve certain aspects of their software like documentation that traditionally receive less effort that the actual code writing.

An interesting functionality is the possibility of creating community-specific virtual software stacks using the software catalogues. Once the profile is defined, it can be kept updated as the products evolve and dedicated community integrator could provide pre-packaged appliances to be shared for example through appliance marketplaces like the one implemented by StratusLab.

State-of-the-Art

The first approach being considered to address the described issues and provide the desired functionality is to exploit lessons learned from successful open source software communities.

Most of the software used in scientific research and developed by academic institutes is generically "open source" in the sense that it uses some type of OSI license. However, it takes more that source code and a license to have a "community". The general definition of an open source software community is a group of developers and users interacting to produce free-software. The interaction among users and developers, the sharing of resources and common objectives and the benefits deriving from sharing are of course fundamental to have a community and not just software in a repository.

We can distinguish primarily between four different types of open source communities:

- **Technology-specific (or horizontal) projects**: this type of communities includes projects focused around a specific technology or framework which all members contribute to. Usually the membership rules are quite stringent both in technological and legal terms. For example it's not uncommon to have to adopt a mandated IP model and a license for all contributing products. Notable examples of this category are the Apache Foundation, the Eclipse Foundation or the Drupal Association

- **Operating system distributions**: communities focused around different flavours of Linux operating systems have been among the first to emerge and have in many cases enabled most profitable open source business models. Although in general there is no formal membership into these communities, the engagement rules to contribute are quite strict and require a peer-review level of competence and quality for both contributors and products. Most notable examples of this category are Fedora, EPEL, Debian, CentOS, etc.

- **Services and tools**: these open source communities usually provide a software application and often services based on that application. They have dual usage models, whereby access to the service is free for personal or non-commercial use, while professional use is charged a fee. Most notable examples include SourceForge, GitHub, Zarafa and many others.

- **End-to-end (or vertical) open source communities**: at the end of 2011 Andrew Aitken, president of the Olliance Group, a leading open source consulting firms, wrote an article about the appearance of "super-communities" or communities of communities. The super-communities instead of focusing on a particular piece of open source technology are built around the entire end-to-end supply chain of an industrial sector, like the aerospace industry (Polarsys launched by Airbus), stock exchange management (OpenMama launched by the New York Stock Exchange) or OSEHRA (electronic healthcare records launched by the US Department of Veteran Affairs). The Olliance Group predits that this kind of communities will rapidly increase in number. Microsoft launched in 2011 the OuterCurve community to host open source software and communities and provide general-purpose IP management services with a focus on Windows applications. In the scientific communities, portal like NanoHub or CyberSKA focus on the general needs of the nano technologies and radio astronomy communities respectively.

The software produced and used in scientific applications is by its nature very diverse. It uses different programming models, technologies, IP and licensing models. In addition the end users, the scientists using the software to perform their research are not overly interest in what technologies are used under the hood, but are very concerned with having a working set of tools. The first three types of open

source communities described about could therefore be used for parts of the software produced in the academic world, but would t bring the level of communication and organization needed to provide the functionality described earlier. A suitable model for ScienceSoft could therefore be the fourth one, where the overall end-to-end software needs of specific scientific communities could be modelled and addressed with a more global approach than just individual pieces of software.

ScienceSoft Organization, Structure, Operations

Based on the preceding analysis the ScienceSoft super-community could therefore be configured as a community of users and developers of software targeted at scientific applications. The community members contribute software and information and provide services to other members. The members within ScienceSoft are organized in focused communities or collaborations around a specific scientific or research topic. People, software, services, etc. are tagged within the super-communities based on their particular focus in order to build community-specific sets of resources. Resource tagging is not exclusive. The same resource can be tagged with more than one community focus, so that it becomes possible to understand the overall usage of that resource within a more general scientific context.

A community portal gives access to the community services like member and product registration and management and the different functionality described earlier in the document.

Most of the common base functionality required to operate the ScienceSoft portal already exists in some form. As a first implementation, the ScienceSoft portal can be configured as an aggregation of such functionality within a coherent container. Functionality like software inventories, source-code repositories, social networking, forums, etc., can be provided in this way using existing open source services like ohloh.net or inventories, Drupal modules for the web based collaboration tools and existing social networks like FaceBook, Linkedin or Google+ for user management.

The community specific services would instead be provided by the members through links or applications running in the portal. Community-specific micro-sites can be easily established from common templates to create well-defined identities and focused sets of people, programs, services, tools, information, etc.

The governance structure for ScienceSoft could therefore be based on two general roles:

- **ScienceSoft maintainers**: the maintainers are in charge of implementing and supporting the portal and the base common features by integrating as much as possible existing functionality from other compatible open source initiatives, where compatible means that the license and usage rights must allow the integration within the ScienceSoft portal. The maintainers could also provide support for general items like community templates, although it can be envisaged that such functionality would be also contributed by other community members
- **ScienceSoft contributors**: any ScienceSoft community member providing information, software or services to ScienceSoft or one or more communities hosted within ScienceSoft. This category includes the people responsible to defined and maintain collaborations or communities within ScienceSoft. For example the EMI Collaboration, the Grid and Cloud Security Group, the Earthquake Modelling and Prediction project could be hosted communities with one or more person managing their definition, memberships, activities, software stacks, etc.

Although the initial organization and governance structure of ScienceSoft should be as lean and lightweight as possible, it is foreseen that depending on the success and evolution of its activities it could become in the future a Not-for-Profit Foundation on the model of existing successful open source foundation.

Current Activities and Timeline

The requirements and possible implementation strategies for Science Soft are currently being investigated by the ScienceSoft Steering Committee. Initial requirements and desired functionality have been discussed in occasion of the ScienceSoft Workshop held at CERN in February 2012. The outcome of the workshop has been incorporated in this document. The workshop participants have agreed to keep providing feedback and collaborate in the definition of implementation priorities. The proposed timeline is as follows:

- **March to June 2012**: definition of priorities and identification of volunteers ScienceSoft maintainers
- **July to December 2012**: progressive implementation of a prototype community portal, dissemination, engagement of scientific communities in trying the functionality and providing feedback
- **January to April 2013**: start of the regular activities, further requirements and implementation cycles. Until this date the community is incubated with the EMI project, which provides overall coordination
- **May 2013 onward**: regular operations, fund raising for continuing activities based on the success of the initiative. Phase down and discontinuation if no interest has emerged.

The Steering Committee

Surname	Name	Project	Institute/Company
BEGIN	Marc-Elian	StratusLab	SixSq
BRYCE	Ciaran		SharedObjects SA
CASTELLI	Donatella	iMarine, OpenAIRE	CNR
CECCANTI	Andrea	EMI	INFN
DI MEGLIO	Alberto	EMI	CERN
ESTRELLA	Florida	EMI	CERN
KONYA	Balazs	EMI	Lund University
MANSET	David		Maatg France
MIKULIC	Marko	OpenAIRE	
NEWHOUSE	Steven	EGI-InSPIRE	EGI
RIEDEL	Morris	EMI	FZJ