Project acronym	D4Science-II
Project full title	Data infrastructure ecosystem for science
Project No	239019

Knowledge Ecosystem Operation Report

Deliverable No DSA1.2b

September 2011

SEVENTH FRAMEWORK PROGRAMME
Research Infrastructures

INFRA-2008-1.2.2: Scientific Data Infrastructures

DOCUMENT INFORMATION

Project	
Project acronym:	D4Science-II
Project full title:	Data Infrastructures Ecosystem for Science
Project start:	1st October 2009
Project duration:	24 months
Call:	INFRA-2008-1.2.2: Scientific Data Infrastructures
Grant agreement no.:	239019
Document	
Deliverable number:	DSA1.2b
Deliverable title:	Knowledge Ecosystem Operation Report
Contractual Date of Delivery:	September 2011
Actual Date of Delivery:	14 October 2011
Editor(s):	A. Manzi
Author(s):	L. Candela, A. Manzi, P. Pagano
Reviewer(s):	A. Ellenbroek
Participant(s):	CERN, CNR
Work package no.:	SA1
Work package title:	Knowledge Ecosystem Operation
Work package leader:	CERN
Work package participants:	CERN, CNR, ENG, FAO, FIN, NKUA
Est. Person-months:	3
Distribution:	Public
Nature:	Report
Version/Revision:	1.0
Draft/Final	Final
Total number of pages: (including cover)	28
Keywords:	Ecosystem, Infrastructure, VREs, Status

CHANGE LOG

Reason for change	Issue	Revision	Date
Initial table of contents	0	1	9 Sept
First version	0	2	25 Sept
Second version with Reviewr comments addressed	0	3	3 Oct
Third version with CNR contribution integrated	0	4	14 Oct
Forth version with statistics upgraded	0	5	14 Oct

DISCLAIMER

This document contains description of the D4Science and D4Science-II project findings, work and products. Certain parts of it might be under partner Intellectual Property Right (IPR) rules so, prior to using its content please contact the consortium head for approval.
E-mail: info@d4science-ii.research-infrastructures.eu

In case you believe that this document harms in any way IPR held by you as a person or as a representative of an entity, please do notify us immediately.

The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and publication of this document hold any sort of responsibility that might occur as a result of using its content.

This publication has been produced with the assistance of the European Union. The content of this publication is the sole responsibility of D4Science-II consortium and can in no way be taken to reflect the views of the European Union.

The European Union is established in accordance with the Treaty on European Union (Maastricht). There are currently 27 Member States of the Union. It is based on the European Communities and the member states cooperation in the fields of Common Foreign and Security Policy and Justice and Home Affairs. The five main institutions of the European Union are the European Parliament, the Council of Ministers, the European Commission, the Court of Justice and the Court of Auditors. (http://europa.eu.int/)

D4Science-II is a project partially funded by the European Union

LIST OF ABBREVIATIONS

D4Science-II	Data infrastructure ecosystem for science
DoW	Description of Work
EC	European Commission
EM	Environmental Monitoring
FARM	Fishery and Aquaculture Resource Management
gHN	gCube Hosting Node
HDFS	Hadoop's Distributed File System
IS	Information System
JRA	Joint Research Activity
MS	Messaging System
NA	Network Activity
NGI	National Grid Initiative
ROC	Regional Operations Centres
SA	Service Activity
VO	Virtual Organisation
VRE	Virtual Research Environment
WSRF	Web Services Resource Framework

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

SUMMARY

This deliverable reports the activities carried out by the D4Science-II Service Activity (SA) to deploy and maintain the project production ecosystem. This ecosystem can be defined as the set of hardware, software, data collections, and procedures that together provide reliable collaboration environments to the D4Science-II user communities.

This document documents the work carried out during the second year of the project (M13–M24) to deploy and maintain the production ecosystem. This includes: (1) the description of the infrastructure, sites, and nodes allocated to the ecosystem, (2) the results of applying the procedures defined for deployment, certification, monitoring, accounting, support, and finally (3) the description of the environments deployed to satisfy the requirements of the project user communities.

EXECUTIVE SUMMARY

The objective of the D4Science-II Service Activity is to make available and maintain a distributed ecosystem to support the activities of the project's user communities. These communities access the ecosystem via the exploitation of Virtual Research Environments (VREs) that satisfy concrete needs of the communities' Virtual Organizations (VOs).

This report documents the work done to maintain the D4Science ecosystem during the second year of the D4Science-II project. It describes:

- The different infrastructures that compose the ecosystem,
- An update of the activities carried out to manage the infrastructure nodes with respect to the first year,
- An update of the activities carried out for monitoring and accounting of the infrastructure,
- The activities carried out to provide efficient support to the ecosystem users,
- The status of the VOs and VREs deployed to satisfy the D4Science-II scientific scenarios.

The D4Science ecosystem is composed of many infrastructures in which the D4Science infrastructure is the most important, and mandatory for the overall orchestration of the ecosystem. Different from the other infrastructures whose resources are managed by external organizations, the D4Science infrastructure resources are managed by the project partners. As a consequence this report describes the ecosystem in general but provides detailed information only for the D4Science infrastructure itself.

As already described in the first version of the deliverable (DSA1.2a) the scientific scenarios served by the ecosystems are: INSPIRE, DRIVER, FCPPS, AquaMaps, and ICIS while the infrastructures that compose the ecosystem are: D4Science, EGEE (EGI), AquaMaps, GENESI-DR, DRIVER, and INSPIRE.

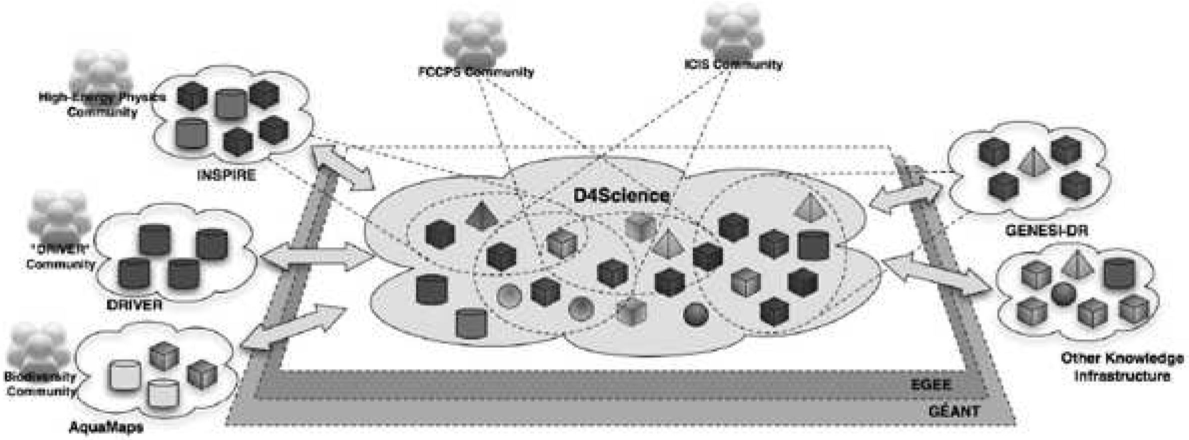

Figure 1 - D4Science ecosystem scenarios and infrastructures

From the Service Activity point of view, and as defined in the project Description of Work, the D4Science ecosystem must serve the different scientific scenarios according to the following milestones:

- Jan 2010 (M4) – AquaMaps, FCPPS, and ICIS;
- Sep 2010 (M12) – AquaMaps, FCPPS, ICIS, INSPIRE, and DRIVER;
- Mar 2011 (M18) – AquaMaps, FCPPS, ICIS, INSPIRE, and DRIVER;
- Sep 2011 (M24) – AquaMaps, FCPPS, ICIS, INSPIRE, and DRIVER.

During the second year the focus was on setting up VOs and VREs for the INSPIRE and DRIVER Scenario. These scenarios required the integration of the D4Science, EGEE (EGI), and DRIVER **infrastructures** in the ecosystem. The developments started on the DRIVER and INSPIRE scenarios have been completed before the end of the first year of the project, and the two scenarios have been officially deployed in the production ecosystem on September 2010. In general the 5 scenarios have been further improved throughout the project lifetime and in addition a new scenario has been supported starting from June 2010. The gMan environment has been deployed in production to serve the needs of group of individuals and/or institutions in the context of the Arts and Humanities scenarios.

As originally defined in deliverable DSA1.1a and further described in DSA1.1b, the D4Science infrastructure gathers a number of **nodes** provided by SA1 work package members. In total four sites contributed to the infrastructure: CNR, FAO, FIN, and NKUA. At the end of the project the sites made available by the project partners provide a total of 43 physical machines, offering 903 GB RAM, 47,4 TB disk space, and 228 processor cores. These nodes were exploited to run the gCube system delivered by the D4Science SA3 work package and the gLite middleware as released by the EGEE (EMI) project. As a result the infrastructure made available 83 gCube nodes and had access to 3706 gLite nodes.

The management of the infrastructure was facilitated by the definition and implementation of clear procedures for **monitoring** and **accounting**. A number of monitoring tools were deployed allowing different infrastructures roles to visualize the status of their resources and to be actively notified when problems occurred. An accounting tool was also put in production providing relevant statistics about the users' exploitation of the infrastructure.

A clear **support** procedure was also defined and put in production during the course of the project. This procedure ensured an efficient response to all incidents affecting the ecosystem operation. A total number of 443 tickets were submitted, 44.4% as which had a high priority. All tickets were properly closed and documented.

From the perspective of the ecosystem user different **environments** have been deployed and made available satisfying the requirements expressed by the different project scientific scenarios. In total 5 VOs and 12 VREs were made available.

1 NODES OPERATION

This section describes the hardware resources that compose the D4Science infrastructure, the core infrastructure of the D4Science ecosystem, and the activities carried out during the project lifetime to operate such nodes. These activities include the deployment and upgrade of gCube, gLite and Hadoop in the infrastructure, the certification of the infrastructure nodes, and the management of infrastructure downtimes.

The D4Science infrastructure is organized in three node types:

- **gCube nodes** are the hardware resources able to run gCube services. The gCube software includes a particular web service container, the gCube Hosting Node (gHN), and a set of services and libraries that provide the functionality to create, manage, and exploit VREs;
- **gLite nodes** are computing and storage nodes running gLite software. gLite is the middleware provided by the EGEE project which continues to be developed in the context of the EMI project. By running gLite, these nodes provide core grid functionalities such as file-based storage, distributed computation of applications, etc. gLite nodes are exploited by gCube services that then provide higher-level functionalities through the D4Science VREs.
- **Hadoop nodes** are servers running Apache Hadoop Software which provides a distributed filesystem (HDFS) that can store data across thousands of servers, and a means of running work (Map/Reduce jobs) across those machines, running the work near the data. gCube Services can execute Hadoop Map Reduce Jobs using the gCube Execution engine which implements a particular adaptor to interface to Hadoop.

1.1 Hardware Resources

There are currently four sites providing hardware resources to the production infrastructure:

- **CNR** – Pisa, Italy
- **FAO** – Rome, Italy
- **FIN** – Kiel, Germany
- **NKUA** – Athens, Greece

The 43 physical machines provided are inline with the plans defined at the beginning of the project and allowed the deployment and availability of all VOs and VREs requested so far by the project SA2 and NA5 work packages.

The following table provides detailed information about the contribution from each site.

	#	CPU	RAM GB	Disk TB	Cores
CNR	10	Sun v20z 2x AMD Opteron	80	1.5	20
	2	Supermicro 6016T-NTRF4+: 2x Intel(R) Xeon(R) CPU E5630	144	4	32
	2	Dell PowerEdge R715: 2x AMD Opteron	256	8	48
	2	Supermicro Intel Core 2 Quad-core	8	8	8
	2	Sun X4140 2x AMD Opteron Six-core	128	1.2	24
	4	Sun X4100 2x AMD Opteron Quad-core	128	1.2	32
	4	Intel Core 2 Quad-core	28	15	16
	2	Sun X2100 AMD Opteron Dual-Core	16	1	4
	2	Asus: AMD Opteron 242	12	0.64	4
	2	Asus: AMD Athlon 64 X2 Dual-core	16	1	4
	1	Sun Cobalt LX50 Intel Dual Pentium III	3	0.15	2
FAO	1	2 Intel Xeon Quad-core	8	0.3	8
FIN	1	Intel Xeon E5504	12	1	4
NKUA	5	Dual Core Intel	40	2.9	10
	3	Intel Xeon	24	1.5	12
Total	**43**		**903**	**47.39**	**228**

Table 1 - Hardware resources by partner

1.2 Software Deployment

As introduced before, the resources allocated to the D4Science production infrastructure were exploited to host gCube, gLite and Hadoop nodes. Many of the hardware resources were virtualized in order to provide a higher number of gCube, gLite and Hadoop nodes. Due to the different nature of the three types of nodes the deployment and upgrade of software in these nodes is also distinct.

gCube Nodes

Of the four sites providing resources to the infrastructure, four sites deployed gCube nodes in their site. The common component to all gCube nodes is the gCube Hosting Node (gHN). The gHN is responsible to manage the deployed Running Instances (RIs) of the different gCube services.

In order to satisfy the requests from the project user communities for VOs and/or VREs, new nodes were continuously added to the infrastructure. Table 2 shows this increase in the number of nodes (gHNs) and a constant number of RIs, reflecting the utilization from VOs and VREs of the available gHNs. The decrease of RI from M13 to M15 is due the new search subsystem that has been moved from a distributed architecture to a centralized service.

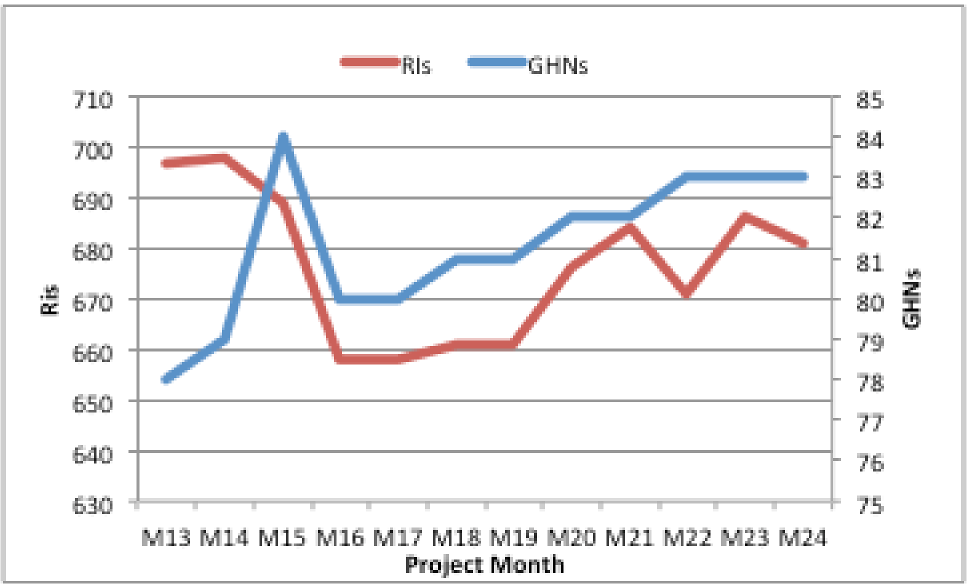

Figure 2 - Evolution of gHNs and RIs

The gCube release deployed in production at the end of the project is the release 2.5.1. During the second year of the project 5 major gCube releases were deployed in production (2.1.0, 2.2.0, 2.3.0, 2.4.0, 2.5.0)[1]. Due to the need to quickly fix critical defects affecting the production environment six gCube maintenance releases were also deployed (2.0.2, 2.2.1, 2.2.2, 2.3.1, 2.4.1, 2.5.1). Table 2 provides more information about the gCube major releases.

Release	Date	Integration Time	Subsystems	SAs
2.1.0	3 Nov 2010	12	23	209
2.2.0	12 Nov 2010	4	23	212
2.3.0	3 Mar 2011	24	23	229
2.4.0	6 Jun 2011	31	21	237
2.5.0	29 Jul 2011	29	21	243

Table 2 - gCube releases deployed in the infrastructure (M13-M24)

The chart below displays the information reported in Table 2 to highlight how the time needed to integrate a release increased along the year, reflecting the need to deliver always higher quality releases to SA1 WP and it's proportinal to the size of the release. The release dates reflected the evolution of the software with regular provision of code from the JRA work packages.

[1] The releases 2.6.0 and 2.7.0 integrated by SA3 team have not been deployed in production at the end of M24.

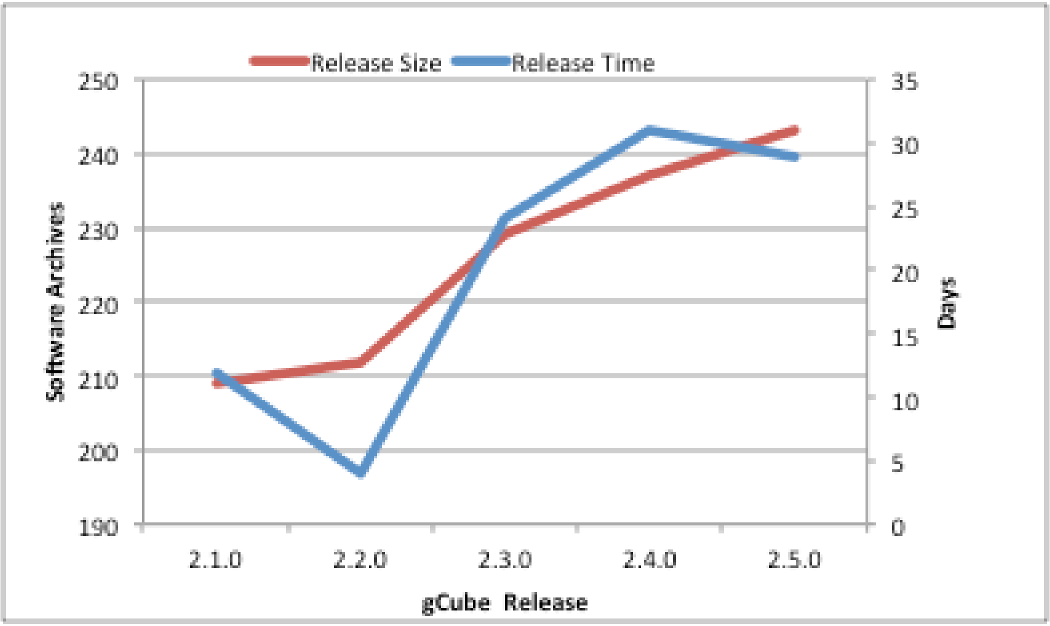

Figure 3 - gCube releases size and release integration time

This automatization of procedures is on the software deployment work. Figure 4 shows the number of days needed to deploy any new gCube release in the infrastructure. In average a normal release required 3 days (4 days during the first year of the project) and maintenance release 1 day. It should be noted that usually the upgrade to a new release requires modifications to all infrastructure nodes. A full upgrade of hundreds of nodes in such short time is only possible due to the remote node management features provided by the gCube services.

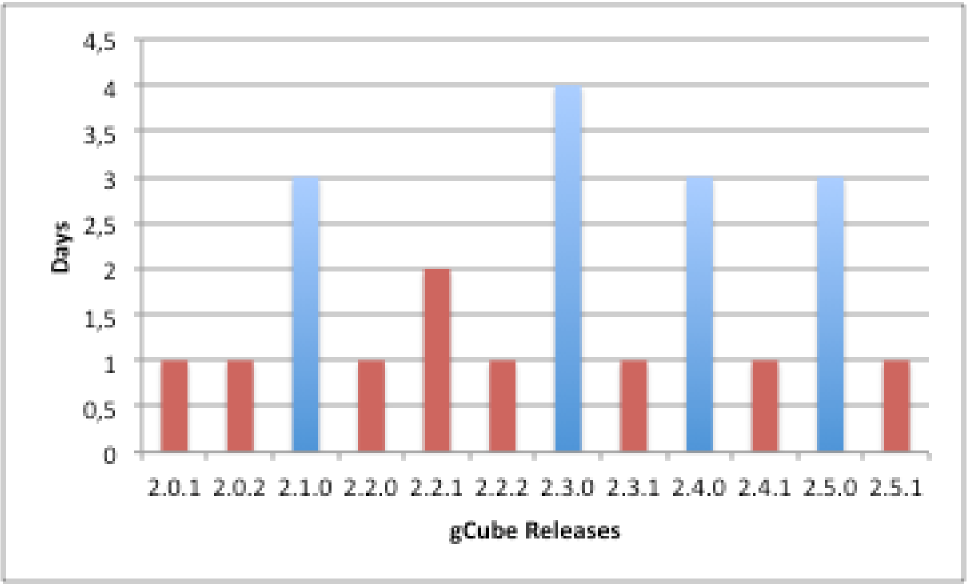

Figure 4 - gCube deployment time

gLite Nodes

Of the four sites providing resources to the infrastructure, two sites deployed gLite nodes in their site. Table 3 presents how the different gLite services have been distributed between these sites.

	CE	WN	sBDII	SE	WMS	LFC	VOMS	MyProxy
CNR	☑	☑	☑	☑	☑	☑	☑	☑
NKUA	☑	☑	☑	☑	☑	-	☑	-

Table 3 - gLite nodes per partner

The services marked in green represent a primary instance and in orange a secondary instance. The primary instances are the first ones contacted by the clients, which automatically switch to secondary instances in case of the primary's unavailability. All gLite sites provided by D4Science partners are also registered as sites of the EGI.eu project production infrastructure.

Besides the sites provided by project partners, other gLite sites of the EGI.eu production infrastructure also support the D4Science VO (d4science*.reseach-infrastructures.eu*) . These sites are: csTCDie, INFN Bari, INFN Trieste, and Taiwan.

In particular they provide access to 3706 gLite worker nodes distributed as follows:

- CNR – 10
- NKUA - 10
- INFN Bari – 1443
- INFN Trieste – 1983
- ASGC Taiwan – 240

In addition the sites provide access to storage resources equal to 4.3 TB distributed as follows:

- CNR – 51 GB
- NKUA – 280 GB
- csTCDie – 1923 GB
- INFN Trieste – 1983 GB
- Taiwan – 61 GB

Almost all gLite nodes provided by the above sites run gLite release 3.2. All patches for gLite 3.2 released during the last year were deployed in D4Science gLite nodes.

Hadoop Nodes

The Hadoop software is deployed in one of the four production sites. The following table summarizes the features of the cluster:

Partner	JobTracker	Slaves	HDFS Size	Ram	CPUs
CNR	☑	☑ ☑ ☑	1.2 TB	32 GB	32

Table 4 - Hadoop clusters features

The Hadoop Map/Reduce framework has a master/slave architecture. It has a single master server or jobtracker and several slave servers or tasktrackers, one per node in the cluster. The jobtracker is the point of interaction between users and the framework.

The Hadoop's Distributed File System is designed to reliably store very large files across machines in a large cluster. Hadoop DFS stores each file as a sequence of blocks, all blocks in a file except the last block are the same size. Blocks belonging to a file are replicated for fault tolerance.

The Hadoop software version running in production is the 0.20.0.

1.3 Nodes Certification

All nodes that compose the D4Science infrastructure must be certified before they can be exploited by the ecosystem.

gCube Nodes

Concerning the gCube nodes, there are no certification problems to report. All new nodes allocated to the infrastructure were kept up and running and registered in the gCube Information System with the correct activation status. With no certification problems to report all gHN containers were always available to host any service of the gCube system.

gLite Nodes

All nodes running gLite services were certified by the EGEE Regional Operation Centers. With the transition to the EGI.eu project the Regional Operation Centers (ROC) have been replaced by the National Grid Initiatives (NGIs) which perform certification and monitoring tests of the sites belonging to a specific country. The two D4Science sites running gLite services (CNR and NKUA) kept their services up and running satisfying the service availability levels required by EGI.eu, therefore no new certification problems were raised by the 2 related NGIs (NGI_IT and NGI_GRNET).

Hadoop Nodes

No certification procedure is applied to Hadoop nodes.

1.4 Nodes Downtime

The following chart presents the total number of downtimes declared by the sites that compose the D4Science infrastructure.

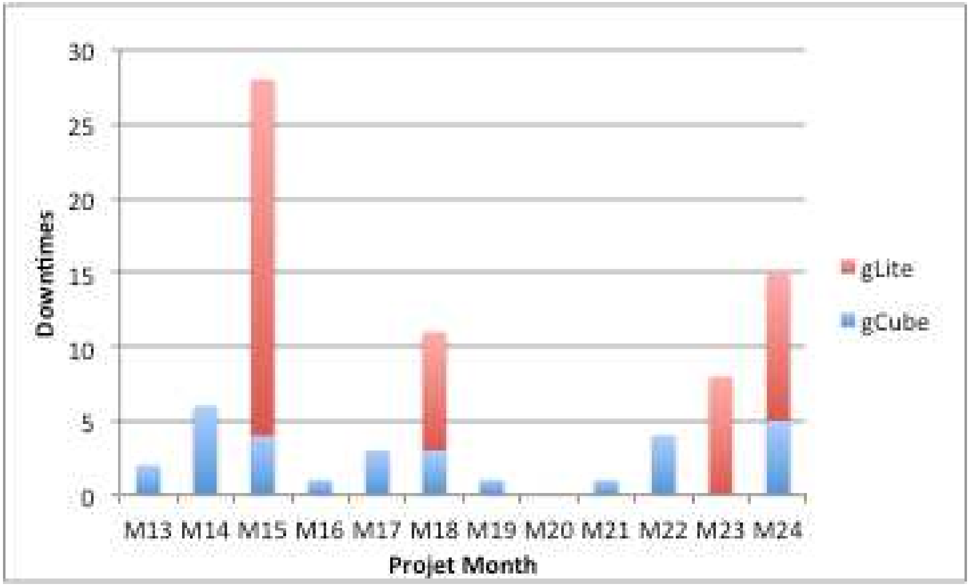

Figure 5 - Infrastructure nodes downtime

It should be mentioned that these downtimes were caused either by scheduled network interventions or by infrastructure upgrades. The downtimes never lasted for more than a few days so the production infrastructure was never affected by long unavailability periods. In downtimes motivated by infrastructure upgrades, only one VO was down at any time. No Hadoop downtimes were necessary during the project lifetime, while the high number of gLite downtimes can be explained by the new procedure introduced by the EGI.eu project in terms of downtimes, which have started to be calculated by Services and not by Sites.

2 MONITORING AND ACCOUNTING

In the D4Science II project a number of monitoring and accounting tools have been exploited:

- for gLite nodes the tools provided by the EGEE (EGI.eu) project: GOCBD, SAM, MyEGI.
- for gCube nodes two categories of tools : based on information collected by the gCube Information System (IS), and based on information produced/consumed by the gCube Messaging System (MS). While the gCube Messaging System provides tools for monitoring and accounting, the gCube Information system only provides tools for monitoring.
- for Hadoop nodes the tools are based both on Hadoop solutions and the gCube Messaging System.

This section presents the tools developed by the project (the tools developed by gLite and Hadoop are described in [8] and [9]) and provides some statistics regarding the infrastructure load and infrastructure usage during the project second year.

2.1 Base Technology

As introduced before the monitoring and accounting tools for gCube nodes are based on two main gCube subsystems:

- gCube Information System
- gCube Messaging System

gCube Information System

The gCube IS subsystem has been further enhanced during the second year of the project, but no new components were developed, therefore please refer to [7] for a brief description.

gCube Messaging System

The gCube Messaging subsystem has been further enhanced and new components have been developed or extended in order to cover most of the monitoring and accounting project's requirements.

The whole subsystem is composed by the following components:

- Central Broker:
  - Message broker – receives and dispatches messages;
- Producer Side:
  - Messages – defines the messages to exchange;
  - Local producer – provides facilities to send messages from each node;
  - Node monitoring probes – produces monitoring info for each node;
  - Node accounting probes – produces accounting info for each node;
  - Portal accounting probes – produces accounting info for the portal;
  - System accounting library – allows gCube services to account service specific information.
- Consumer Side:
  - Messaging consumer – subscribes for messages from the message broker, checks metrics, stores messages, and notifies administrators;
  - Portal accounting portlet - displays portal accounting information;
  - Node accounting portlet - displays node accounting information.

With respect to new components, the System Accounting Library has been developed in order to let gCube services account service related information. In particular the library has been integrated by the Workflow and Process execution Engine, in order to account job executions statistics (see 2.4).

2.2 Infrastructure Load

As presented in section 1.2 the number of gHNs running in the gCube nodes of the infrastructure was in the order of 70-80. These gHNs are exploited for the deployment of other gCube services to serve particular VOs and VREs. Figure 6 shows the load on the infrastructure from these VOs and VREs.

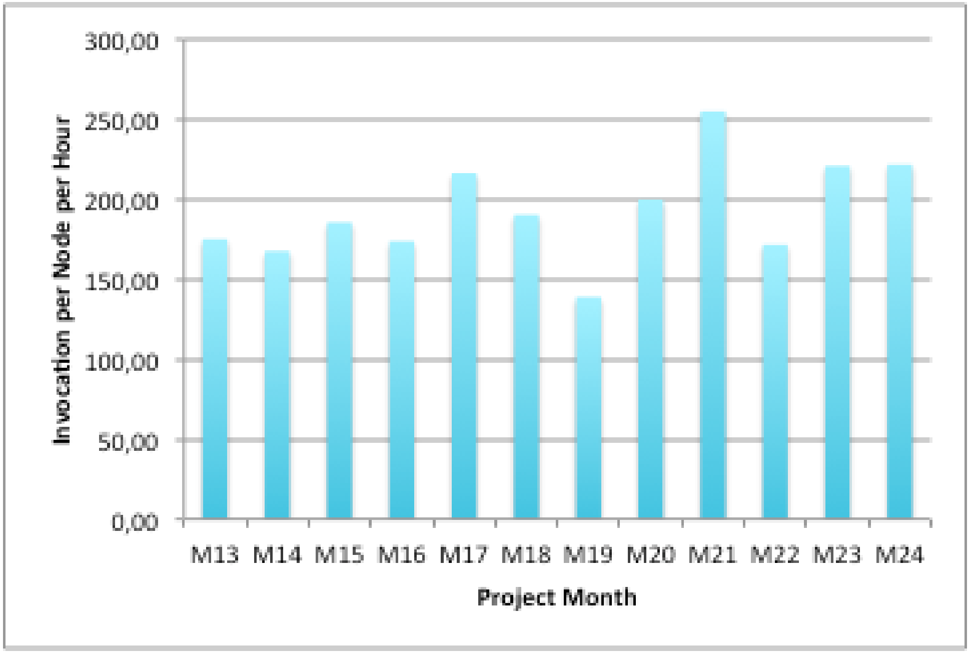

Figure 6 - Infrastructure load: average number of RIs calls

Due to the characteristics of the deployed monitoring tools it is possible to extract detailed information about the infrastructure load. If Figure 6 provides an overview of the average node utilization per hour, Table 5 depicts the total number of calls organized by gCube subsystem.

Subsystem	Calls
Information System	131229611
Content Management	251839
Index	237007
Search	102361
Application	85125
Execution	44507
VRE Management	32567
Metadata Management	24355
Personalisation	11685

Messaging	6683
DIR	1184
Open Search	927
Data Transformation	594
Annotation	388

Table 5 - Infrastructure load: total number of RIs calls

From the table above the large impact of the gCube Information System components in the overall infrastructure load is clear.

2.3 Infrastructure Usage

From the infrastructure exploitation viewpoint, by relying on the deployed accounting tools, it is possible to verify a continuous usage of the infrastructure by its users.

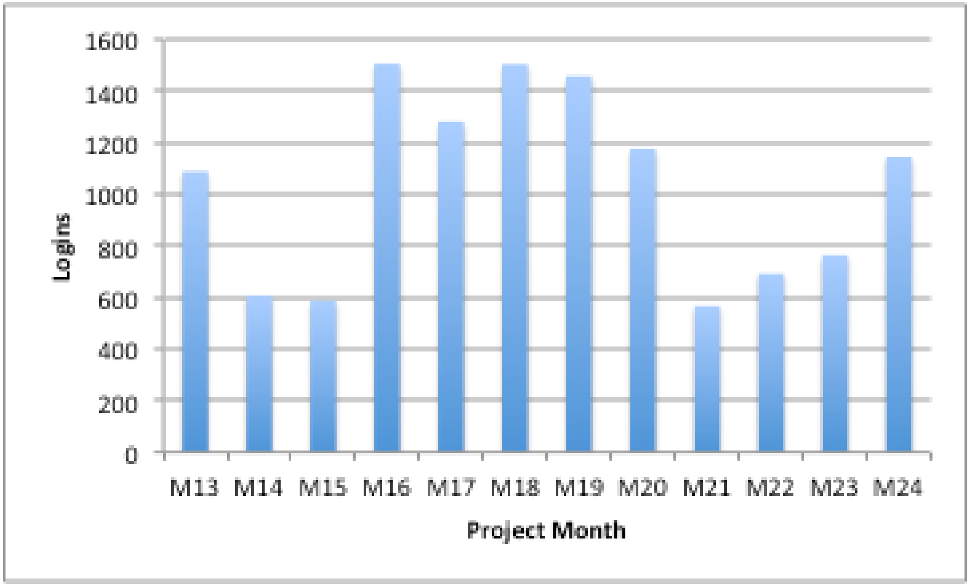

Figure 7 - Infrastructure usage: total number of logins

Figure 7 shows such usage by presenting the number of logins from the D4Science portal by month. As for the infrastructure load, more detailed statistics can also be produced for the infrastructure usage. Table 6 for example explains how much the portal users have exploited the different VOs and VREs offered by the infrastructure. This numbers however only takes into consideration the actions of the users on a subset of the functionality offered by the portal: Search, Time Series and Workspace.

Operation Type	Count
Advanced Search	4199
Browse	2538
Simple Search	2276
Workspace	1795
Content Retrieval	563
Time Series	217

Table 6 - Portal Operations

2.4 Infrastructure jobs statistics

The D4Science Ecosystem has been enhanced during the second year of the project in order to execute workflows trough the Workflow and Execution engine.

The Execution subsystem is capable to combine processes into workflows over the three types of nodes deployed on the infrastructure (or interfaced by other infrastructures) : gCube, gLite and Hadoop.

The exeucutions accounted since February 2011 using the mechanism offered by the gCube Messaging System (the System Accouting Library) are summarized in the following table:

Job Type	Total	Average Execution time (s)
gCube	1456	892
gLite	829	3072
Hadoop[2]	82	70

Table 7 : Infrastructure Job Statistics

It is evident that the average execution times of gLite executions suffered a lot from the job queuing and the resource sharing with other VOs, while the gCube and Hadoop nodes were dedicated only to D4Science users.

[2] The number of jobs for Hadoop does not take into account the Hadoop jobs directly executed on the cluster without contacting the Workflow and Process Execution engine. In that case the jobs could not be accounted for.

3 PRODUCTION SUPPORT

This section describes the activities carried out to provide support to the operation and exploitation of the D4Science ecosystem by its different user types: VO Administrators, VRE Managers, Site Managers, etc. The ecosystem support activity is based on the incident management procedure. This procedure follows the ITIL methodology for incident management and has been adopted since the beginning of the project and it has been enhanced during the project lifetime.

The incident management procedure description is available on the previous period deliverable [7] while detailed information about the procedures, the people involved, and the workflow of the different activities, is available in the D4Science infrastructure website [6] under the incident management section.

Even if there is a common incident management procedure, some small differences apply whether the procedure is applied to gCube or gLite nodes.

- gCube and Hadoop Nodes: All incidents related to the exploitation and deployment of gCube and Hadoop nodes are tracked according to the incident management procedure;
- gLite Nodes: The incidents related to the usage of gLite nodes are followed according to the incident management procedure. The only exception concerns the incidents related to the deployment of gLite nodes, which are submitted directly to the regional NGIs.

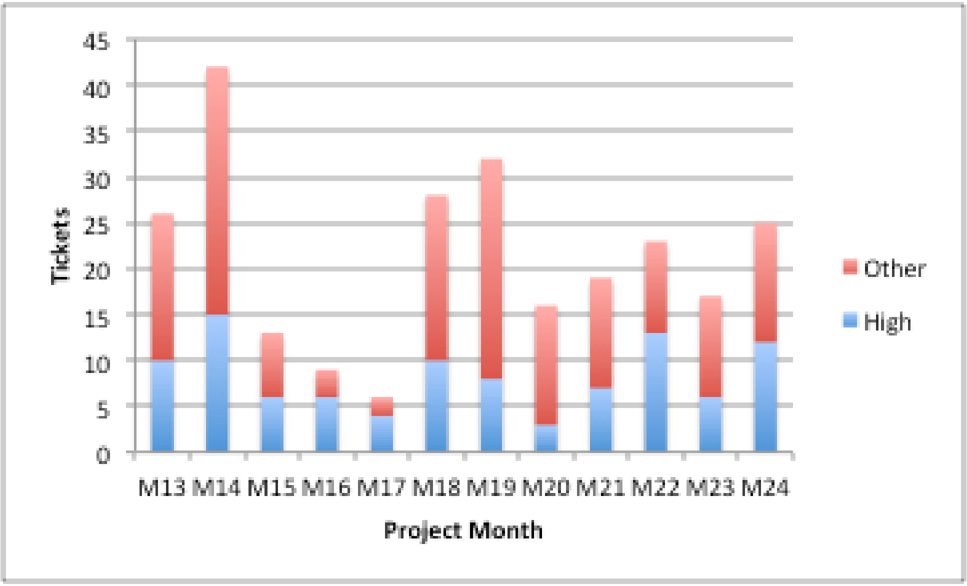

Figure 8 - Incident tickets by priority

A total number of 256 tickets were submitted during the project second year. From this total, 39,6% (100 tickets) was high priority incidents. Figure 8 and Figure 9 provide detailed information about the total number of submitted tickets, the number of high priority tickets, and the average number of days the tickets remained opened.

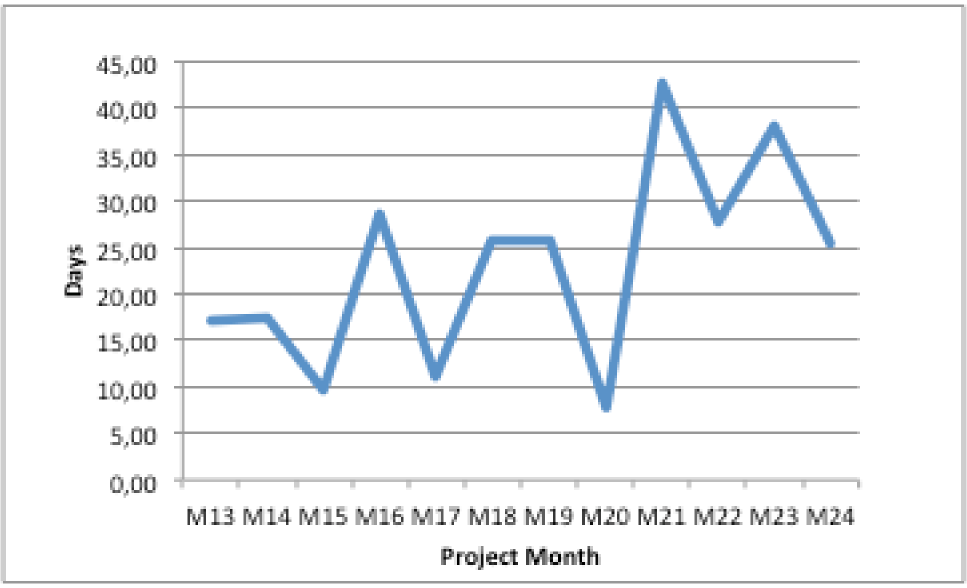

Figure 9 - Incident tickets average resolution time

The number of days required to close incident tickets till M20 has always been in the time interval of 10-25 days, while the peaks starting from M21 can be justified by the high number of old incident tickets which have started to be fixed with the end of the project approaching.

Concerning the affected environments, Figure 10 shows the distribution of the recorded incident tickets across the support VREs. Many tickets were common to all existing VREs. Looking to VRE-specific tickets the most affected VREs are the ones under the FARM VO (FCPPS, ICIS, and AquaMaps). Such high numbers of incidents are expected as these VREs are the most exploited of the infrastructure.

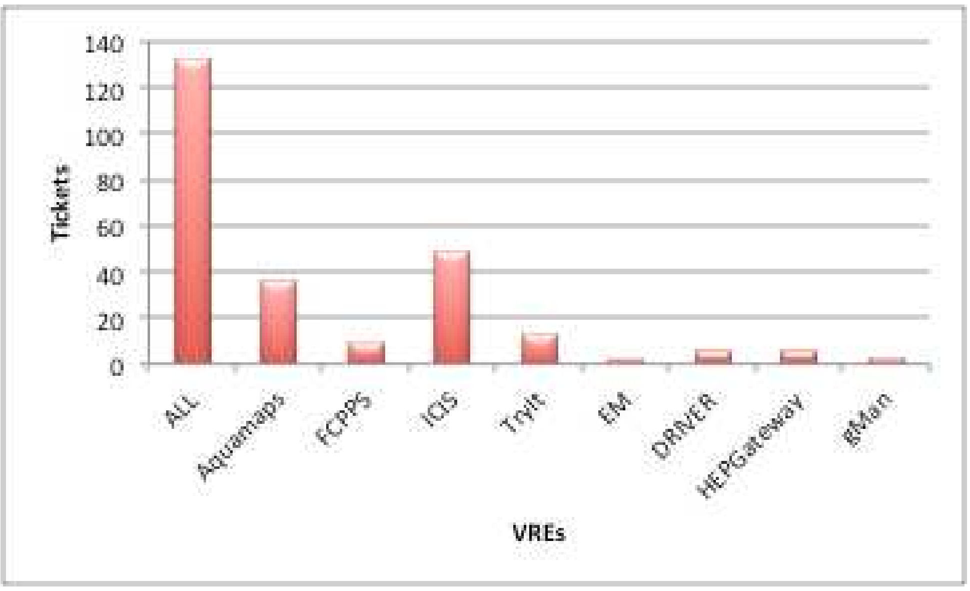

Figure 10 - Incident tickets by VRE

4 ECOSYSTEM ENVIRONMENTS

Virtual Organisations (VOs) and Virtual Research Environments (VREs) are, from the infrastructure operation point of view, sets of resources and users grouped together by sharing policies with the goal to serve the needs of a certain scenario.

The set of VOs and VREs the D4Science-II ecosystem operates, as well as their evolution in terms of resources involved and services offered, is mainly a consequence of the requirements captured in the context of the five scientific application scenarios the project is focused on – i.e. INSPIRE, DRIVER, AquaMaps, FCPPS and ICIS. These requirements are documented in "Communities Practices and Requirements" [1]. Specified requirements are carefully analysed and transformed into a deployment plan as documented in "VOs and VREs planning" [2]. This development plan is then implemented by a dedicated team and a living report is produced to document the activities performed [3]. In addition to the five scenarios introduced above, other VOs and VREs have been deployed to serve the needs of "external" communities having expressed their interest in these services.

This section provides a brief report on the VOs and VREs deployed and maintained from October 2010 to October 2011[3]. During this reporting period five Virtual Organisations have been deployed and operated in the context of the D4Science-II infrastructure[4]:

- The **FARM** Virtual Organisation has been created for the Fisheries and Aquaculture Resources Management communities. This VO supports a large number of application scenarios from these communities such as the production of Fisheries and Aquaculture Country Profiles, the management of catch statistics including the harmonisation across data-sources, the dynamic generation of biodiversity maps and species probability maps, the analysis of vessel trajectories;
- The **Ecosystem** Virtual Organisation has been created to serve the needs of any community interested in exploiting the D4Science-II Ecosystem capabilities. It is not conceived to serve the needs of a specific community focusing on a research topic. It aims at promoting a cross-fertilisation among the various communities partaking to it. The pool of resources is more heterogeneous than those of the other Virtual Organisations supported since it is not driven by a specific research topic;
- The **INSPIRE** Virtual Organisation has been created to serve the needs of the High Energy Physics community through the INSPIRE scenario. This VO supports a number of scenarios ranging from the execution of large scale data processing facilities oriented to build resources supporting the INSPIRE service backend to cross collection search use case;
- The **Arts-Humanities** Virtual Organisation has been created to serve the needs of groups of individuals and/or institutions in the context of the Arts and Humanities scenarios mainly focusing on the study of ancient documents.
- The **gCube Apps** Virtual Organisation has been created for demonstration purposes. This Virtual Organisation is not conceived to serve a specific community, rather it has been created to host a number of VREs focusing on specific applications ranging from ecological niche modelling to time series management and vessel trajectories analysis.

VO	VREs	Users
FARM	4	43
Ecosystem	3	43

[3] The snapshot of the previous period is documented by DSA1.2a [7].

[4] The gCubeApps VO has been deployed after the official end of the project, therefore no related statistics have been reported on chapters 1,2 and 3.

INSPIRE	1	19
Arts-Humanities	1	22
gCube Apps	4	14

Table 8 - D4Science-II VOs detailed information

Twelve Virtual Research Environments have been deployed and operated, each in the context of a specific Virtual Organisation[5]:

- The **AquaMaps** Virtual Research Environment (in the FARM VO) is for providing fisheries and aquaculture scientists with facilities for producing and accessing species predictive distribution maps showing the likelihood that a certain species or a combination of species will live in specific regions or areas;
- The **Fisheries Country Profiles Production System (FCPPS)** Virtual Research Environment (in the FARM VO) is for fisheries and aquaculture authors, managers and researchers who produce reports containing country-level data. It provides seamless access to multiple data sources, including their annotation and versioning and permits production of structured text, tables, charts and graphs from these sources to be easily inserted into custom reporting templates that can support multiple output formats;
- The **Integrated Capture Information System (ICIS)** Virtual Research Environment (in the FARM VO) offers fisheries statisticians a set of tools to manage their data. Statisticians produce statistics from often very different data sources, and need a controlled process for the ingestion, validation, transformation, comparison and exploitation of statistical data for the fisheries captures domain;
- The **Vessel Transmitted Information (VTI)** Virtual Research Environment (in the FARM VO) is for marine biologists willing to analyse vessel activities over space and time by taking into account environmental data;
- The **DRIVER** Virtual Research Environment (in the Ecosystem VO) is for users willing to access the content offered by the DRIVER infrastructure (an infrastructure built by aggregating scientific publications in journal articles, dissertations, books, lectures, report, etc. from 250+ repositories scattered over 30+ countries) in the context of a VRE as to benefit from additional facilities;
- The **Environmental Monitoring (EM)** Virtual Research Environment (in the Ecosystem VO) is for making available various satellite data and services for consuming such data. It resulted from the fusion two Virtual Research Environments operated in the previous period, i.e. GCM and GVM;
- The **TryIt** Virtual Research Environment (in the Ecosystem VO) is conceived to serve demonstration and training activities. It gives access to a set of sample content and supports the most common functionality a gCube-based VRE might be able to support, e.g. search, browse and personal work space management;
- The **HEPGateway** Virtual Research Environment (in the INSPIRE VO) is conceived to support users willing to access the INSPIRE High Energy Physics Information System, i.e. the one-stop-shop for the HEP community research publications. Through this VRE users might benefit from the gCube facilities for managing information objects retrieved through such a service, e.g. store them in their workspace, share with coworkers, annotate.
- The **gMan** Virtual Research Environment (in the Arts-Humanities VO) is conceived for the activities of the homonymous project. It serves as an investigation into the support that gCube can offer to research activities that originate in the Humanities, particularly in the study of ancient documents;

[5] The DocumentWorkflow, EcologicalModeling, TimeSeries and VesselActivitiesAnalyzer VREs have been deployed after the official end of the project; therefore no related statistics have been reported on chapters 1,2 and 3.

- The **DocumentsWorkflow** Virtual Research Environment (in the gCube Apps VO) is conceived to provide its users with a working environment focused on the gCube facilities for managing Document life-cycles. It exploit the facilities offered by the gCube Business Documents Workflow Management Suite enabling the production of reports that require a collaborative activity of several actors;
- The **EcologicalModeling** Virtual Research Environment (in the gCube Apps VO) is conceived to provide its users with a working environment focused on the gCube facilities for producing species distribution maps resulting from the processing of data on species characteristics and environmental observations. The resulting maps are actually rich information objects containing PNG images, GIS layers as well as metadata;
- The **TimeSeries** Virtual Research Environment (in the gCube Apps VO) is conceived to provide its users with a working environment focused on gCube facilities for managing time series. This environment supports the load of time series objects, the curation and validation by relying on authoritative code lists, the sharing of such objects with coworkers, the production of graphs, the visualization through a GIS service;
- The **VesselActivitiesAnalyzer** Virtual Research Environment (in the gCube Apps VO) is conceived to provide its users with a working environment focused on gCube facilities for managing vessel trajectories. This environment support users in loading and curating their own vessel trajectories, enriching such data with bathymetry and FAO Area, sharing with co-workers, analysing such objects by producing maps on vessel activities and fishing monthly effort;

VRE	Users
AquaMaps	36
FCPPS	27
ICIS	29
DRIVER	16
EM	13
TryIt	33
HEPGateway	16
gMan	18
DocumentsWorkflow	6
EcologicalModeling	11
TimeSeries	11
VesselActivitiesAnalyzer	11

Table 9 - D4Science-II VREs detailed information

REFERENCES

[1] Ellenbroek, A. "Communities Practices and Requirements". DNA5.1 D4Science-II Project Deliverable, March 2010

[2] Candela, L.; Pagano, P. "VOs and VREs Planning". DSA2.1 D4Science-II Project Deliverable, February 2010

[3] Koltsida, P.; Kakaletris, G.; Candela, L. "VOs and VREs Operation Activity Report". DSA2.2 D4Science-II Project Deliverable, September 2010

[4] Andrade, P.; Candela, L.; Manzi, A.; Pagano P. "Infrastructure Operation Report". DSA1.3c D4Science Project Deliverable, January 2010

[5] D4Science-II website: http://www.d4science.eu

[6] D4Scienc-II production infrastructure wiki: https://service.wiki.d4science-ii.research-infrastructures.eu/service/index.php/SA1_Home

[7] Andrade, P.; Candela, L.; Manzi, A.; Pagano P. "Knowledge Ecosystem Operation Report". DSA1.2a D4Science-II Project Deliverable, September 2010

[8] MyEGEE project Website : https://twiki.cern.ch/twiki/bin/view/EGEE/MyEGEE

[9] Hadoop Cloudera Distribution : http://www.cloudera.com/hadoop/