COMBINATION OF COLLABORATIVE PROJECT AND COORDINATION AND SUPPORT ACTION

Integrated Infrastructures Initiative project (I3) proposal
Infrastructures Call 4

FP7-INFRASTRUCTURES-2008-2

Data Infrastructures Ecosystem for Science

D4Science-II

Date of preparation: September 2008
Version number *(optional)*:

Participant no.	Participant organisation name	Part. short name	Country
1 (Coordinator)	GEIE ERCIM	ERCIM	FR
2	Consiglio Nazionale delle Ricerche	CNR	IT
3	National and Kapodistrian University of Athens	NKUA	GR
4	European Organisation for Nuclear Research	CERN	CH
5	Engineering Ingegneria Informatica SpA	ENG	IT
6	University of Strathclyde	BDM-USTRATH	UK
7	The Food and Agriculture Organisation of the United Nations	FAO	IT
8	Fishbase Information & Research Group Inc.	FIN	PH
9	4D SOFT Számítástechnikai Kft	4D SOFT	HU
10	Terradue s.r.l.	TERRADUE	IT

Work programme topics addressed
INFRA-2008-1.2.2: Scientific Data Infrastructure

Name of the coordinating person: Jessica Michel
e-mail: jessica.michel@ercim.org
fax: +33 4 92385011

Name of the scientific coordinator: Donatella Castelli
e-mail: donatella.castelli@isti.cnr.it
fax: +39 050 3153464

Proposal abstract

D4Science-II will develop technology to enable interoperation of diverse data e-Infrastructures that are running autonomously, thereby creating e-Infrastructure Ecosystems that can serve a significantly expanded set of communities dealing with multidisciplinary challenges whose solution is currently beyond reach. Furthermore, D4Science-II will bring together several scientific e-Infrastructures established in the areas of biodiversity, fishery resource management, high energy physics, etc., to set up a prototypical instance of such an ecosystem. This will support several critical scientific scenarios that are distinct but also feed into and enrich each other in nontrivial ways. Finally, possibly in collaboration with appropriate international bodies and initiatives, D4Science will take steps to ensure sustainability of the Ecosystem, some of them based on and synchronized with the dissemination, training and standardization activities of the project.

Technologically, D4Science-II constitutes a continuation of the DILIGENT and D4Science projects, which have initiated an effort towards using existing network, grid, and repository infrastructures to deploy on top of them a pan-European research e-Infrastructure that will enable unlimited generation and dissemination of scientific knowledge. As the next step towards this goal, D4Science-II will transform the current, operational D4Science e-Infrastructure into the lynchpin of an e-Infrastructure Ecosystem, holding together and mediating between all participating e-Infrastructures through programmatically-available interoperability services.

The D4Science-II ecosystem will include among others, the GENESI-DR and DRIVER repository e-Infrastructures, and important thematic repositories maintained by international organizations, e.g., INSPIRE and AquaMaps. The project will create Virtual Research Environments that will offer significantly enhanced services to scientist without incurring high development and maintenance costs.

Table of contents

Section 1: Scientific and/or technological quality, relevant to the topics addressed by the call

1.1 Concept and objectives

> ***"Urban social pressure in Brazil drives deforestation. Carbon from fires adds thermogenic gases to the atmosphere. The capacity of the canopy to retain moisture is diminished, and the rainforest shifts into the process of desertification. Surface soils & nutrients are eroded and make their way via rivers to the oceans. Oligotrophic waters are temporally enriched. Resulting blooms of the microalyal coccolithophores sequester carbonates as they form scales. The nutrients are exhausted, the algae die, the scales settle to the ocean floor. This is the start of the geological burial of calcium carbonates, perhaps the main counter-force to anthropogenic additions of carbon dioxide to the atmosphere."***
>
> ***"Modelling the impact of human activities on net carbon dioxide balance is a complex process that could only be performed by having access to a multitude of information sources and tools."***
>
> ***"Knowledge Ecology is knowledge generated by analysing information from diverse sources for a more effective analysis"***
>
> Thomas Garnett, Second GRL 2020 Workshop, Pisa (Italy), March 2008

Where will one find all the necessary resources (original information, analysis software, and processing power) to experiment with the appropriate models, perform the relevant computational experiments, and generate the knowledge required to solve such a critical global problem? Invariably, the relevant knowledge comes from many diverse scientific disciplines, the resources that are potentially available for attacking the problem are scattered across different research communities, and (as of late) these resources are managed within heterogeneous, independently-operated Electronic Infrastructures (e-Infrastructures). The latter are often dealing with multi-disciplinary, multi-purpose, multi-form resources themselves, but at a much smaller scale than what is required. Moreover, any thoughts of a scaled-up all-inclusive e-Infrastructure can be quickly dispersed, for rather obvious sociological, financial, administrative, and technological reasons: the concept of *THE ONE* global e-Infrastructure is as plausible as a unique natural language or a unique car model for everyone. Interoperation of existing, autonomous, heterogeneous e-Infrastructures appears to be the only solution, leading to the creation of ***e-Infrastructure ecosystems***, where the individual resources are harnessed and aggregated for a multiplicative effect on information availability, processing power, and more importantly scientific perspective: the much desired ***Knowledge Ecology*** finally obtains life within and is fed by e-Infrastructure ecosystems, much like Natural Ecology (whose problems it is often trying to address) obtains life from our own Natural Infrastructure Ecosystem, the Earth.

D4Science-II intends to deliver exactly what is mentioned above to help addressing scientific and societal challenges like the one described by Thomas Garnett. It will develop technology and methodologies that will enable **sustainable interoperation** of multiple, diverse and heterogeneous data e-Infrastructures that have been established and are currently running autonomously, thereby creating e-Infrastructure ecosystems that can serve a significantly expanded set of communities dealing with complex, **multidisciplinary challenges** whose solution is beyond reach with existing resources. Furthermore, D4Science-II will use the existing D4Science e-Infrastructure as a hub to bring and hold together several established scientific e-Infrastructures and, thus, set up a prototypical instance of such an e-Infrastructure ecosystem. This will be offered to researchers from several different scientific areas, e.g., **biodiversity, fishery resource management, and high energy physics**, to support originally distinct scientific scenarios that nevertheless feed into each other and enrich each other with services that none of them had at its disposal individually. The scenarios will be served by innovative **Virtual Research Environments** (VREs) that will be set up to provide the runtime framework required to support researchers *in extracting further meaning from masses of data stored in institutional, national or community repositories*. Finally, D4Science will study issues of sustainability of e-Infrastructures and e-Infrastructure ecosystems.

Through the mediating role of the D4Science e-Infrastructure, the ecosystem will "*increase the scale of federation and interoperation of digital repositories, consolidating synergies with the underlying e-Infrastructures"* and will *"reduce costs, increase the users' base and bridge across multidisciplinary communities, enabling cross-fertilisation of scientific results and favouring innovation.*" The overall result will be "*a more efficient way for all scientists to work on global research challenges that would otherwise be difficult to address, rationalising at the same time the investments in expensive resources and fighting digital divide*".

Europe will emerge as a **two-fold beneficiary** of this effort. First, **European scientists** in the aforementioned communities will be the initial users of the first ever e-Infrastructure ecosystem, taking advantage of unprecedented computational power and information richness and addressing their research problems with several new methodologies thus afforded. Second, scientists around the globe will be drawn to the ecosystem for their own work while other ecosystems (and possibly recursively, higher-level ecosystems of interoperating ecosystems) will be formed based on the techniques developed in the project, if not the D4Science e-Infrastructure and software itself. This will increase and reinforce the **European leadership** in the relevant technologies as well as the impact and global relevance of the D4Sience e-Infrastructure and the ecosystem based on it as a unique platform for supporting science.

Background and Detailed Motivation

Modern patterns of scientific enquiry require collaborations among parties that are widely dispersed and otherwise autonomous. Collaborations are often cross-discipline and may rely on virtual research environments that make available data, processing and interaction intensive workflows to translate instrument observation into information, information into knowledge, and knowledge into the need for further observation.

Across the whole cycle, the functional and non-functional requirements that characterize collaborative workflows - particularly *cost-effectiveness*, *security*, and *autonomic adaptation* to highly dynamic environments - exceed the capacity of commodity technologies and ad-hoc developments alike. As a result, a remarkable amount of worldwide research and development over the past decade has concentrated on the deployment of *e-Infrastructures* of human, hardware, software, and data resources for the support of innovative research environments.

From the beginning, infrastructural efforts have rallied around the design principle of *resource sharing*, most explicitly captured in the vision of Grid computing. Over time, its interpretation has become increasingly holistic: *like computing cycles, storage, and data before, infrastructures are now expected to enable transparent sharing of domain and application services*. The expectation suggests a layered approach to the design and deployment of infrastructures: higher-level infrastructures build upon lower-level ones and broaden the scope of their support for resource sharing. To reach the application domain, in particular, *last-generation infrastructures virtualize the facilities of underlying middleware systems towards increased transparency, interactivity, content/knowledge-orientation, and development support*. D4Science is one of the distinguished examples of such principled and integrated approach to the deployment of an application-level infrastructure.

A fundamental design concern in infrastructural efforts - one that cuts across the growing stack of initiatives - is related to the degree with which heterogeneous and autonomously operated resources may be made to interoperate both within and *across* infrastructures. There is growing evidence, in particular, that the requirements raised by cross-discipline research collaborations may not be always satisfied within the boundaries of a single infrastructure, regardless of how wide in geographical scale and large in aggregate capacity this may prove to be. Rather, the expectation is that *scientific collaborations will need to span across multiple infrastructures as they involve scientists and resources from different institutions, disciplines, and countries*. As current infrastructural efforts either approach or already strengthen the production-level quality of their services, *it becomes increasingly pressing to consider the issue of cross-infrastructure interoperability and along the entire spectrum of resource sharing abstractions that they may uphold*. The risk is ultimately to constrain the vision of resource sharing within the artificial boundaries of regional developments and to force an equally artificial partition upon the work of scientists.

D4Science-II Vision and Objectives

The previous observations suggest the compelling need of introducing a more powerful model in which e-Infrastructures are not isolated but dynamically interoperate and influence each other as the components of an ecosystem. In order to create such an ecosystem, the project will focus on the deployment and exploitation of mechanisms and policies for the interoperability of application-level infrastructures and infrastructural services. In the vision that drives this proposal, *the possibility of any two infrastructures exchanging resources (using the D4Science e-Infrastructure as the mediator) will complement their individual strengths at a contained cost, thereby accelerating the adoption and broadening the coverage of each within a cycle that increases the mutual advantage at each iteration*. Furthermore, the collective benefit will be proportional to the number of interoperable infrastructure pairs, for *the gain of each infrastructure will then transitively propagate to the others as an opportunity for growth and innovative exploitation*. In this sense, **interoperability will foster the coordination of otherwise autonomous application-level infrastructures and infrastructural services into a true Knowledge Ecosystem**.

The D4Science e-Infrastructure is in a privileged position to act as the driving force behind the formation of the envisioned ecosystem. Given its layered approach to infrastructure building, it can conveniently mediate access to the computational facilities of the underlying EGEE infrastructure and, in turn, the connectivity fabric of the GÉANT infrastructure. Likewise, it can export unique functionality for the definition, operation, and autonomic management of VREs, which contextualize and greatly simplify the controlled use of shared resources. Such capabilities appeal to many application-level infrastructures, repositories, and other infrastructural services that *(i)* have been extremely successful at aggregating resources of interest to their communities, but *(ii)* for strategic as well as practical reasons, have excluded the provision of advanced computational and collaborative services from the scope of their current design, and in some cases, from their future roadmap.

Given all the above, the main objective of the proposed project is **to advance the current D4Science e-Infrastructure into a pivotal element that, through its capabilities and mediating role, will be capable of drawing numerous infrastructural initiatives and scientific communities within the scope of a Knowledge Ecosystem.** To succeed in its overall mission, the project will address the concrete objectives outlined below.

OBJECTIVE 1: Interoperability of the D4Science e-Infrastructure with Other Data e-Infrastructures

D4Science-II constitutes a continuation of the **DILIGENT** [21] and **D4Science** [17] projects, which have initiated an effort towards deployment of a "*pan-European scientific e-Infrastructure on top of existing network, grid, and repository infrastructures*". The ultimate goal of this chain of work is "*to provide scientists with virtual research environments*" for "*unlimited generation and dissemination of scientific and technical knowledge*". As the next step towards this goal, D4Science-II will transform the current operational D4Science e-Infrastructure into the lynchpin of an e-Infrastructure ecosystem, holding together and mediating between all data e-Infrastructures[1] that participate in the ecosystem. Initially, these will include among others the **GENESI-DR** [32] and **DRIVER** [22] repository e-Infrastructures, created in the framework of FP7 projects, and important thematic multi-type repositories maintained by large stakeholder international organizations, primarily **INSPIRE** [41] and **AquaMaps** [45]; in time the goal is to attract many other European and international e-Infrastructures as well. These data services manage several millions of information objects of different types, ranging from metadata and textual documents, to maps, sensor, statistical and GIS data. Close collaboration with standardization bodies, such as **W3C**, **OGF**, and **ETSI,** together with early adoption of new emerging standards will ensure that the interoperability solutions identified will have wide applicability.

Figure 1 shows the Knowledge Ecosystem envisioned, i.e., interoperable data e-Infrastructures, repositories, and scientific communities exploiting the services provided. It also illustrates the dual role played by D4Science e-Infrastructure, i.e., virtual aggregator of resources available in interoperable e-Infrastructures,

[1] For simplicity, in the following, we will use the term "data infrastructure" to mean a generic class of services giving access to data sources. These comprise elements termed sometimes repository infrastructure, (federated) repositories, (federated) digital libraries, archives, etc.

and provider of these resources back to the participating e-Infrastructures and, through those, to complex VREs serving cross-domain scientific communities.

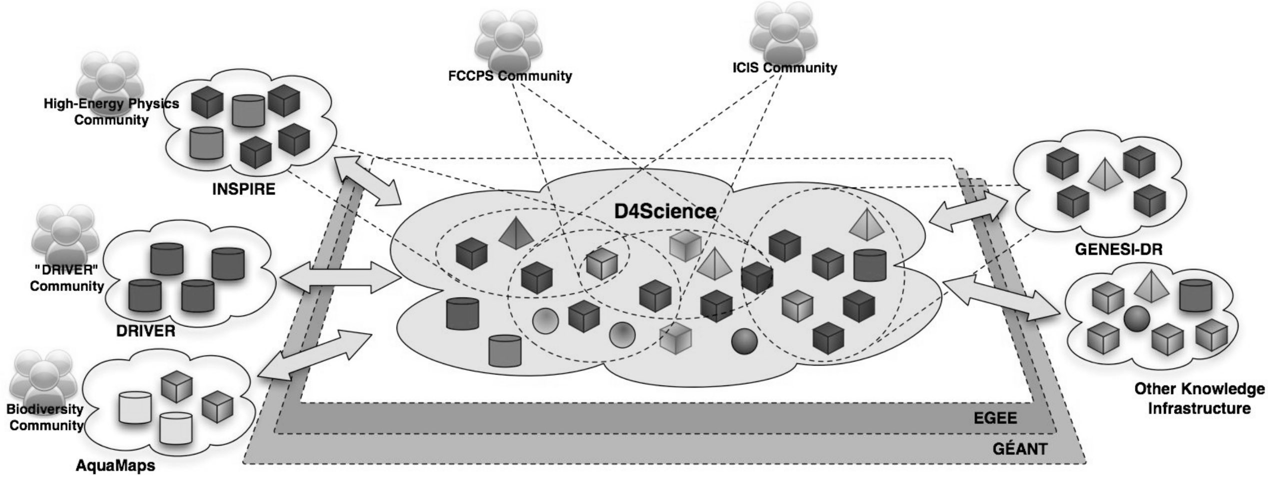


Figure 1. D4Science as a linchpin of a knowledge ecosystem

Enhanced with rich interoperability functionality, which will be available **programmatically** as well, the D4Science e-Infrastructure will come out of the project ready to play a brand new set of roles. It will allow data e-Infrastructures and other data-centric applications to **outsource the implementation of specific functionality** to it, a process that will become increasingly attractive economically as more resources and advanced services become available in the ecosystem. The D4Science e-Infrastructure will be the vehicle through which the other components of the ecosystem will be able to exploit **EGEE resources** for executing computationally intensive and potentially repetitive tasks in a completely transparent way without incurring high sustainability costs. It will also be able to act as a **broker of resources** in the Knowledge Ecosystem, enabling data e-Infrastructure to **transparently exploit the resources of each other**, again without implementing expensive ad-hoc solutions.

OBJECTIVE 2: Support for Scientific Scenarios Related to Global Research Challenges

By taking advantage of the huge amount of cross-domain resources that will become available through the interoperability mechanisms and policies established, D4Science-II will create several Virtual Research Environments that offer significantly enhanced services to their users without incurring high development and maintenance costs. These VREs will cover the needs of five very different scientific scenarios, for which interacting with a Knowledge Ecology and sharing across e-Infrastructures play a fundamental role in enhancing the quality and efficiency of the underlying research. Each of these scenarios responds to the concrete demands of a specific multi-disciplinary scientific community, which plans to make the final outcome of D4Science-II available in its **own production working environment** at the end of the project. The scenarios exemplify the main patterns of exploitation of the aggregate pool of resources that become available within an ecosystem based on the D4Science e-Infrastructure. Collectively, such patterns characterize a Knowledge Ecology in which individual parties act in the role of resource providers and/or resource consumers, and where the collective benefit becomes a strong implication of as well as a necessary condition for individual gain. The VREs that will be created and the corresponding scenarios that will be supported are briefly analysed below.

The **INSPIRE VRE** will respond to the needs raised by developers of the **INSPIRE service**. This new repository for High-Energy Physics (HEP), a research infrastructure in itself, intends to offer to its worldwide user basis, counting over 30,000+ scientists, more complex functionality than earlier HEP repositories, which had technical, architectural, and computational limitations. Part of this new functionality is based on **computationally intensive processing of bibliometric information** for advanced information search and retrieval through the exploitation of multi-dimensional search spaces and for assessing the impact of scientific results through hybrid metrics. Instead of adopting a "silo mentality" and solve these problems in-house by adding their own processing power and hard-wired algorithms, INSPIRE designers have opted for exploiting the D4Science facilities. D4Science-II will create a VRE that will be accessible programmatically and will provide the virtual context, i.e. appropriate content, tools and computational resources, for supporting the INSPIRE service. As a result, INSPIRE will not face any large computational bottleneck (maintenance, processing power, and expertise) when providing such innovative functionality. Moreover, the tools developed in this VRE will become **available to other communities facing similar problems** in other environments.

The **DRIVER VRE** will enhance the off-line capabilities underlying the DRIVER portal, which currently harvests, aggregates, and curates approximately 700,000 documents from 150 repositories in 20 countries, as well as the on-line services offered to its users, which include search, user profiling and recommendation. In its present form, the DRIVER e-Infrastructure, from which the portal has been built, lacks the necessary hardware and software resources to handle any computationally-intensive or storage-heavy **high-value and innovative end-user services**, e.g., cutting-edge **search**, advanced user profiling and bibliometric document analysis for **personalization**, etc. Furthermore, the DRIVER portal gives access to a respectable but still relatively limited number of documents, making the creation of **enhanced publications** that recursively point to other data and documents not sufficiently appealing. D4Science-II will create a programmatically accessible VRE that will offer **access to computationally-intensive services,** including some related to interoperability itself, i.e., metadata brokerage, data transformations, and ontology-based mapping. It will also support the concept of enhanced publications, which under appropriate policies, would consist of DRIVER content potentially linked to **any other data available in the ecosystem**. Finally, it would incorporate the bibliometric-data-mining and metrics-calculation **services implemented for the INSPIRE scenario** to provide new sophisticated information retrieval and personalisation functionality.

The **AQUAMAPS VRE** will support the **AquaMaps service**, which produces maps that result from fitting geo-referenced species occurrence data with known distribution of physico-chemical parameters of the aquatic milieu. The potential of this service is currently hampered by difficulties in accessing the necessary data and by the lack of the processing power needed to generate maps using large datasets. D4Science-II will create a VRE that will exploit data obtained from multiple sources (like GENESI-DR, OBIS, NEON and GBIF), specific services, and the computational capability of the Grid to support the AquaMaps service in the generation of **more precise, accurate, synthetic and predictive maps**. The D4Science-enabled interoperability will also make it much easier for other infrastructures to **take advantage of the predictive data generated by the AquaMaps VRE**.

The **FCCPS VRE**[2] has been developed to allow the FAO Fisheries and Aquaculture Department to both include additional data streams into the production of country-level fisheries and aquaculture reports and better manage those streams. It will provide an environment where scientists can query, access, analyze, and report on a broad range of fishery indicators available in heterogeneous data sources. D4Science-II will dynamically extend this VRE, profiting from the larger set of resources that will become available. By exploiting the interoperability offered, especially since this will be doable programmatically, this VRE will **facilitate the requested data sharing** across fishery, biology, commodity, and taxonomic information systems that have already embraced the sharing approach, such as the Ocean Biogeographic Information System (OBIS) and the Global Biodiversity Information Facility (GBIF). Furthermore, it will make possible to maintain at least some of the intermediate data generated when producing the relevant country reports, which are very valuable to scientists but are now discarded. The end goal is to **both serve and receive a greater variety of data** to cover more countries and serve more communities.

The **ICIS VRE** brings together a group of regional fishery bodies (RFBs) in charge of monitoring and managing fisheries and aquatic resources in the world's oceans. It provides a framework for the import and harmonization of these datasets as well as species-distribution data generated algorithmically from environmental and occurrence data and by expert knowledge. VRE specific functionalities allow for catch reallocation scenarios to **improve coarse-grained catch data** by combining it with species occurrence data. D4Science-II will extend the scenario with improved species occurrence data due to the synergy with AquaMaps; it will also allow the inclusion of Vessel Monitoring System Catch data at extremely high resolution to improve aquatic resource assessments. Dealing with these new types of data will require

[2] FCPPS and ICIS are actually extensions of VREs already operational in the D4Science e-Infrastructure.

extensions of the current tools, which will in turn require the power of the ecosystem to run. In addition, its programmatic interface for interoperability will give the dynamic data import/export capabilities demanded by the community for using harmonised re-allocated data both within D4Science and within their own legacy systems.

OBJECTIVE 3: Sustainability of the D4Science e-Infrastructure, Exploitation and Promotion of the Ecosystem

D4Science-II will take steps to ensure the sustainability of the D4Science e-Infrastructure ecosystem (and therefore, of the D4Science e-Infrastructure, as its centrepiece). To become a backbone of excellence in European research, the life of the ecosystem will have to extend beyond the life-cycles of this (and any potential future) project. In this direction, D4Science-II will establish synergies with initiatives such as EGI and the DRIVER Confederation and investigate relevant models for sustainability. Still it is already clear that sustainability of the e-Infrastructure ecosystem will only be ensured if there are always new e-Infrastructures that greatly enhance their individual capabilities by joining it and staying with it. In turn, this is critically dependent on these e-Infrastructures bringing new scientific communities that enthusiastically embrace the D4Science-II ecosystem because they need to use its unique cutting-edge services to solve their problems. Hence, it becomes evident that any effective way to sustainability must be deeply intertwined with the **dissemination**, **consulting**, **training**, **technological support**, and **standardisation** activities of the project. D4Science-II aims at defining strategies to **promote the emerging ecosystem approach** and the use of the D4Science production infrastructure as the e-Infrastructure mediator, and to **demonstrate the advantages** thus obtained for the scientific communities represented in the consortium. Profiting from the strong networks of the D4Science partners, which include International Organizations participating in many international projects and initiatives on global research challenges, D4Science intends to percolate these advantages to other fields of science, at a national, regional, pan-European, and international level, thereby .progressively extending its interoperability links and clientele far beyond the current consortium. Ultimately D4Science-II plans to build on the benefits offered to these scientific communities to engage them in the enrichment and long-term shepherding of the ecosystem, effectively combining the fate of research excellence with the performance of the e-Infrastructure ecosystem. Many European and international projects/initiatives have already approached the D4Science project to establish synergies and exploit the capabilities of its production e-Infrastructure, giving good indications about its sustainability prospects[3].

1.2 Progress beyond the state-of-the-art

D4Science-II will operate within a landscape of intensive worldwide activities in infrastructure deployment. This Section overviews those that are most closely related to the aims of the project and highlights the advances to the state of the art that the project will bring about in the target applications domains and in the area of cross-infrastructure interoperability at large.

E-Infrastructures

Initiatives for the deployment of e-Infrastructures are well underway across the globe. Long-term national and international deployment efforts for network, computational, and data Grids may be found in Europe (GÉANT [31], EGEE [24], DEISA [19], NGS [56], D-Grid [20], NDGF [54]), the United States (TeraGrid [69], OSG [63]), China (CROWN [15]), Japan (NAREGI [53]), India (Garuda [28]), Australia (APAC [4]), and the countries of the Pacific Rim (PRAGMA [65]).

Focusing on higher levels of abstraction, infrastructural support shows a greater variety of forms, both in the context of individual disciplines (domain-specific services) and, more horizontally, within the scope of specific functionalities (cross-domain services). Of direct relevance to D4Science-II are a number of repository services that are key to the evolution of applications currently hosted on the D4Science infrastructure (e.g. NeOn [55], OBIS [14], GBIF [29], GOOS [38], UN-DATA [73], COMTRADE [72]) and, most noticeably, the infrastructural initiatives introduced in Section 1.1 and selected for inclusion in the knowledge ecosystem targeted by the project:

[3] The Letters of Support reproduced in Appendix B exemplify some of the liaisons already established.

GENESIS-DR. In line with long-term initiatives for Earth Observation (EO) in Europe (GMES [37]) and worldwide (GEOSS [34]), GENESIS-DR is an infrastructure for the integration, validation, preservation, and uniform dissemination of data which originates from space, airborne, and in-situ sensors and is stored – along with related tools, models, and knowledge artefacts – in digital repositories maintained at a number of Earth Centers dispersed all over Europe [32].

INSPIRE. It is a new-generation repository offering an integrated, enhanced, and personalized access to the entire corpus of the High-Energy Physics (HEP) literature [41]. The service federates a number of key content and metadata providers within the HEP community (SPIRES [67], ArXiv [5], CDS [12], ADS [1]) and is positioned to act as the main reference tool for the HEP community worldwide. It will serve a user basis of over 30,000 researchers, in Europe, the United States, China, Japan and beyond.

AquaMaps. It is a service for the generation, standardized dissemination, and mapped visualization of model-based, large-scale predictions of currently known natural occurrence of marine species [45]. The service reflects a holistic approach to species distribution modeling which resorts to expert knowledge and habitat usage data to compensate for the potential unreliability and/or biases which are associated with sampling and collation of species occurrence data [46]. A joint project of FishBase [27] and SeaLifeBase [64], AquaMaps is widely used within the biodiversity community as a basis for species distribution inference.

DRIVER. It is an infrastructure for the integration and dissemination of metadata about widely distributed, autonomously managed, and cross-discipline content, particularly research outputs natively stored in institutional digital repositories [22]. It offers a successful demonstration of the potential of e-Infrastructures for sustaining aggregation services that, at much higher cost at any given scale, have already proved popular in the practice of Digital Libraries [49]. Today, through the aggregated metadata, DRIVER offers uniform, personalized, and instantaneous access to an information space distributed across more than 170 repositories and 20 EU countries.

D4SCIENCE. Among application-level infrastructures, the case of **D4Science** is unique rather than prototypical. Its control-oriented and cross-domain content-oriented services – to store, describe, annotate, search, select, merge, and transform information – leverage the functionality of the gLite middleware [36] in order to virtualize the computational and storage resources of the pan-European EGEE infrastructure and, in turn, the connectivity fabric of the underlying GÉANT infrastructure. The infrastructural services are organized into a distributed software system, *gCube* [30], which provides an end-to-end solution for the interactive construction and lifetime management of: *(i)* the whole infrastructure, *(ii) Virtual Organizations* of community-bound services and resources which operate within the scope of the infrastructure (VOs), and *Virtual Research Environments* (VREs) for the execution of community-specific applications within the scope of specific VOs. The solution includes:

- the dynamic selection, configuration, and secure orchestration of infrastructural services, service-oriented applications, and resources across the infrastructure, specific VOs, and specific VREs;
- the strategical deployment of infrastructural and application services across the infrastructure, their monitoring and their redeployment on demand according to QoS requirements and in response to failures;
- the provision of sophisticated administration and development tools – dedicated containers, application frameworks, and domain-level libraries – to simplify the evolution of the system, the installation, configuration, and maintenance of the infrastructure, and the development of applications which target gCube as their runtime platform.

Currently, the D4Science infrastructure hosts a number of scientific applications for communities that operate in the areas of Environmental Monitoring and Fisheries and Aquaculture Resources Management. Among these, the following VREs from the Fisheries VO are particularly instrumental to the goals of the project:

- The *Fishery Country Profiles Production System* (**FCCPS**) is a prototype application for the semi-automated generation and life-cycle management of complex reports that profile the state of fisheries in individual countries. Synthesizing in a country-specific format data that originates in a number of heterogeneous and autonomously managed sources – including catch data, vessel data, trade data, and biodiversity data – the generated profiles are vital information sources for fishery managers, fishery economist, and aquaculture specialists and provide key inputs to decision-making and advocacy activities related to the sustainable use and conservation of fish stocks.

- The *Integrated Capture Information System* (**ICIS**) is a prototype application for the integration, harmonization, re-allocation, and uniform dissemination of regional and global catch statistics originally compiled and hosted in a variety of distributed data sources. In particular, the application offers GIS services based on pre-computed and dynamically computed species distribution maps, and emphasizes the on-demand configuration of their inputs to reflect the complementary perspectives of marine biologists, oceanographers, climatologists, GIS experts, socio-economists, and fishery managers on the overall assessment of aquatic species.

Cross-infrastructure Interoperability

Interoperability across infrastructures is the current focus of a number of international standardization activities, but so far it has been mostly considered in relation to the interface and semantics of key Grid middleware services. It is within this context that international standardization efforts such as OGSA [26] and the OMII-Europe [59] initiatives are best understood: the first rallying consensus on the specification of abstract interfaces to core middleware services; the second re-engineering implementations of key OGSA specifications out of the components currently available in widely deployed middleware platforms. OGSA is the flagship output of a broader and well-established forum for standardization of Grid technologies, the OGF [61]. Within the OGF, the GIN group is devoted to the organization, management, and demonstration of ad-hoc, bilateral interactions between production-level Grids (e.g. EU-INDIAGRID, EGEE-OSG, EGEE-DEISA).

Considerably less attention has been given to the problem of resource sharing across application-level infrastructures, where the lack of a reference architecture for infrastructural services and the heterogeneity of the data handled by domain-specific services are expected to result in additional challenges of systemic, syntactic, structural, and semantic interoperability [68]. In particular, the deployment of mechanisms for cross-infrastructure interoperability will introduce problems of heterogeneity of application-specific data models, metadata models, and ontologies. These issues are well-known in the area of distributed Digital Libraries, where they have motivated the adoption of a number of official or de-facto standards for data representation, packaging, exchange, harvesting, and search, both within application domains (e.g. ISSCAAP [16], ISSCFC [16], ISSCFG [16], ISSCFV [16], KML [60], GML [42], NAF [71] in Fishery Management alone) and across application domains (e.g. DC [23], DIDL [43], METS [51], MPEG [52], OAI-PMH [47], OAI-ORE [57], OWL [76]). Such issues, however, remain largely latent in the Grid discourse.

D4Science-II will be one of the first systematic attempts to tackle the problem of interoperability from the perspective of a high-level infrastructure. In line with the vision and objectives presented in Section 1.1, the main contribution of the project will be *the construction and exploitation of the first ecosystem of application-level infrastructures which accommodates the whole spectrum of requirements for the interoperability of the involved parties*.

For D4Science, the pivot role within the ecosystem will translate in a much stronger value proposition for the communities directly or indirectly served by the infrastructure. In particular, the methodology adopted by the project to build the ecosystem (cf. Section 1.3) will augment the infrastructure *with mechanisms and policies for interoperability that are sufficiently general to promote the exchange of resources and the export of its functionality with other infrastructures, repositories, and services beside those within the purview of the project.*

For the communities served by the infrastructures that comprise the ecosystem, *the exploitation of the aggregated resources will result in innovative applications that the infrastructures cannot currently support in isolation from the ecosystem and that – for reasons related to risk, cost, and scope – they often exclude altogether from their roadmap for evolution*.

Through the mediating role of D4Science, in particular, DRIVER, INSPIRE and AquaMaps will be able to draw from the computational and storage resource that are available in EGEE, and to use them to perform a range of processing-intensive analyses over the primary and secondary data sources that are already *referenced within their rich publications models but not otherwise exploited*. In both cases, the exploitation is hampered by stringent requirements of resource pooling which lie outside the design assumptions of the infrastructures, and which typically exceed the relatively modest ICT profile of the digital repositories they federate. Beside the lack of an adequate physical fabric, the functionality required to virtualize, aggregate and manage data and processing intensive computations is also missing. Native provision of such functionality remains unforeseen, partly due to its cost and partly because it would risk to compromise the stability of services that are already in production. As a result, sophisticated bibliometric analyses, usage pattern detections, topical clusterings, advanced approaches to indexing and distributed retrieval are all possibilities that remain latent within the infrastructures.

To provide its users with better biodiversity maps, *AquaMaps will follow a similar pattern of exploitation of the knowledge ecosystem to increase the sophistication of its predictive models. In addition, it will feed into those models the geo-morphological and ecological data sources that originate in GENESIS-DR and GOOS and will become available and interoperable within the ecosystem*. Though already innovative in its holistic approach to species distribution modelling, the service is currently under-resourced to address well-known modelling issues that affect the overall reliability and usefulness of its maps. In particular, lack of access to adequate computational facilities and relevant data sources impedes the adoption of improved modelling techniques. For example, current algorithms renounce high precision to avoid the cost of mapping species distribution at small scales. For analogous reasons, the accuracy and usefulness of current maps suffer from the inconvenience of recomputing them more frequently (e.g. to reflect the review of experts), comparing them with other maps (e.g. to assess existence and extent of deviations, to group them taxonomically or thematically with respect to species), and producing them simultaneously at local, regional, national, and global levels (e.g. to support decision-making processes for the design marine protected areas). Computational problems co-occur with problems of limited interoperability with external data sources that may be fed into the multi-faceted model used by the service. Some auxiliary sources are already exploited by the service, if not fully integrated (e.g. FishBase, OBIS, GBIF). However, other key sources of geo-morphological and ecological data (bathymetry, salinity, temperatures, productivity, etc) still lay beyond its reach. Clearly, solving the associated interoperability problems would further increase the requirements for data and computationally intensive processing**.**

In turn, the improved maps will feed back into D4Science and enable improved application services for the communities already served by the infrastructure. Specifically, the ICIS VRE *will gain access to novel and more varied data sources – species distribution maps from AquaMaps, physical ocean data from ARGOS sensors maintained by GOOS, climate and sea surface data from GENESIS-DR, ecosystemic data from OBIS/GBIF, and Vessel Monitoring data from its Regional Fisheries Management Organisations – to improve the precision of its harmonization and spatial re-allocation criteria and to explore their extension towards ecosystemic data*. With an independent but otherwise similar pattern of resource sharing, the *FCPPS VRE will be able to extend the geographical coverage and the dissemination reach of its country profiles by exchanging legal regulation, natural resource, economic, and social data with UN agency systems such as UN-DATA and commodity and taxonomic information systems such as OBIS, GBIF, and COMTRADE*.

Overall, the project will deliver applications that are of great interest and immediate usefulness to the communities served by the parties involved in the knowledge ecosystem, and that exemplify important patterns in the cost-effective exploitation of the resources and functionality collectively aggregated within the ecosystem. Accordingly, *the project will offer a model of ecosystem building and exploitation that is suitable for adoption by other initiatives within the broader context of global infrastructure development*.

1.3 Methodology to achieve the objectives of the project, in particular the provision of integrated services

In order to meet the stated objectives, D4Science-II will perform the activities for the implementation and usage of the knowledge ecosystem by: (*i*) enhancing the D4Science e-Infrastructure technological and operational capabilities (Objective 1); (*ii*) setting-up a number of VREs for scientific application scenarios in the domain of the D4Science e-Infrastructure (Objective 2) and (*iii*) identifying appropriate models for successful outreach and sustainability (Objective 3).

As far the enhancement of the technological capabilities are concerned, D4Science-II will **extend gCube** (the D4Science e-Infrastructure enabling system) with the interoperability mechanisms that are required to support the realization of the ecosystem. In particular, it will introduce new **mechanisms for minimizing the *ad-hoc* development** of solutions to discover, access and use resources published by different heterogeneous data e-Infrastructures. The design of these mechanisms will be strongly driven by the experience acquired in the framework of the D4Science project.

In the majority of cases, it will be not realistic to assume that these e-Infrastructures will change their systems and their access, validation and security policies in order to participate to the knowledge ecosystem, at least in its initial phase. Therefore, particular attention will be given to the identification of interoperability solutions that *preserve as much as possible the autonomy* of the participating data e-Infrastructures.

Further, the exploitation of D4Science resources will be promoted by transforming the gCube VRE mechanism into a **commodity** that can be programmatically consumed by other interoperable e-Infrastructures. Thus, any e-Infrastructure, as any other third-party service, will be able to exploit VRE operation and management capabilities for the controlled virtualization and contextualization of shared resources. In particular, these actors will be enabled to outsource to D4Science VREs the processes that are required to extract knowledge from data stored in a variety of collections and in the scope of different e-Infrastructures. The only limitation to the access and usage of these data will be determined by the policies stated by their owner e-Infrastructure. Once generated, the new knowledge can be maintained by a D4Science content service or by a service of the requesting e-Infrastructure. In either case, D4Science will exploit the realised interoperability to disseminate the new knowledge to all the other components of the knowledge ecosystem.

As far as the operational aspects, the project will maintain the D4Science production e-Infrastructure by guaranteeing its quality of service. Moreover, it will identify the **procedures** needed for the definition of agreements that clarify relationships and responsibilities between the D4Science-II knowledge ecosystem and other data infrastructures sharing services, clients, and resources.

The first nucleus of the knowledge ecosystem set up by D4Science-II will consist of the four main components mentioned above, i.e. GENESI-DR, DRIVER, INSPIRE and AquaMaps, and many other repository services (cf. Appendix A). Collectively, such components make available a large number of data collections (cf. Appendix A). These components are very heterogeneous with respect to many factors, including the type and organisation of content, the kind of supported functionality, the representation and management of user profiles, the policies, and, more generally, the adopted technology. All these aspects will be taken into account when developing the mechanisms that enable interoperability.

For what concerns objective 2, **a number of innovative Virtual Research Environments (VREs) serving challenging application scenarios** will be created by exploiting the D4Science-II knowledge ecosystem. Some of these VREs will be consumed by scientific communities operating in the D4Science infrastructure domain, while others will be consumed by the services of other scientific e-Infrastructures. In either case the ultimate beneficiaries will be the scientists since through D4Science-II they will have on-demand access to the large resources of the Grid; they will have at their disposal a working environment in which they can transparently exploit cross-domain shared resources; and finally, they will have better opportunities to export the produced knowledge and promote discipline cross-fertilization.

Regarding objective 3, all of the above activities will be performed taking into account that the project's long-term desired outcome is a broadly populated ecosystem serving a growing number of scientific communities. In order to create the conditions for achieving this objective, the project will give special attention to the factors that may influence its success. In particular, the project will include activities, such as the participation to standardization bodies, that will facilitate the adoption of emerging standards. Focused dissemination and training activities will also be largely conducted to promote the knowledge ecosystem vision. Finally, as success will strongly depend on the sustainability of the D4Science infrastructure and the entire ecosystem, an appropriate plan for future sustainability will be defined by establishing synergies with other on-going European initiatives.

The rest of this section elaborates further on the activities and methodologies described above. It is organized as follows. Section 1.3.1 describes the technical and service activities that will be performed to enhance the facilities offered by the D4Science e-Infrastructure. Section 1.3.2 introduces the characteristics of the data e-Infrastructures that will become interoperable with D4Science in the target knowledge ecosystem. Section 1.3.3 introduces the scenarios that will be addressed by the project and how these will exploit and contribute to the knowledge ecosystem; it also illustrates the work that will be carried out to create the VREs and the new facilities that they will offer to the target scientific communities. Section 1.3.4 discusses the activities of the project to promote the knowledge ecosystem among other players, e.g. e-Infrastructures and scientific communities, and to identify appropriate sustainability models for it. Finally, Section 1.3.5 provides an overview of the work packages that will implement these activities and on the main interactions that will take place among them.

1.3.1 Building the Ecosystem: D4Science Enhancement

As previously discussed, D4Science is the first model of an application level e-Infrastructure to define an explicit abstraction for the Virtual Research Environment. With its VREs the infrastructure can serve different application scenarios by letting scientists choose from a pool of community resources that are needed to support specific collaboration needs. Such resources range from computational and storage resources – largely and implication of interoperating with EGEE – to data collections and cross-domain content-oriented services whether legacy services or services developed specifically for D4Science. It is because of its VRE management facilities that the D4Science infrastructure can ply a pivot role in the formation and operation of the target knowledge ecosystem. In particular, the D4Science e-Infrastructure will become the *aggregator* of all the resources that populate the knowledge ecosystem and, through the implementation and operation of VREs, a *broker* for disseminating those resources within the ecosystem. To fulfil these expectations two main lines of enhancement activities will be carried out for D4Science: (*i*) the **enhancement of gCube**, its **enabling technology**, and (*ii*) the **consolidation and improvement of its management policies and procedures**.

gCube [30] is the distributed service-oriented software system originally implemented during the DILIGENT project to prove the effectiveness of application-level e-Infrastructures for Earth Observation and Cultural Heritage communities [11]. This system has been designed and implemented by extending the specifications of the *Open Grid Service Architecture (OGSA)* [26] toward the specific goals and requirements of an application level e-Infrastructure. The principles borrowed from OGSA are: (*i*) the *grid approach to resources sharing*, concerning the integration, virtualisation and management of services and resources in a distributed and heterogeneous environment and (*ii*) the *service as the coarse-grained unit of functionality in a distributed environment,* in particular its assumptions of loose coupling and alignment with open Web standards. The architectural extensions were related to the introduction of services for with the provision of (*i*) advanced content management facilities [66][11], to store, describe, annotate, transform and retrieve compound and multimedia information objects, and (*ii*) Virtual Research Environments definition and management facilities [6]. One of the objectives of the ongoing D4Science project is the consolidation and enhancement of the gCube technology to address (*i*) new requirements that emanate from *Earth Observation* and *Fishery and Aquaculture Resources Management* communities and (*ii*) stringent QoS requirements of service provision within a *production-level* e-Infrastructure. In particular, consolidation activities seek to improve: (*i*) *performance* and *reliability* of all gCube services; (*ii*) *VREs definition and management* mechanisms (e.g., enhancing resources discovery and dynamic deployment); (*iii*) *scalability* and *granularity* of *security mechanisms* (e.g. Access Control Lists for fine-grained sharing policy); (*iv*) *e-Infrastructure management* tool support (e.g., a multi-view monitoring system); (*vi*) *processes*[4] *definition and orchestration* tool support (e.g., usability of process modelling tools, interoperability with other applied technologies such as OGSA-DQP [2]), (*v*) *advanced information management* facilities (e.g. information model extensions for new types of information), (*vi*) *information retrieval* capabilities (e.g. support for non-cooperative information sources and OGSA-DAI/WS-DAIX [3] compliance); (*vii*) *presentation services* functionality (e.g. single-sign on); and (*viii*) *enhanced reports*[5] *definition and generation* facilities (e.g. enhanced templates) [10]. D4Science-II will start from the results of

[4] Also known as *workflows*, are services built by aggregating existing services into execution flows.

[5] Also known as *living document*, are innovative Information Objects whose constituents are either pre-filled (e.g. a human edited descriptive text) or dynamically generated (e.g. a table or graph obtained by processing a pool of data).

D4Science consolidation and enhancements activities – e.g. the gCube data model will be ready to handle the majority of information types that are expected in D4Science-II – to further improve the gCube. The following enhancements will be implemented in D4Science-II:

- design and development of a gCube ***Interoperability Framework***. The challenge of interoperability in the context of a knowledge ecosystem is a non trivial, for it raises different classes of requirements (for *data* vs. *information* vs. *knowledge* interoperability) against different types of information and thus requires different *approaches* that rely on different *technologies*. In spite of this complexity, D4Science-II will define a framework that enriches the gCube system with off-the-shelf solutions and reference implementations for common interoperability issues, and thus provide gCube developers with innovative facilities for developing specific applications. The primary content of the framework will be based on the approaches and solutions identified while addressing the interoperability issue in the context of the knowledge ecosystem target infrastructures (cf. Section 1.3.2). Candidate technologies include, but are not limited to OGSA DAI and its follow-up technologies / specifications (WS-DAIX), OGSA-DQP;
- reinforcement of the gCube ***resources model*** to provide ecosystem components with a uniform, comprehensive representation of the resources – and their distinguishing features – that are pooled within the knowledge ecosystem. As the resource pool is intrinsically large and heterogeneous – including data sources, collections, services, computing and storage resources – the extensibility of the model is a core requirement. Moreover, since such the representation granted by the model will allow humans and services to *understand* the resource and to perform informed actions, they must be kept in sync with the true and current status of the resource;
- consolidation and enhancement of the gCube ***Security Framework***. An innovative set of components and interfaces aiming at automating and homogenising the federation of digital identities and access control across different administrative domains will be defined and implemented. As a result the gCube VO Management components will be equipped with facilities providing users and services with a uniform set of interfaces to manage security in a cross-technology and cross-domain environment;
- consolidation and enhancement of the *gCube Application Support Layer* toward a gCube ***Access Interface***. gCube is already equipped with a component that abstract over the gCube internals and aim to provide a unique access point to its features. This component was initially conceived to serve the needs of graphical user interface developers willing to implement applications for consuming gCube facilities without delving with the full complexity of the system. In D4Science-II these mechanisms will be extended in a new direction to support service providers willing to build upon gCube facilities and resources in order to implement their business logic via a programmatic access. For instance, gCube Information Objects can be exposed through the OAI-ORE mechanism [57], the complex WSRF based gCube Information Space can be accessed by plain web services. These innovative exploitation patterns are appropriate for a knowledge ecosystem and will be used to satisfy the requirements of some of the D4Science-II scenarios (e.g. to outsource some computationally intensive processes to the ecosystem or seamlessly accessing cross-domain heterogeneous collection of data (cf. Section 1.3.3));
- reinforcement of the ***data (***and ***metadata) representation, abstraction*** and ***virtualisation*** mechanisms toward a true *information broker/mediator service*. Information is probably the most important resource in a knowledge ecosystem. Because of the *design and implementation autonomy*, information represented in very different ways and with very different technologies. Metadata are attached to it in order to explicitly capture data peculiarities and promoting data consumption. Unfortunately this solves the problem of information heterogeneity only partially, rather it shifts it toward metadata representation and interpretation. gCube is already equipped with a set of facilities for managing multiple data and metadata formats by which data and metadata may be pre-computed or generated on demand. These mechanisms need to be improved in various aspects including the performances of the approaches, the set of available virtualisation and reconciliation mechanisms, the application domains and formats supported. Exploitation of cutting edge approaches for information interpretation, such as ontology based inference is among the facilities to be exploited in the D4Science-II activities;
- extension of the ***gCube Information Retrieval framework***. gCube is equipped with a very powerful and extensible Information Retrieval framework. New modules will be added to such a framework to make it interoperable with the new kind of information sources partaking the ecosystem (e.g. modules for

exploiting the OpenSearch protocol [62] to issue queries against OpenSearch compliant data sources, or exposing its operators as OGSA-DAI sources, or extending its query language to capture new complex forms of data, or spanning its queries in external distributed query engines via standard compliant languages;

- consolidation and enhancement of the ***gCore Framework***, i.e. the framework supporting the development of high-quality gCube services and service clients.[6] This foundational component will be reinforced with additional abstractions and reference implementations of common design patterns that will improve further the development of gCube services and their quality.

In addition to these development activities, software engineers will continuously monitor the status of ***(emerging) standards*** and design innovative solutions and approaches benefitting from their exploitation in gCube services.

Besides the gCube software, D4Science-II project results will produce an operational application level e-Infrastructure.

The setup of the D4Science infrastructure is a work that was initiated by the DILIGENT project. In 3 years lifetime, DILIGENT built a first prototype, a testbed, that proved the feasibility and value for user communities to dynamically allocate resources (computing, data and services) and compose ad-hoc virtual research environments. D4Science moved this concept into production in an infrastructure that aggregates, shares and dynamically allocates computing resources, services and community specific content or service resources.

In D4Science-II, this production infrastructure will evolve into a knowledge ecosystem of collaborating e-Infrastructures. This will be achieved by enhancing the enabling technology, gCube, but also extending its infrastructure management activity with **policies and procedures governing its interoperation** with collaborating scientific infrastructures, repositories and services. These are needed for the definition of agreements that clarify relationships and responsibilities between the D4Science-II knowledge ecosystem and other data infrastructures sharing services, clients, and resources. They will define which part of another infrastructure will be accessible by D4Science services (e.g. access to copyrighted resources is permitted in restricted contexts only) and how (e.g. only local elaborations of data, security policies, frequency of updates, etc.). On top of this, a formal resource sharing policy will have to regulate the overall sharing and usage of the interfaced community specific resources, as well as the access and allocation of the D4Science computing resources. This is needed in order to ensure a peaceful and satisfactory cooperation between the infrastructure user communities and the resource providers.

In addition, during the D4Science-II project lifetime, the operation of the D4Science infrastructure will also be enhanced in the technological support by: (*i*) improving the infrastructure operations looking for more efficient (less amount of effort) and effective (better results) ways to deliver production-quality services. This will be done by applying **best practices procedures for services operation** and by requesting JRA to produce a **higher level of automation and operability** of the gCube software; (*ii*) **optimising the resources usage** for VOs and VREs exploitation.

There is in fact the need to improve on processes as times change and things evolve. In the specific case of the D4Science-II project, the application of good operating practices will touch on several aspects, between them service delivery and service support, and might induce to investigate other domains as the maturity and needs of the infrastructure evolve.

Together with the procedures, a higher level of automation and operability will also be required in the software operated in the infrastructure, gCube, in compatibility with its evolution. In particular, operational requirements will be formulated and implemented in relation with the VOs and VREs administration portal, the core management tool of the infrastructure. This will have the double effect of reducing the technical support needed by the VRE managers and VO managers to operate the infrastructure, as well as improving the service support activity with more efficient first level support and incident management process.

---

[6] This application framework allows gCube services to abstract over functionality lower in the web services stack (WSRF, WS Notification, WS Addressing, etc.) and to build on top of advanced features for the management of state, scope, events, security, configuration, fault, service lifetime, and publication and discovery.

Thirdly, there is also the need to further investigate on the infrastructure services deployment requirements and on the strategy to optimise the usage of the available resources and increase services reliability and performances (e.g. with service replication).

Finally, a long-term **sustainability strategy** for the resources feeding the infrastructure will be investigated. Cooperation with other initiatives, like EGI, will be established. The user communities vision of the long-range future of their members, and the links to the environmental, economic, and social aspects of the community will be taken into account.

1.3.2 Building the Ecosystem: Participating Infrastructures

Objective 1 will be met by constructing an ecosystem composed by the enhanced version of D4Science and a number of other prominent e-Infrastructures and repositories. These serve a huge number of scientists by making available several millions of heterogeneous primary and secondary data. A brief description of the four main services that will be integrated in the ecosystem is given below. Each description highlights some of the issues that will be taken into account when realising interoperability solutions between the services and D4Science.

1.3.2.1 GENESI-DR

As a heterogeneous digital **Earth Science repository e-Infrastructure**, GENESI-DR establishes a dedicated infrastructure providing transparent access and allowing Earth Science communities to easily and quickly derive objective information and share knowledge based on all **environmentally sensitive domains**. It adopts a **common subset of well-known standards for geographic metadata** which has been agreed in the open source geospatial community. This subset and the search interface that it implements are designed to comply with the Implementing Rules for Discovery Services in the European Union drafted by the Infrastructure for Spatial Information in Europe in a minimal way. The system supporting GENESI-DR provides quality solutions for the machine-readable and repeatable properties of data: unique identifiers which can be used to annotate the provenance and processing history of data sets; URLs to represent license constraints; contact details for people responsible for data sets. To achieve this, a minimal model for metadata about geographic information, **DCLite4G** ("Dublin Core Lightweight Profile for Geospatial"), is used. This model is a Dublin Core Application profile defined following the Draft Implementing Rules for the European SDI Directive. It consists of: (*i*) an abstract information model; (*ii*) a reference implementation exploiting common standard vocabularies (e.g. Dublin Core, GeoRSS); (*iii*) a namespace used to define extra properties needed to usefully specify the properties of geospatial data. The URL at which the namespace above lives, contains an OWL ontology showing the structure of metadata schemas and providing mappings between them, where possible.

From the technical point of view, GENESI-DR exploits the open source ***duetopia*** software [70], developed by Terradue srl. This software supports:

- The publication of metadata into well-known registries and indexes
- The syndication of data between peers automatically
- The discovery of data relevant to an application
- The establishment of clear rights to reuse data, minimising negotiation

The current implementation of the machine-machine interface supported by *duetopia* uses **OpenSearch (with Geo extensions)** to provide an Atom feed of geographic metadata in RDF/XML. A "lego model" approach is taken allowing one to glue open source components to the edges of the existing systems. Taxonomy and ontology are left very open while the use of URI schemes for keywords is recommended. Data annotation is distributed with the emphasis not on how the data is described, but on how it is used.

1.3.2.2 DRIVER

The DRIVER Infrastructure is a Service Oriented e-Infrastructure providing an environment where different organizations can (*i*) find the tools to aggregate **heterogeneous OAI-PMH compliant Institutional Repositories** into uniform shared Information Spaces and (*ii*) build and customize their Digital Library applications to operate over such Spaces. Nowadays, the information space operated by DRIVER counts around 700,000 records out of 150 repositories across 21 countries, and the number of candidate registering repositories is still growing.

The infrastructure builds a Service-oriented environment, where distributed resources are implemented as standard Web Services. Services are registered to a central **Information Service** (Registry) and are orchestrated by special Manager Services so as to match established application functional needs. Specifically, one instance of the DRIVER infrastructure can scale up to arbitrary numbers of service instances and applications, here intended as "regions" of services orchestrated by a Manager Service to satisfy certain functional expectations. Services can be shared across several applications whose boundaries are ensured by secure communications exploiting **XACML authentication and authorization services**. The infrastructure implements the WS-* set of specifications, ranging from service-to-service communication protocols (**SOAP**) to subscription and notification protocols (**OASIS Standards WS-BaseNotification 1.3** [39] and **WS-Topics 1.3** [74]).

The DRIVER application framework is independent of the functionalities offered by the services, i.e. it is open to the introduction of services providing new functionality. Currently, the DRIVER infrastructure offers the services required to build distributed aggregation systems. Such applications enable the construction of information spaces hosting **XML metadata records** harvested from potentially heterogeneous Institutional Repositories. Aggregation Services offer advanced tools for harvesting, cleaning and integrating metadata records according to common, possibly richer and more qualified, metadata record formats. The resulting records are stored and replicated into the Store Services of the infrastructure and then fed into the available Index Services, which in turn expose their content through CQL-SRW standard interfaces. The implementation of the Aggregation, Store and Index Services is independent of the input and output metadata formats, which can be defined and modified at run-time in order to match the Information Space specification at hand. All instances of such services can be discovered through the Information Services and be accessed through their standard interfaces and protocols by authorized consuming applications.

The current operational e-Infrastructure is being extended to enable it to deal not only with simple documents but also with **enhanced publications**, i.e. complex structured ones with bindings to source data. Further, all the services and interfaces are being modified to deal with a new document model able to represent these new types of content.

Currently, the DRIVER e-Infrastructure information space is accessed by three distinct communities, each having its own content and functional requirements, through independent web user interfaces, namely, the **Belgium national portal**, offering search on the Belgium Repository Federation; **Recolecta**, offering search on the Spanish Repository Federation; and the **DRIVER portal**, providing access and advanced functionality over the whole information space.

1.3.2.3 INSPIRE

Today, three main information resources serve the field of High-Energy Physics which counts 30,000+ practitioners: **arXiv** (Cornell, United States) [5], **CDS** (CERN, Switzerland) [12], and **SPIRES** (SLAC, United States) [67]. These include collectively over two million records concerning HEP publications, with full **metadata information** and about half a million Open Access **full-text scientific documents**. As it has been highlighted by a recent survey to which more than 10% of the HEP community answered, 91% of the HEP scholars use one of these community-based system for their information needs. In order to better respond to the emerging needs of this large community, pointed out by the user survey, the four leading HEP laboratories (CERN in Switzerland, DESY in Germany, Fermilab and SLAC in the United States) have embarked on the realization of INSPIRE, a next-generation repository. INSPIRE will be developed in synergy with other partners (notably arXiv), integrating the content of CDS and SPIRES and enriching the metadata of scientific publications in a continuous dialogue with major publishers in the field in Europe and the United States.

A live beta release of INSPIRE is planned for 2009, reproducing the current functionality, and presenting the aggregated content, of arXiv, CDS and SPIRES with immediate add-ons through a Google-like search of both metadata and full-text. INSPIRE will then further develop to meet the user request of advanced tools for citation analysis, hybrid metrics and web2.0 features.

From the technical point of view, INSPIRE is based on the **Invenio Open Source digital library software**

[13] which uses python/mySQL technology. Its records are held in **MARC21 format**, with additional custom fields. Additional tables used for citation analysis are stored in mySQL format. So far, INSPIRE was designed to be accessed via web queries from the end user. In addition its records can be exported in several formats, from XML to **OAI-PMH**. This functionality is designed for remote harvesting of limited subsets of pre-determined records. The rationale for this limitation is that INSPIRE was originally conceived as a **stand-alone application for HEP**.

As part of the D4Science-II project, INSPIRE will grow to become an interoperable part of an ecosystem of e-infrastructures. Benefiting from the capabilities of D4Science and contributing to the entire ecosystem, its wealth of bibliographic content will be of interest *per-se*, beyond the field of HEP, as it constitutes a stand-alone, complete, record of half a century of the scientific corpus of an entire discipline. A number of steps will be done in the project to enable INSPIRE to grow beyond its current inception: a **system of credentials** to allow trusted users to access large part of the records will be designed and implemented; one or possibly more **APIs** will be defined, tested, implemented and tried to allow other trusted partners of the e-infrastructure to perform complex queries in the system, e.g. bibliometrical analysis of all data for an immediate response or for short-listing a subset of records for future analysis. As a result of these extensions trusted members of the e-infrastructure will be allowed to harvest tagged records or part of them.

1.3.2.4 AquaMaps

AquaMaps is a data service, currently oriented to the marine environment, that makes available species range maps plus a variety of thematic maps (taxonomic, climatologic, invasiveness, etc.). Each map is a complex information object. Each point on the map is associated with the following additional information:

- Physical oceanographic parameters like sea surface temperature and primary productivity, derived from the Sea Around Us project, CSIRO, Kansas Geological Survey/NOAA and available in an AquaMaps file (HCAF).
- Occurrence point data with a basic minimum triplet (a scientific name, a geo-referenced location, a date); each element of the triplet may have various level of precision and accuracy. These data are 'crawled' from GBIF, OBIS, FishBase, SeaLifeBase whenever a new map is generated and are kept in a local cache.
- Ecological data at species level like depth and temperature preferences. These data are obtained from FishBase and SeaLifeBase whenever a map is generated or updated, and are stored in an AquaMaps file (HSPEN). These preferences can be edited by experts. The history of such edits is preserved.
- The probability of a species to occur at a certain point. This is the data that is stored in an AquaMaps file (HSPEC) and is used to create the maps. Previous maps are preserved.

All the AquaMaps collections are open source. Currently, several different communities of scientists access these collections, either through the AquaMaps user interface or through the interface of other systems (e.g. FishBase [27], the most widely used biological information system with over one million visitors per month; other species information systems such as SeaLifeBase [64] and GBIF [29]). All the maps and their data parts are downloadable under simple formats.

A generalization of the service to cover also maps from freshwater and terrestrial environments is in progress. This generalization will require access to other datasets disseminated among various institutions.

The current map interface was developed by CSIRO, C-Squares, using a methodology based on latitude-longitude squares. A move to the Google Earth API is ongoing.

The AquaMaps information objects are associated with a DarwinCore2 XML schema metadata description [18].

AquaMaps is developed under MySQL and PHP for the batch production of the species, aggregated (biodiversity) and topic maps, the query interface, and the validation process. It is accessible programmatically as a Web service. Clients of the Web service can include an interactive thumbnail version of a specific map in their own web portal.

1.3.3 Exploiting the Ecosystem: Scientific Application Scenarios

The knowledge ecosystem created through the realised interoperability and through the enhancement of VRE access modalities will enhance the support that the D4Science production e-Infrastructure can offer to scientists. In this new context, as addressed by Objective 2 of the project, new scientific application scenarios will become sustainable. The project will focus on five of them. Others will be analysed as part of feasibility studies. Below, a brief description of these scenarios is given. Each description highlights the requirements addressed and the facilities that the VRE supporting the specific scenario will offer. The new solutions, developed as part of the described VREs, are expected to have a considerable impact in innovating the operational context of the target scientific communities and in fostering new scientific and technological paradigms. All the activities performed in the framework of these VREs will produce new knowledge resources. Some of these resources will be stored in the D4Science own storage, others will be maintained by the diverse ecosystem components. Independently of the solution chosen, thanks to the realised interoperability and the supported virtualisation, the newly generated resources will be made available as sharable entities that can be used in the framework of any other VREs, hence exploitable in other e-Infrastructures. This will be the way in which, as dictated by the notion of ecosystem, the enhancements in a component will also propagate to and influence the surrounding ones. A sixth VRE, ImpECt (Implementation of Environmental Convention) [11][17] developed in the framework of the D4Science project to serve the Environmental Monitoring scientific community, will be maintained operational in the D4Science infrastructure. As it will not require new development it is not included below.

1.3.3.1 INSPIRE Scenario

Requirements: Mining of bibliometric data and calculation of hybrid metrics on the whole corpus of the HEP literature

Users: HEP scientists

Main e-Infrastructures used: D4Science, INSPIRE

Main D4Science exploited capabilities: dynamic service deployment, on-demand processing

Following the analysis of a user survey which elicited thousands of answers [33], INSPIRE plans to offer to its users more complex functionality [50] than existing HEP repositories have been able to do so far due to technical, architectural and computing-power limitations. Two main avenues will be explored, both requiring **computational intensive applications** performing sophisticated analyses on the whole corpus of the HEP literature:

- **Mining of bibliometric data**, such as co-authorship, cites-cited pairs as well as contextual information on the citation in order to offer multi-dimensional searching capabilities. Example applications are the exploration of social-science models describing the construction of scientific networks, the design, tuning and implementation of algorithms to identify experts in the field e.g. to automatically suggest referees/evaluators free from conflicts of interests, the discovery of cross-disciplinary links which are worth pursuing but are hidden in a few precursory works.
- **Calculation of hybrid metrics** to assess the impact of an article, a dataset, an individual, an institution, or a country by combining conventional impact metrics and integrating "new" information, such as server logs, text-mining for references beyond citations (i.e. names or quotes).

Since INSPIRE is currently planned as a repository, such computationally intensive applications, either on demand or in a batch/periodic mode, are beyond its scope. D4Science-II will provide a solution to this problem by transforming INSPIRE into a component of the ecosystem and enabling it to access programmatically a D4Science VRE that will offer the necessary environment for performing the computational intensive analyses. This VRE will serve the requests coming from the INSPIRE services by **dynamically deploying appropriate services in the D4Science production e-Infrastructure**. These services will perform the necessary analyses by accessing selected information from the INSPIRE service.

It is expected that this solution will not only have a strong impact on the INSPIRE system and the HEP scientific community. Other repositories or researchers active in cross-disciplinary fields will be able to try complex bibliometric algorithms on INSPIRE's large and self-contained data sample, accessing the INSPIRE information through its novel API from D4Science and possibly using the computational capacity of the Grid. The text-mining and data-mining services, which also go beyond the initial idea of INSPIRE and are **useful for other users of the ecosystem,** will become available as D4Science resources (see for example, the DRIVER scenario).

1.3.3.2 DRIVER Scenario

Requirements: high-value and innovative end-user services, support for enhanced publications, powerful personalization services

Users: scientists, National Community and Thematic repository portals

Main e-Infrastructures used: D4Science, DRIVER

Main D4Science exploited capabilities: access to the Grid, access to primary data, shared tools

The DRIVER service is now in production since May 2008. Through its portal, researchers can plug into a large shared information space and use scientific content using several user services, including search, data collection, profiling and recommendation. By analyzing the feedback received by early adopters a number of new requirements have been highlighted. Meeting these requirements impose to solve specific technological issues. Among them the followings:

- When trying to build **high-value and innovative end-user services** on top of DRIVER e-Infrastructure, the need for computation and storage intensive processing emerges. For example, specific, cutting edge information retrieval services, (such as inference mechanisms, metadata brokerage etc) can potentially be quite demanding, in terms of logic and resources. The DRIVER e-Infrastructure mainly builds on the repository concept which until recently has left out the need for computation and storage intensive processing.
- Access to external data sets is mandatory to support **enhanced publications**. These documents contain link to original data which might be parts of, or entire, huge data sets. These sets have to be referenced and acquired via formal means (specs for referencing and exchanging them). The DRIVER e-Infrastructure at the moment does not yet include considerable data repositories to support an extensive exploitation of this facility.
- Good quality user services increasingly depend on **personalization capabilities**. The realization of these capabilities requires the availability of precise bibliometric analysis on large collections of data and user behaviour recordings. Only very simple information of this type is available in DRIVER at the moment. More sophisticated analysis tools would be needed.

In the direction of filling the gaps for DRIVER to achieve excellence in its domain, D4Science-II project will build bidirectional gateways to the D4Science production infrastructure which will facilitate the provision of the above facilities. In particular, a DRIVER VRE, accessible programmatically, will be created offering:

- **Access to computational intensive services** able to perform metadata brokerage and data transformations on the D4Science production e-infrastructure, and even ontology-based mapping.
- Support for the creation, storage and management of enhanced publications. By exploiting these facilities DRIVER users will be enabled to create, retrieve and access publications that link DRIVER hosted content with **any other data available under appropriate policies in the ecosystem**.
- Analysis of DRIVER space content through advanced services deployed in the D4Science e-infrastructure. In particular, it is expected that the created VRE will include as accessible resources the **services implemented in the framework of the INSPIRE scenario** for the mining of bibliometric data and the calculation of hybrid metrics. By exploiting the information generated through these services DRIVER will be able to provide new advanced information retrieval and personalisation functionality.

The created VRE will also enable the inclusion of other services, that may be identified in the course of the project, to allow the hosting of DRIVER content on the D4Science e-Infrastructure vast storage resources.

1.3.3.3 AquaMaps Scenario

Requirements: production of more precise, accurate, synthetic and predictive specific habitat maps

Users: biodiversity scientists, fisheries and aquaculture resources management scientists and decision makers, ICIS, FishBase, SeaLifeBase, GBIF

Main e-Infrastructures used: D4Science, GENESI-DR, OBIS, NEON and GBIF

Main D4Science exploited capabilities: access to the Grid, access to earth observation data sources and specialised tools, workflow capabilities

The AquaMaps service permits the biodiversity community to establish/predict species geographic distribution based on species ecological envelopes according to a transparent algorithm that can be handled by species experts. The project that developed the service is steered by an informal group of biodiversity scientists led by Rainer Froese (Univ. Kiel, IfM-GEOMAR, Germany) with data encoding and technical development carried out by the FishBase and SeaLifeBase teams in The WorldFish Center Philippine Office.

AquaMaps complete potential is presently hampered by the difficulty of access to necessary existing data sources and by the computationally intensive nature of the calculations:

- **Precision**: Scientists would like the algorithm be able to compute distribution at lower scales along the coasts (at least 10’ side, and 5’ will be tested). The algorithm used for creating the model currently calculates occurrence in a 30’ x 30’ square. For small-scale fisheries management at these latitudes the scale is too big. The realisation of this feature has not yet been attempted since it would heavily increase the amount of data and quantity of computation needed.
- **Accuracy**: A more dynamic peer review process would be needed to increase map usefulness and credibility. However, map re-computations are processor-intensive and currently prohibitive. The current AquaMaps interface has a facility for experts to amend the computed map – increased processing power would make this feature of much greater use.
- **Interoperability**: The accuracy of the predictive models could significantly be improved by comparing the computed maps versus other published maps like the FAO species distribution maps (from expert knowledge), national maps and others. These resources must be brought together. Better maps are also needed for assessment purposes and catch data reallocation by other VRE’s such as ICIS.
- **Different resolution levels**: Decision-makers designing marine protected areas would be very interested in a variety of focussed maps. In particular, they would like to be able to produce biodiversity maps at local, national, regional and global levels including up to 200,000 species. Due to the computational cost AquaMaps implementation is currently limited to single resolution type maps.
- **Prediction capability**: The AquaMaps model builds on various sources of geo-morphological and ecological data (such as bathymetry, ranges of latitude and longitude, salinity, sea surface temperature, primary productivity, etc.) as well as species occurrences. These sources of data are distributed in various systems such as FishBase, SeaLifeBase, AlgaeBase, OBIS/GBIF, and make use of NOAA satellite images (averaged across a fixed number of years). A knowledge ecosystem that brings this data together would mean much greater efficiency in map preparation.

The possibility to combine Earth observation data and other environmental sources available in the knowledge ecosystem, along with new data access and distributed computing technology creates new opportunities of improvement on the above points. New spatial data sources, Grid computing and fast data access protocols provide mechanisms for an enhanced integration of higher quality data sources into AquaMaps. This includes methods for data assimilation for regional and local models and customizing models/forecasting with in-situ data sources and the integration of Earth observation with new ground based monitoring capabilities.

In order to create the conditions for exploiting such improvements, D4Science-II will set up an AquaMaps VRE that will provide access to the resources needed to improve the predictive model as specified above. In particular, raster images of temperature, salinity and primary productivity according to user’s selected periods/seasons will be generated on-demand by exploiting data, services and processing capabilities accessible through the D4Science e-Infrastructure. The architecture of the AquaMaps service will also be consistently modified to exploit the features provided by the corresponding VRE.

The achieved improvements, including the ability to monitor trends, will have a strong impact on the work of numerous scientific communities. For example, it will enable enhanced monitoring of climate change and a better evaluation of its impact on aquatic populations distribution and abundance, and on aquatic ecosystems. The better quality maps will also have benefits on **ICIS** (see below) which exploit them for generating catches re-allocation. Moreover, the improved quality and variety of maps will be beneficial for all the species information services that exploit these maps (e.g. **FishBase** [27], **SeaLifeBase** [64], **GBIF** [29] and a popular **Internet service** currently under development).

1.3.3.4 FCPPS Scenarios

Requirements: support for timely creation of Country profiles, uniform access to the knowledge produced during the Country Profiles production process

Users: policy makers, fisheries managers, fisheries economists, aquaculture specialists, other UN organizations, biodiversity specialists

Main data e-Infrastructures used: D4Science, FAO FI time-series data; FAOLEX, UN-DATA, COMTRADE, OBIS, GBIF

Main D4Science exploited capabilities: access to the Grid, retrieval and access to heterogeneous data, high processing capabilities, storage and management of compound dynamically generated information objects, resource sharing.

The FAO Fisheries and Aquaculture Department (FI) produces country-level reports to enhance fisheries management decision-making and to promote advocacy in the sustainable use and conservation of aquatic resources. These popular, high-priority reports require complex aggregation, processing, analysis and editing of continuously evolving **data from heterogeneous sources**, e.g. inventories of fish stocks and fisheries, time-series data on capture and production, national and regional legal regulations and commodities trade data. Organizational and technological changes are bringing strong pressure on the department and on the FAO to make this data interoperable within the organization, across the UN and to other organizational partners.

The Fishery Country Profiles Production System (FCPPS) is a VRE being built in as part of the D4Science project[7], which aims at providing an environment where scientists can query, access, analyze and report on a broad range of fishery indicators at country-level and create compound objects whose parts can dynamically generated on-demand. The generation of these parts can also require **large processing capabilities**, like in the case of the generation of environmental maps.

By the end of the D4Science project, the FCPPS VRE will be in production for a selected subset of countries. The operational VREs will form the basis for a thorough revision of the Department's country profiles business model. The early prototyping of this scenario has highlighted the importance of that the **knowledge produced during the country profiles production workflow may have value for other scientists**. Much knowledge is produced in the intermediate steps but only the final aggregated product is published as a report. FCPPS seeks to give partner data providers/consumers on-demand dynamic access to the underlying resources that make up the reports such as time-series statistics and fisheries and resources inventories. There is currently no mechanism to service these requests, which must be satisfied with ad-hoc and through a variety of disunited interfaces, both at user and API level.

D4Science-II will address these requirements by making available in the FCCPS VRE services for cataloguing, storing in the production e-Infrastructure, retrieving and making available both the underlying and the produced knowledge items through the single, standardized access interface built by the project. The programmatic usage of this VRE and the inherent resources sharing supported by D4Science, will enable the re-use of this information which is currently trapped in various legacy systems.

The wider availability of knowledge source services paired with enabled sharing of the large amount of statistical, legal, management and economic knowledge will enable FAO data consumers to better benefit from FCPPS data. This is an important issue for FAO, in particular following both the recent independent external evaluation recommending the FAO position itself as a **"Knowledge Exchange" organization** and the creation of **UN-DATA** [73], a new internet-based data service for the global user community provided by the United Nations Statistics Division (UNSD) of the Department of Economic and Social Affairs

[7] The first public release of the FCPPS VRE is planned for November 2008.

(DESA). Other interested groups come from the fishery, biology, commodity and taxonomic information domains, some of which have already embraced this approach such as OBIS and GBIF.

1.3.3.5 ICIS Scenarios

Requirements: Improving the quality of catches re-allocation

Users: fisheries managers, fisheries and fish stock assessment scientists, ecologists

Main e-Infrastructures used: D4Science, GENESI-DR, AquaMaps, OBIS, OBIS, GBIF, GOOS

Main D4Science exploited capabilities: interoperability with different data sources, data harmonization, large processing capabilities

There is an increasing demand for global catch data with finer geographical resolution which take a greater number of environmental variables into account. Notable examples are:

- The UN recommendation to FAO to "revise its global fisheries statistics database to provide information for the stocks to which the Agreement applies, as well as to high seas discrete stocks on the basis of where the catch was taken." (Document A/CONF/.210/2006/15, Annex, Paragraph 25, 26);
- Frequent requests to separate catch data between Exclusive Economic Zones (EEZs) and high seas;
- The Coordinated Working Parties (CWP) recommendation to respond to the UN recommendation by integrating catch statistics that Regional Fishery Bodies (RFBs) and FAO have collected, maintained, and disseminated, based on rules established in the previous CWPs (CWP-22 Report).

In order to begin addressing the demand for **greater interoperability between regional statistical data sets**, regional and global capture and distribution information of aquatic species, from a number of Regional Fisheries Management Organisations (RFMOs) and international organisations (FAO, WorldFish Center) have been integrated as resources in the D4Science e-Infrastructure. Within the scope of this common framework a specific VRE responding to the needs of fisheries scientists making regional and global aquatic resource assessments has been created. This VRE supports these scientists in performing **spatial re-allocation of catches** between EEZ and high seas[8]. In this environment fishery scientists can use data coming from known (**FAO species distribution maps**) and computed (**AquaMaps**) species distributions to enable the configuration of algorithms and filters that re-allocate harmonized coarse-grained catch data coming from multiple data providers. Integrated into the this framework are query, comparison and output services that allow users to define custom data outputs and output services such as tables, charts and maps. It is expected that

The D4Science project will deliver a prototype VRE where FAO global and regional statistical databases, as well as regional data bases of two independent Regional Fishery Bodies – ICES and NAFO, will be accessible uniformly through a shared data model. The VRE will use custom filtering algorithms to dynamically re-allocate data between EEZ and high seas using known and computed species distribution via a shared GIS data model using FAO species distribution maps and World Fish Center Aquamaps.

During the ICIS requirements gathering meeting (Rome, June 2008) the fishery scientific user community highlighted new opportunities for improving the quality of catches re-allocation not foreseen by the planned D4Science project activity. These opportunities will be developed in the D4Science-II project in particular by:

- **Integrating fine-grained Fishing Vessel Monitoring System data** and related data such as fishing operations' logbook data, cruise survey data and data collected through port state measures. These type of data are useful not only for the fisheries community but also for other scientific communities (e.g. biodiversity community, see AquaMap scenario above). A simplified access to these types of data will enable tighter regional monitoring, improved scientific assessment and better fisheries management.
- **Exploiting related repositories of ecosystem information** such as OBIS, GBIF and GOOS and climate and sea surface products coming from GENESI-DR
- **Being able to both provide and consume data programmatically**. Partners want their current systems to be interoperable with ICIS so that they can place their resources within D4Science and work

[8] The first release of this VRE will be made publicly available in November 2008.

within the VRE, but also have easy programmatic access to their datasets to be used within their legacy systems as required.

1.3.4 Supporting the Ecosystem: Outreach and Sustainability

The success of the D4Science-II will depend not only on the quality of the technical solutions that it will develop but also on its capability to stimulate interest around the notion of knowledge ecosystem. To this aim the project will perform several networking activities dedicated to attract both scientific communities and organizations managing data e-Infrastructures. In performing these activities it will largely exploit the huge number of contacts brought by the project partners at European and International level. In particular, as preferred policy, the project will work on the **co-organization of dissemination and training activities** with relevant third parties in order to maximize the impact of these events. Other EU project meetings, scientific conferences, and especially communities' meeting and data-Infrastructure events, will be monitored in order to establish appropriate liaisons.

The promotion activities performed will necessarily take into account the characteristics of the data e-Infrastructures organizations addressed. The majority of them play in a market characterised by clear objectives and by precise business plans. Capturing and keeping the attention of these players alive for the duration of the project and beyond require the adoption of carefully selected outreach strategies. In particular, the project intends to strongly rely on **training and live demonstrations** of results by exploiting its production e-Infrastructure and the operational VREs. A key issue to be addressed to foster the wide adoption of D4Science by the scientific communities and the growth of the knowledge ecosystem is sustainability. To address it the project plans to liaise with other on-going initiatives that are already facing this issue in more traditional e-infrastructure contexts. Notable examples of these initiatives are the **European Grid Initiative** (EGI) (see Appendix B.4) and the **DRIVER Confederation** (see Appendix B.2). The current D4Science project already collaborates with both initiatives by contributing with its experience in data e-Infrastructures. D4Science-II will introduce in the current panorama the new notion of knowledge ecosystem. In the knowledge ecosystem vision the success of a data e-Infrastructure is influenced by the success of the others since each of them affects, and is affected by, the modification and updates on the delivery of the other e-Infrastructures' offerings. This may open the way to the adoption of alternative sustainability solutions. For example, the fact that each ecosystem e-Infrastructure has its own community that uses and enhances the ecosystem capabilities might imply that all these communities could federate to provide the support for the ecosystem operation.

D4Science-II plans to strengthen its collaboration with EGI and the DRIVER confederation by also providing them with the vision of the many other collaborating large data e-Infrastructures involved in the construction of the knowledge ecosystem. These will be invited to take part in the discussion under the umbrella of the two initiatives and, possibly, also through specific events organized by the project.

1.3.5 D4Science-II Activities Overall Strategy

The objectives of the project will be achieved through the cooperation of Networking, Service and Joint Research Activities.

The overall objective of the D4Science-II Networking Activities is to *(i) promote the building* of the envisaged knowledge ecosystem by establishing cooperation channels with the key players and *(ii) foster the exploitation* of the resulting ecosystem by scientific communities involved in global research challenges. This objective will be reached by (*i*) involving key e-Infrastructures and large research communities in the processes governing the development and usage of the knowledge ecosystem, (*ii*) disseminating project outcomes to a large and heterogeneous audience, (*iii*) training the various players and decision makers, and (*iv*) delivering a true and proper knowledge ecosystem built by relying on the knowledge gathered during D4Science, the technology developed and the advices and feedback received by providing the community with it during the project lifetime.

The D4Science-II Service Activities aim at providing the key players and large research communities internal to the project or attracted through the Networking Activities with the services implementing the envisaged Ecosystem and Virtual Research Environments. This will be done by putting in place activities, procedures, mechanisms and resources – either human or inanimate (e.g. hosting nodes, data collections, services) – needed to guarantee a 24/7 service availability. In particular, *(i)* the procedures and mechanisms designed in D4Science to operate the resulting e-Infrastructure will be revised to accommodate the needs of the new scenarios, (*ii*) the activities needed to populate the ecosystem with resources will be enhanced to deal with the large pool of resources resulting from federating the target e-Infrastructures, and *(iii)* the mechanisms leading to the definition and operation of VREs will be reinforced to deal with the strong demand for them arising in the target scenarios.

Finally, the D4Science-II Joint Research Activities will implementation the technology supporting the vision predicated by the D4Science-II project. This will be done by having the D4Science technology – gCube – as a firm base and expanding it so as to meet the novel expectations emerging from the D4Science-II application scenario. In particular, a considerable effort will be invested in implementing the technology needed to bridge the target e-Infrastructures and services with D4Science; at the same time the approaches to render D4Science facilities exploitable by service providers in programmatic way will be empowered. Thus, the innovative services envisaged in the target scenario become possible and will be implemented to demonstrate the effectiveness of the knowledge ecosystem in challenging research scenarios.

Moreover, since D4Science-II continues the work of the D4Science project by further progressing toward the deployment of *a pan-European e-Infrastructure implementing the knowledge ecosystem* exemplified above (cf. Section 1.1), it is expected to build upon the D4Science achievements and experiences gained. Thus all the project activities have been designed to properly manage a smooth transition of services between D4Science and D4Science-II. In particular, a six-month overlapping between the two projects is envisaged thus the D4Science-II activities explicitly continuing D4Science relative activates will start at month 7.

The three activities just described are organised into cooperating work packages whose interaction is depicted in Figure 2.

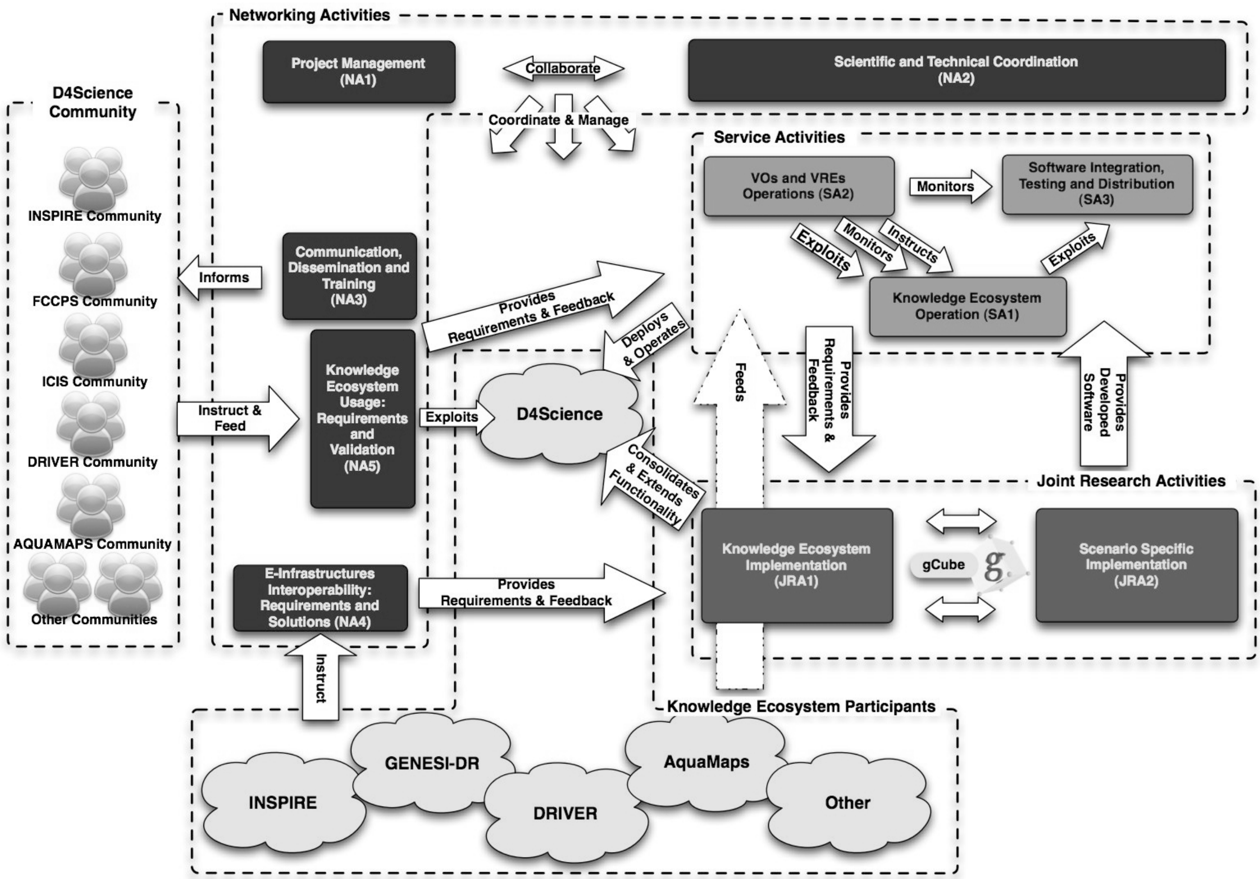


Figure 2. D4Science-II Activities Overall Strategy

The three activities will operate in synergy toward the achievement of the D4Science-II objectives. In particular, the following work packages hold a critical role in linking the activities assigned to each area of the project:

- *NA2 Scientific and Technological Coordination*; this work package organises the overall scientific and technical activities, monitor the achievements and recommend appropriate actions enhancing the activities effectiveness toward the project objectives;
- *NA4 Knowledge Ecosystem Building: Requirements and Solutions*; this work package promotes the development of common standards and solutions allowing for sharing resources across the boundaries of the constituent e-Infrastructures;
- *JRA1 Knowledge Ecosystem Implementation*; this work package implements the technology realising the approaches and standards identified by NA4 and reinforces gCube with new features transforming it into a system supporting the Ecosystem approach;
- *NA5 Knowledge Ecosystem Exploitation*; this work package promotes the consumption of the facilities offered through the implemented Ecosystem and their testing in the concrete and challenging research area;
- *JRA2 Community Specific Resources Implementation*; this work package implements innovative services benefiting from their operating in the knowledge ecosystem and providing the target scenarios with new features that cannot be currently implemented in the Infrastructure and Services their communities were used to;
- *SA1 Knowledge Ecosystem Operation*; this work package organises the deployment and operation of the proposed knowledge ecosystem, i.e. the reinforced version of the D4Science Infrastructure as well as of the services bridging the participating Infrastructures;
- *SA2 VOs and VREs Definition and Operation*; this work package manages and supports the activities needed to provide the target scenarios with the Virtual Research Environments supporting them.

The complete list of D4Science-II work packages, their descriptions, their relations, their constituent tasks and the related timing is detailed in the rest of this proposal (cf. Sections 1.4, 1.5, and 1.6).

1.4 Networking Activities and associated work plan

1.4.1 Overall strategy

The objective of the Networking Activities (NAs) is to coordinate the interoperability needs of the repositories and infrastructures participating in the project, and by extension the scientific communities (e.g., biodiversity, high-energy physics, and fisheries and aquaculture resources management) that use them. The NAs will place an emphasis on awareness raising, training, adoption of standards and mitigation/validation of community requirements. The Networking Activities also encompass the administrative and technological coordination of the project at a global level.

To achieve this objective a set of activities are planned and organised into a number of interacting work packages:

- *NA1 – Project Management* will be dedicated to the overall project administration and finance, with a particular focus on quality assurance. Sustainability will also be considered in order to create a community framework that ensures the long term viability of the proposed approach.
- *NA2 – Scientific and Technical Coordination* will cover the coordination activities to be carried out by the scientific and technical managers, including the operation of the Project Management Board (PMB) (cf. Section 2.1).
- *NA3 – Communication, Dissemination and Training* will disseminate project results and knowledge in order to raise awareness within the largest pool of user communities, including the training of software engineers, administrative personnel and end users, from within the project's user communities as well as from other scientific communities.
- *NA4 – E-Infrastructures Interoperability: Requirements and Solutions* will develop common approaches and interoperability solutions for promoting an open exchange of resources within and among the ecosystem constituents, namely D4Science and the other main application-level infrastructures the ecosystem federates, GENESI-DR, DRIVER, INSPIRE, and AcquaMaps. Standards will cover several levels, from low-level communication protocols and encoding schemes, to data models to classification systems. While performing such an activity the project will set up a forum for the discussion, agreement and promotion of such standards.
- *NA5 – Knowledge Ecosystem Usage: Requirements and Validation* will involve the D4Science-II target communities and their exploitation of the developed knowledge ecosystem. An iterative process will be put in place, in which requirements feed VREs configuration. VRE's are tested by the communities, and the results feed back into refinements and improved configurations.

Figure 3 shows the interactions among the Networking Activities work packages and the relationships of NA with the other project activities.

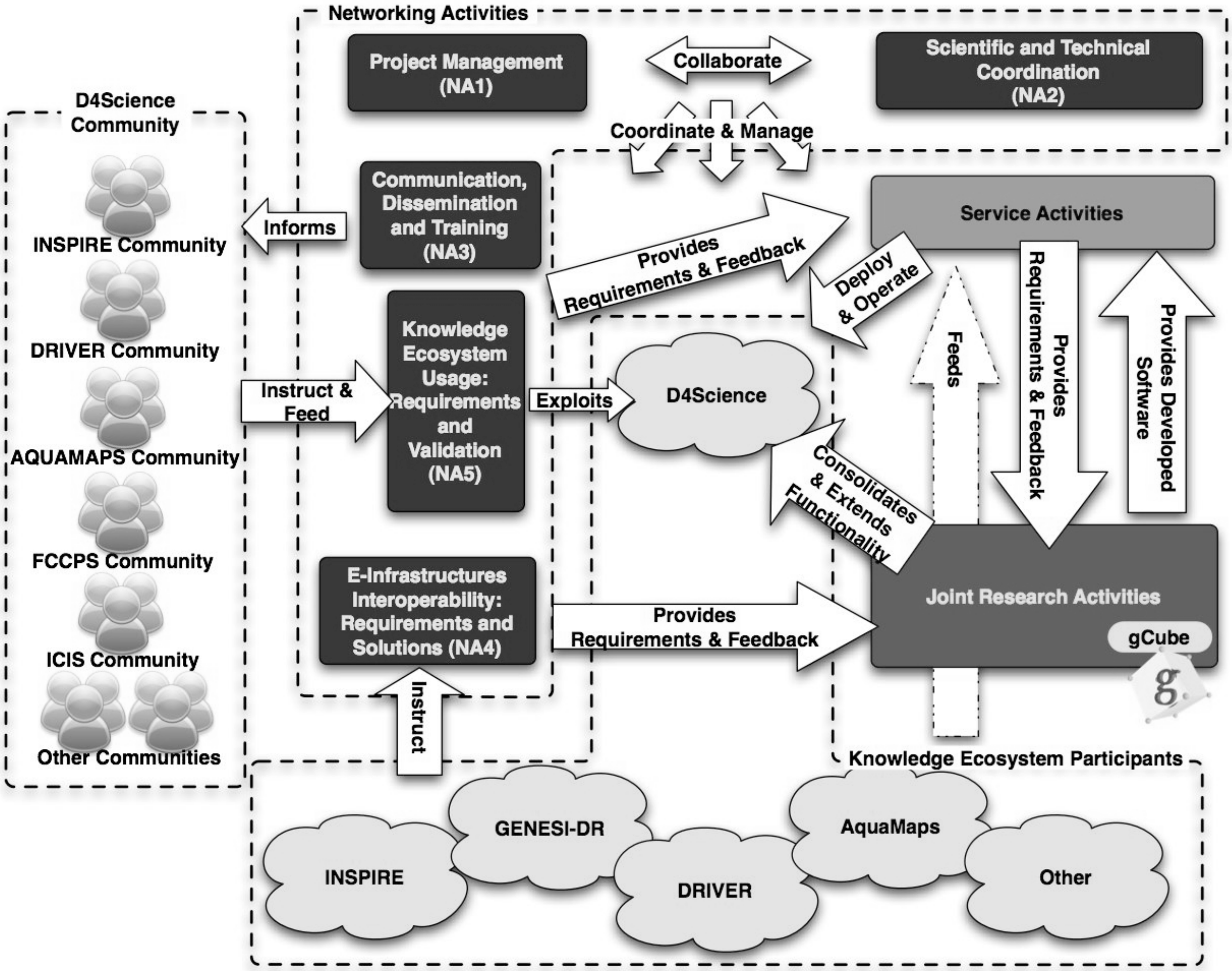
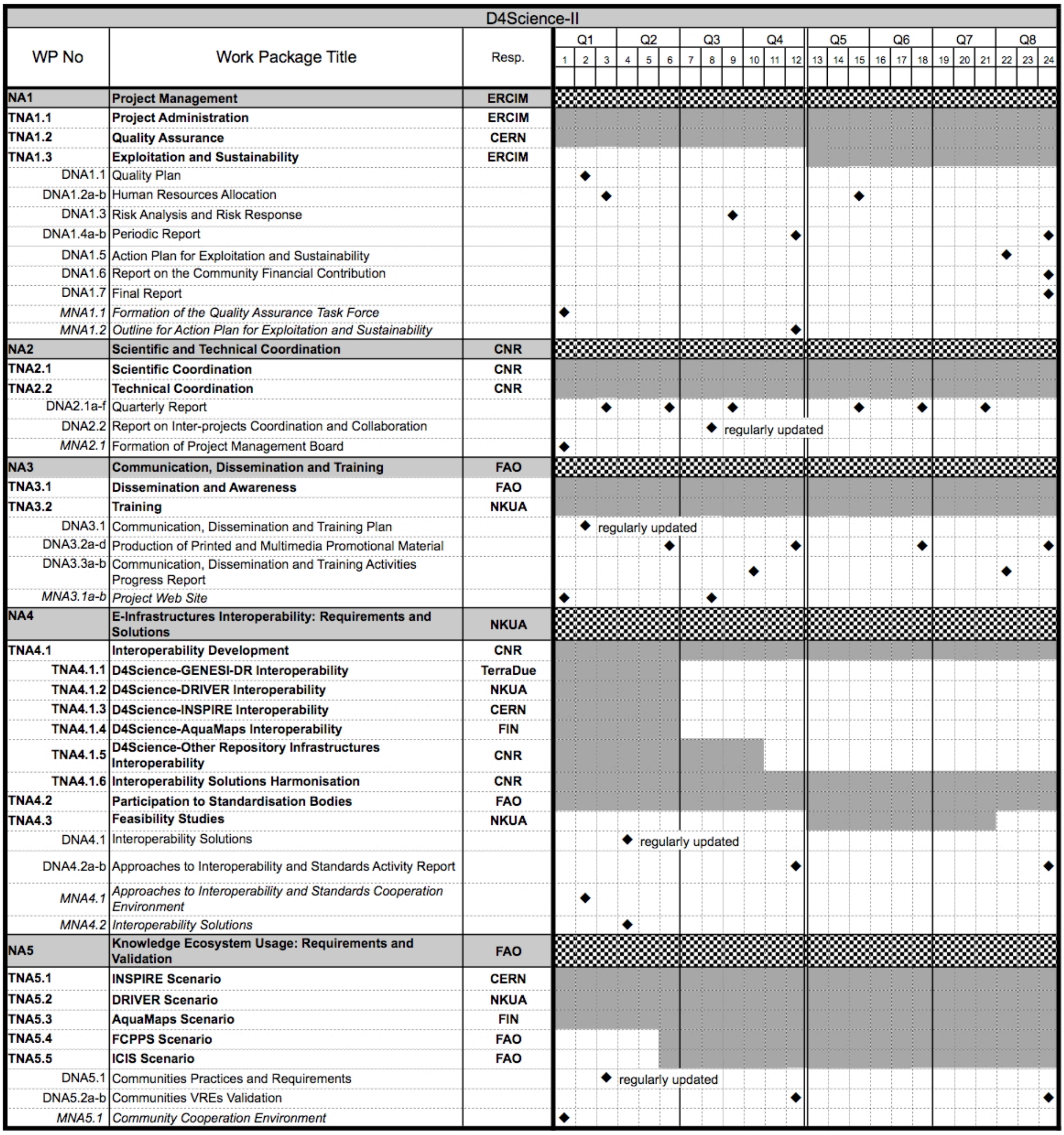


Figure 3. Networking Activities Overall Strategy

In this figure, community requirements, pulled in through NA5 *Knowledge Ecosystem Usage: Requirements and Validation,* inform the Service Activities which serve to configure the D4Science-II knowledge ecosystem and the VREs. If such requirements can be satisfied through existing technology, Service Activity will address them; otherwise Service Activity will feed the Joint Research Activities with requests for technology enhancement (cf. Section 1.5.1). Meanwhile, e-Infrastructure/repository involvement in NA4 *E-Infrastructures Interoperability: Requirements and Solutions* nourish the Joint Research Activities with requests about the technology needed to build knowledge ecosystem. NA3 *Communication, Dissemination and Training* will actively train user communities and disseminate project results to a wide audience. Finally, NA1 *Project Management* provides the support and tools to manage the process at the global project level while NA2 *Scientific and Technical Coordination* provides the necessary scientific and technical overview and guidance to keep the project true to the overall vision.

1.4.2 GANTT Diagram



1.4.3 Detailed work description

1.4.3.1 Work package list

Work package No	Work package title	Type of activity	Lead participant No	Lead participant short name	Person-months	Start month	End month
NA1	Project Management	MGT	1	ERCIM	24	1	24
NA2	Scientific and Technical Coordination	COORD	2	CNR	14	1	24
NA3	Communication, Dissemination and Training	COORD	7	FAO	33	1	24
NA4	E-Infrastructures Interoperability: Requirements and Solutions	COORD	3	NKUA	78	1	24
NA5	Knowledge Ecosystem Usage: Requirements and Validation	COORD	7	FAO	46	1	24
	TOTAL				195		

NA1 - Project Management

The first of the Networking Activities' work packages will ensure the overall efficient operation of the project, while guaranteeing that the level of quality meets the expectations raised by the project. It includes the work of the project Coordinator, or Administrative and Financial Director (AFD), and the Assurance Task Force. Scientific and technical coordination activities are presented in a separate dedicated work package (NA2).

GEIE ERCIM will serve the role of **Coordinator**, acting as the formal point of contact with the European Commission. The Coordinator is responsible for consortium, financial, contractual, ethical and innovation management. Project administration includes the careful monitoring of resource expenditures and the fulfilment of EU reporting and contractual obligations. The conflict resolution procedure defined in Section 2.1 will be handled by the Coordinator. These activities are grouped under Task NA1.1 which will produce DNA1.2a-b, DNA1.4a-b, DNA1.6 and DNA 1.7, dedicated to the reporting obligations as emphasized in the Grant Agreement with the European Commission.

Quality assurance is an essential element of NA1. The objective of this task is to ensure the monitoring and assessment of progress and results across the project. A task leader will coordinate Quality Assurance Task Force (QATF) to ensure the delivery and adherence to a Quality Plan that will define the project's main quality processes, such as deliverable preparation and internal review; periodic review preparation and post-review follow-up; activity-specific process; basic rules for D4Science-II publications; practicalities concerning the functioning of the management boards; etc.; as well as the tools and metrics that will be applied in many of these processes. The task force will obtain input from the work package leaders in order to measure the project's progresses. Thus, the task dedicated to Quality Assurance (TNA1.2) will perform risk identification and analysis impacting the global Integrated Infrastructure Initiative (I3) project, and measure the project's progresses across activity areas (i.e., networking, service, joint research). As indicated, the QATF will produce a Quality Plan (DNA1.1) for use in the project, and perform a Risk analysis and risk response (DNA1.3) for monitoring purposes.

Exploitation and sustainability is the third task (TNA1.3) encompassed by the Project Management activities. Sustainability is concerned with ensuring on a long-term basis that the artefacts produced by D4Science-II knowledge ecosystem will not only continue to exist beyond the project duration, but will have been adopted by large and influential user communities and that many others will have learned about the ecosystem's capabilities. In order to achieve these objectives, close collaboration will be required with NA3 (Communication, Dissemination and Training) and NA4's standardisation activities. Within the first project year, a workshop addressing sustainability issues will be organized in conjunction with a larger joint e-Infrastructures event. Proceedings from the workshop, as well as recommendations originating from the D4Science-II high-level External Advisory Board and input solicited from the standardisation bodies will be synthesized into a detailed outline for an Action Plan to be established as a milestone (MNA1.2) in month 12. The specialties of the scientific domains involved, the envisaged open access to information and the nature of implementation planned, all raise special challenges to planning for exploitation, necessarily multi-faceted, looking into different options of the knowledge ecosystem "marketing", be they services, software or information. The Action Plan for Exploitation and Sustainability (DNA1.5), to be delivered in M22, will work to evaluate the strengths and weaknesses, costs and benefits, of the D4Science-II product, and making comparisons to those of the adjacent solutions' where relevant.

NA2 - Scientific and Technical Coordination

This work package will cover the coordination activities to be carried out by the Project and Technical Directors, Project Management Board (PMB) and Technical Board (TB) (see Section 2.1). High quality scientific and technical management will be one of the key success factors of the D4Science-II Integrated Infrastructure Initiative (I3). The activity performed in this work package will build on experience gained by some of the proponents in the framework of the DILIGENT and D4Science projects. In particular, the split between the scientific and technical management and the introduction of a **Project Management Board (PMB)** composed of directors and managers that act as channels between the different project operational areas will be very effective for on-time monitoring and governing of the project towards meeting its objectives. The formation of the PMB with a schedule of anticipated meetings for the two year duration (MNA2.1) is an important milestone to be achieved in the first month of the project.

This structure has been designed to meet our major focus, i.e., to ensure that the management of D4Science-II provides the continuous environment to ensure the implementation the knowledge ecosystem and its effective exploitation to serve scientific communities operating in both the D4Science and in other e-Infrastructures domains. The achievement of this objective will be essential for the effective exploitation of the results by the communities addressed by the projects and by many other communities of scientists.

The **Project Director** (PD) will supervise progress achieved as defined by the Description of Work, create and maintain the necessary conditions for successful and effective collaboration among partners, distribute information pertinent to the consortium, and arrange any necessary ad hoc meetings. Jointly with the Project Management Board, the PD will make strategic decisions and recommend appropriate actions to correct delays and unexpected situations. Additionally, the PD will initiate and maintain interactions with other 6th and 7th Framework projects and other R&D national/international programmes.

The **Technical Director** (TD) will have the responsibility of monitoring the day-by-day progress of the project's technical activities (i.e., technical aspects across the networking, service and joint research activities) and ensuring their proper interaction. The TD will chair the Technical Board (see Section 2.1) in charge of making recommendations for all technical activities across work packages. Additionally, the TD will be responsible for detailed technical effort re-planning in the face of project deviations and for performing technical risks analysis and contingency planning.

The Project Director will be Dr. Donatella Castelli (Scientific Coordinator of the DILIGENT and D4Science projects) and the Technical Director will be Dr. Pasquale Pagano (Technical Director of the DILIGENT and D4Science projects), both of CNR (their research interests and backgrounds are provided in Section 2.2.2). The Project Director will be supported by the Project Office, and in particular, a project assistant hired by CNR.

This work package will produce quarterly reports (DNA2.1a-f) to encourage information sharing on the progress of the work achieved towards the scientific and technical objectives of the project. A report on the interaction with other projects and programmes (DNA2.2a-b) will be produced twice during the life of the project.

NA3 - Communication, Dissemination and Training

The Communication, Dissemination and Training activity is intended to evolve with the life of the project. The objectives of this work package will be to disseminate project results and knowledge in order to raise awareness with the largest pool of potential users and partners, and to train software engineers, administrative personnel, technicians and end users both internal to the project and from scientific communities or other e-Infrastructures.

The work package will collaborate with the other NA work packages, NA1 for sustainability, NA2 for inter-projects collaboration, and in particular, NA4 and NA5 to stimulate interest around the notion of knowledge ecosystem, attract scientific communities and to promote cross-community collaboration. Historically separate disciplines are increasingly discovering shared interests and shared data needs. The outreach aspects of NA3 will be combined with the standards activities of NA4 to encourage cross-discipline communication.

The first activity will be to develop a Communication, Dissemination and Training Plan (DNA3.1) that will define the project's key messages and serve as a blue-print for reaching the targeted audiences. The brand identity defined for D4Science will be maintained. The Communication, Dissemination and Training Plan will encompass the three main steps of dissemination (i.e., Awareness, Understanding and Action) and will equally focusing on training. Communities involved in the project must have sufficient knowledge of the project's infrastructure in order to efficiently exploit it. Champions and decision makers in all areas will be targeted for a deeper understanding of the project activities and for training on the proposed set of tools. Training will address the diverse needs raised in the contexts of **development and exploitation of the infrastructure**. Training to new user communities by bringing them in contact with the instruments placed at their disposal will be a main activity. These scientific communities will not only become acquainted with the use and administration of the infrastructure, but will also learn to extend and enhance it with new resources, be it machinery, archives or services. As an example, INSPIRE is interested in showcasing to other repositories how it could grow its services through the on-demand processing power of the Grid accessed through the interoperability tools developed in this project.

Detailed metrics will be developed to measure dissemination activities and to evaluate the outcome and effectiveness of training events.

Communication, Dissemination and Training activities progress reports (DNA3.3a-b) will be produced at months 10 and 22 to highlight activities accomplished in the period, including achievements and lessons learned and to explain deviations from the Communication, Dissemination and Training Plan.

The necessary framework for dissemination and training activities will be created through a visible and **dynamic public web site** (community platform) and a range of promotional and training materials and instruments. The web site will enable the dissemination, discussion and sharing of results and continuing project progress and will also be a tool to communicate with the scientific communities involved.

A **proactive program** will be set up to optimize synergies between projects and raise awareness of the project's mission and objectives at targeted events. In order to maximise the impact of these events, the project will co-organize dissemination and training activities with relevant third parties. The project will actively participate in joint activities and future meetings related to the e-Infrastructures (e.g., with the European Grid Initiative (EGI) and the DRIVER Confederation) and other related areas including the participation and contribution to relevant working groups (e.g. W3C, OGF and ETSI on standards) and related scientific conferences (e.g., ECDL, JCDL, EGEE).

Other future EU project meetings (e.g., RI Concertation Events), scientific conferences, and especially communities meetings (e.g., with Regional Fishery Bodies, OBIS and AquaMaps) and cross communities meetings (DRIVER and INSPIRE), and data-Infrastructure events, will be monitored or co-organized during the project lifetime in order to establish appropriate liaisons.

By exploiting its production e-Infrastructure and the operational VREs, the project intends to strongly rely on **training and live demonstrations** of results.

Given the importance of a project public web site as a communication and dissemination channel, it is a milestone (MNA3.1a-b) to inform the web community about the D4Science-II project at M1. To reinforce the D4Science brand and stress the continuity of the D4Science and D4Science-II projects, the D4Science web site will gradually integrate and combine information from both projects. At month 8, the common web site will provide full technical capabilities.

The project will provide input for relevant European Commission initiated dissemination activities (e.g., press releases, news bulletins, brochures, success stories, posters, web-based publications, multimedia material, etc). In this context the project's dissemination messages will also reflect its broader societal and economic impact.

All of the communication and dissemination activities will be augmented by the submission of papers to targeted publications and events, including journals, magazines and conferences and the preparation of press releases where appropriate.

The project’s dissemination material will be regularly updated to provide the latest version of its status and achievements. This will be reflected in deliverable DNA3.2a-d.

NA4 - E-Infrastructures Interoperability: Requirements and Solutions

The objective of this work package is to investigate on possible approaches to the various interoperability issues inherent in the construction of a knowledge ecosystem, and then identify and promote proper solutions. Such solutions will have as much as possible to **rely on existing standards** (*de facto* standards also) and **common practices** for a series of reasons, including the well-recognised cross-domain interoperability role assigned to standards, the adequate users’ experience on which any standard is based, , and, obviously, the “don’t reinvent the wheel” principle. The concrete and effective scenario to be considered is characterized by that the main players forming the ecosystem, i.e. D4Science and the target e-Infrastructures (cf. Section 1.3.2), are known *a-priori*; and their representatives - namely domain experts and software engineers driving the design and development of the systems implementing these e-Infrastructures and the policies governing them - actively participate to this activity by contributing with their true experiences and expertises. To make the WP meet its very ambitious objective in this scenario, three main activities are planned: (*i*) the operation of a *forum* for domain experts to discuss on the various interoperability issues and develop common approaches and standard-based solutions for them, (*ii*) the active *participation to standardisation bodies*, and (*iii*) the execution of a series of *feasibility studies* to investigate on the cost-of-participation to the developed ecosystem for new e-Infrastructures and data services.

For what concerns the **forum for developing interoperability solutions**, D4Science-II will set up and operate a task force consisting of domain experts and software engineers that have been directly involved in the design and development of the knowledge ecosystem e-Infrastructures (TNA4.1). The activities of such a task force will be supported by a cooperation environment allowing them to effectively and remotely collaborate without the need to physically meet (mailing lists, a shared workspace and a wiki for presenting the task force outcomes are fundamental tools hosted by such an environment (MNA4.1). The interoperability issues to be resolved range from content-related problems – probably the most expected, primary and wide issue in a knowledge ecosystem – to user, functionality, policy, quality and architecture-oriented issues. The task force will initially identify the core aspects of cross-e-Infrastructure interoperability then perform a throughout survey of existing approaches and standards for the identified core problems. This activity will initially analyse the problem per e-Infrastructure pair, e.g. the interoperability between D4Science and GENESI-DR (TNA4.1.1) is investigated by a different working group than the one dedicated to study the problem between D4Science and INSPIRE (TNA4.1.3). Subsequently the identified solutions and approaches are distilled and harmonised by the joint task force (TNA4.1.6) in which representatives of all the groups work to identify the minimal pool of solutions to be implemented to realise the knowledge ecosystem. In the context of this joint task force the interface for opening the D4Science-II knowledge ecosystem facilities to its consumers for a programmatic consumption will be identified.

For what concerns the **participation to standardisation bodies** (TNA4.2), D4Science-II will actively contribute to this particularly critical interoperability-oriented activity by (*i*) experiencing and commenting on emerging standards and (*ii*) participating in working groups either existing or forthcoming. In particular, the following standardisation bodies will be among the primary target for this activity: W3C, OGF, OGC and ETSI (see Letters of Support - Appendix B). Connections with these standardisation bodies have already established during the D4Science project as well as during per-partner activities. Following these contacts, the D4Science partners were requested to participate in forthcoming standardisation working groups (Open Grid Forum Europe Letter of Support – cf. Appendix B.9).

For what concerns the **feasibility studies** (TNA4.3), they will involve the e-Infrastructure playing a primary role in the D4Science-II knowledge ecosystem together with other e-Infrastructures and data services willing to partake in them. For instance, the Biodiversity Heritage Library is a relevant data source and some of the current scenarios will benefit from its availability in the D4Science-II ecosystem (see Letter of Support - Appendix B.1). This task will analyse requests for interoperability with other e-Infrastructure and data services arising in the context of the target scenarios and determine the viability of implementing these activities according to project technologies and approaches.

The results of this WP's activity will initially manifest in a milestone (MNA4.2) representing the initial pool of interoperability solutions and approaches identified by this task force. Such a milestone will drive part of the development and technology advances to be put in place by the Joint Research Activities (namely, JRA1 cf. Section 1.6.1). In addition to that milestone, the work package will produce two deliverables: one dedicated to document the interoperability solutions and approaches identified by the task force (DNA4.1); the second deliverable dedicated to document the activity of the task force including the participation to standardisation bodies and the feasibility studies (DNA4.2).

NA5 - Knowledge Ecosystem Usage: Requirements and Validation

Exploitation is meant here as usage of the capabilities of the enhanced D4Science to support the emerging needs of scientists. This work package has a two-fold objective to *(a)* manage the requirements for the realisation of the knowledge ecosystem and *(b)* validate the knowledge ecosystem as a whole.

Derived from the community element the **requirements** of each case from the scientific user point of view will be **collected** and **analyzed**. When envisaging a knowledge ecosystem, dynamicity in data and data outputs comes as a natural consequence and demands that the interoperability cases involve multi-directional flows. Thus the capacities for contribution from each case will also be collected and analyzed, driven initially by the already identified opportunities for cross-domain exploitation of resources and knowledge. The delivery of advanced repository tools built by INSPIRE within the DRIVER repositories, the cross-domain flow of data, information and knowledge among earth and life observation communities and the delivery of vast computational and storage resources to the individual cases are some of the examples of such cross-domain exploitation of artefacts that will be made possible via D4Science-II cases. The result of the collection of requirements from the fives addressed scenarios and their analysis will be made available to developers through the deliverable DNA5.1 *Communities Practices and Requirements.* This deliverable will be continuously updated in the first part of the project until requirements are clarified and appropriately refined. The analysis of the requirements of each case will be validated against the technological challenges and options identified by the related work packages and the capacities and scope of the ecosystem backbone. As a result, the entire procedure identifies the aspects that each case needs to handle and drives their prioritization, in the constrained time and resource space of the implementation.

It is against these requirements that the progress of the entire implementation will be **validated**. This is part of a repeating cycle of analysis / design / implementation / validation steps, where the involvement of the community exploiting each case has a guiding role. The exploitability of the VREs and their conformance to requirements will be validated by the interested parties, i.e. the user communities and developers and administrators of the third-party services exploiting them. The outcomes of this evaluation will be reported in DNA5.2 *Communities VREs Validation* which will be realised in two versions (M12, M24).

The work package will work together with NA3 and NA4 to promote cross-community collaboration. Historically separate disciplines are increasingly discovering shared interests and shared data needs. The outreach aspects of NA3 will be combined with the standards activities of NA4 to encourage cross-discipline communication. A web based cooperation environment (MNA5.1) will be set up and maintained for the entire life of the project to enable not only collection of requirements and feedback, but also to enable cross-community collaboration. Such environment will consist in a mix of tools ranging from a shared workspace (e.g. BSCW) to mailing lists, blogs, and wikis making the collaboration between a highly distributed group of people easy and effective.

1.4.3.2 List of Deliverables

Del. no.	Deliverable name	WP no.	Nature	Dissemi-nation level	Delivery date (proj. month)
DNA1.1	Quality Plan	NA1	R	PU	M2
DNA3.1	Communication, Dissemination and Training Plan	NA3	O	PU	M2 (regularly updated)
DNA1.2a-b	Human Resource Allocation	NA1	R	PU	M3, M15
DNA2.1a-f	Quarterly report	NA2	R	CO	M3, M6, M9, M15, M18, M21
DNA5.1	Communities Practices and Requirements	NA5	O	PU	M3 (regularly updated)
DNA4.1	Interoperability Solutions	NA4	O	PU	M4 (regularly updated)
DNA3.2a-d	Production of printed and multimedia promotional material	NA3	O	PU	M6, M12, M18, M24
DNA2.2	Report on inter-projects coordination and collaboration	NA2	O	PU	M8 (regularly updated)
DNA1.3	Risk analysis and risk response	NA1	R	PU	M9
DNA3.3a-b	Communication, Dissemination and Training activities progress report	NA3	R	PU	M10, M22
DNA1.4a-b	Periodic report	NA1	R	CO	M12, M24
DNA4.2a-b	Approaches to Interoperability and Standards Activity Report	NA4	R	PU	M12, M24
DNA5.2a-b	Communities VREs Validation	NA5	R	PU	M12, M24
DNA1.5	Action Plan for Exploitation and Sustainability	NA1	R	PU	M22
DNA1.6	Report on the Community financial contribution	NA1	R	CO	M24
DNA1.7	Final report	NA1	R	PU	M24

1.4.3.3 Work package descriptions

Work package number	NA1	**Start date or starting event:**					M1		
Work package title	Project Management								
Activity Type	MGT								
Participant number	**1**	2	3	4	5	6	7	8	9	10
Participant short name	**ERCIM**	CNR	NKUA	CERN	ENG	BDM-USTRATH	FAO	FIN	4D SOFT	TERRA DUE
Person-months per participant	15	4	1	3			1			

Objectives

This work package is dedicated to the overall project administration and quality assurance. The objectives include the overall efficient operation of the I3 consortium, innovation management, careful monitoring of resource and financial expenditures and fulfilment of contractual obligations and reporting. Quality Assurance objectives include the operation of a Quality Assurance Task Force; definition and monitoring of standards, procedures and metrics; assessment of progress; risk monitoring; and definition and monitoring of a project position on gender equality.

The operation of the Members General Assembly will be chaired by the Administrative and Financial Coordinator and thus falls under the responsibility of this work package.

Description of work

Work package leader: ERCIM;

TNA1.1: Project Administration

Task leader: ERCIM; Participants: CNR;

- Consortium evolution, including management of the consortium agreement;
- Monitoring of resources and financial expenditures, including distribution of the Community contribution;
- Contractual and ethical management, including EU reporting;
- Innovation management, including guidance concerning IPR issues;
- Establishment of intra-project communication and information networks;
- Conflict resolution;
- Monitoring gender equality;
- Operation of the Members General Assembly; including meeting support and follow-up.

TNA1.2: Quality Assurance

Task leader: CERN; Participants: CNR, NKUA;

- Coordination of a task force on Quality Assurance for the definition and distribution of a Quality plan, including tools and metrics;
- Encourages and verifies that standards, procedures and metrics are defined, applied and evaluated;
- Reports on progress achieved within the project;
- Adopts a procedure for identifying, estimating, treating and monitoring risks;
- Define a statement on the “promotion of equality”.

TNA1.3: Exploitation and Sustainability

Task leader: ERCIM; Participants: CNR, CERN;

- Solicits project-specific recommendations from the External Advisory Board on long-term sustainability;
- Co-organizes a restricted workshop to address “exploitation and sustainability” issues with other e-Infrastructure initiatives;

- Produces a detailed outline and Action Plan for Exploitation and Sustainability;
- Obtains feedback from standardisation bodies;
- Investigates the creation of a foundation or association to extend beyond the project duration.

Deliverables

- *DNA1.1 Quality Plan* (M2) defines: (*i*) Quality processes (e.g., deliverable preparation, review preparation and post-review follow-up, activity-specific processes, etc.), tools and metrics, (*ii*) Tracking of progresses;
- *DNA1.2a-b Human Resources Allocation* (M3, M15) informs the other beneficiaries and the Commission of the persons who shall manage and monitor the work of the project, and their contact details;
- *DNA1.3 Risk Analysis and Risk Response* (M9) assesses the probability and impact of risks occurring and to determine the appropriate type of risk response; define a contingency plan if re-planning is not sufficient in face of the actual outcome of work performed and impact of related risks;
- *DNA1.4a-b Periodic Report* (M12, M24) comprises:
  - An overview, including a publishable summary, of the progress of work towards the objectives of the project, including achievements and attainment of any milestones and deliverables identified in Annex I. This report should include the differences between work expected to be carried out in accordance with Annex I and that actually carried out;
  - An explanation of the use of the resources;
  - A financial statement (Form C), from each beneficiary together with a summary financial report consolidating the claimed Community contribution of all the beneficiaries in an aggregate form, based on the information provided in Form C by each beneficiary;
- *DNA1.5 Action Plan for Exploitation and Sustainability* (M22) documents the road map for the continued support and extension of the knowledge ecosystem in the community and plans for post-project collaboration;
- *DNA1.6 Report on the Community Financial Contribution* (M24; due 30 days after receipt of the final payment) informs the Commission on the distribution of the Community financial contribution paid to each beneficiary
- *DNA1.7 Final Report* (M24) comprises:
  - A final publishable summary report covering results, conclusions and socio-economic impact of the project;
  - A report covering the wider societal implications of the project, including gender equality actions, ethical issues, efforts to involve other actors and spread awareness as well as the plan for the use and dissemination of foreground.

Work package number	NA2			Start date or starting event:				M1		
Work package title	Scientific and Technological Coordination									
Activity Type	COORD									
Participant number	1	**2**	3	4	5	6	7	8	9	10
Participant short name	ERCIM	**CNR**	NKUA	CERN	ENG	BDM-USTRATH	FAO	FIN	4D SOFT	TERRA DUE
Person-months per participant		11	1	1			1			

Objectives

The objective of this work package is to coordinate the overall scientific and technical activities of the project. In particular, it will supervise the project across all activities by coordinating interactions, monitoring the time schedule and recommending appropriate actions. Detailed resource allocation and scheduling for technical activities will also be prepared.

This WP includes the operation of the Project Management Board (chaired by the Project Director), a flexible governing unit responsible for decision-making with regard to project strategy, and the Technical Board (chaired by the Technical Director), responsible for day-to-day operational management the D4Science-II e-Infrastructure.

Description of work

Work package leader: CNR;

TNA2.1: Scientific Coordination

Task leader: CNR; Participants: N/A;

- High level supervision of the project from the scientific point of view;
- Strategy development and decision-making;
- Monitoring the time schedule and the timing of the related activities;
- Recommendation of appropriate actions to correct delays;
- Creation and maintenance of the conditions necessary for successful and effective collaboration;
- Assessment of work and achievements of the work packages;
- Interaction with other 6th and 7th Framework projects and other R&D national/international programmes;
- Operation of the Project Management Board.

TNA2.2: Technical Coordination

Task leader: CNR; Participants: NKUA, CERN, FAO;

- Coordination of the technical aspects across all activities (e.g., Networking, Service Activities, Joint Research Activities) by collaborating with the relative Managers (cf. Section 2.1);
- Detailed technical effort re-planning to face project deviations;
- Technical risks analysis and contingency planning;
- Operation of the Technical Board.

Deliverables

- *DNA2.1a-f Quarterly report* (M3, M6, M9, M15, M18, M21) informs the Commission periodically of the progress of work towards the scientific/technical objectives of the project, including achievements and attainment of any milestones and deliverables identified in this proposal. These reports should include the differences between work expected to be carried out in accordance with the proposed plan and that actually carried out;
- *DNA2.2 Report on inter-projects coordination and collaboration* (M8-regularly updated) reports on interactions with FP6 and FP7 projects and other R&D national/international programmes on inter-

project coordination and collaboration.

Work package number	NA3			**Start date or starting event:**			M1			
Work package title	Communication, Dissemination and Training									
Activity Type	COORD									
Participant number	1	2	3	4	5	6	**7**	8	9	10
Participant short name	ERCIM	CNR	NKUA	CERN	ENG	BDM-USTRATH	**FAO**	FIN	4D SOFT	TERRA DUE
Person-months per participant	2	9	8	3		3	7			

Objectives

The Communication, Dissemination and Training activity is intended to evolve with the life of the project. Part of this activity will go beyond contractual obligations.

The objectives of this work package are to disseminate project results and knowledge in order to raise awareness within the largest pool of user communities and to design, plan, support, conduct and evaluate the training of software engineers, administrative personnel and end users, from within the project user communities and from other scientific communities. The first activity will be to develop a communication, dissemination and training strategy that will define the project's key messages, image and style and serve as blueprint to assure reaching targeted audiences. The second activity will be the creation of the necessary infrastructure for dissemination and training activities through the D4Science-II visible and dynamic public web site and a range of promotional materials. The third activity will be the preparation of a proactive program to raise awareness of the project's mission and objectives at targeted events as well as to disseminate and share results and continuous project progress. A specific action in awareness raising will be to transmit the objectives and activities of D4Science-II to other projects as the basis to develop and encourage synergies.

Description of work

Work package leader: FAO;

TNA3.1: Dissemination and Awareness

Task leader: FAO; Participants: ERCIM, CNR, NKUA, CERN;

- Development of a Communication, Dissemination and Training Plan;
- Definition and development of a D4Science-II brand very similar to D4Science (e.g., mission statement, slight update of the logo, etc.);
- Design, content and upkeep of the single D4Science/ D4Science-II website;
- Design and production of publicity material (e.g., posters, flyers, brochures);
- Design and development of sharing mechanisms to foster knowledge exchange among members of the D4Science-II new user communities;
- Organisation of workshops/panels in targeted conferences;
- Organisation of community meetings;
- Planning of joint events with relevant projects (concertation);
- Preparation of press releases;
- Submission of papers to targeted publications and events, including journals, magazines and conferences;
- Provide input for relevant European Commission initiated dissemination activities (e.g. press releases, news bulletins, brochures, success stories, posters, web-based publications, multimedia material).

TNA3.2: Training

Task leader: NKUA; Participants: CNR, CERN, BDM-USTRATH, FAO, TERRADUE;

- Advising the targeted audience in the adoption and exploitation of the services offered by the project and its products;
- Design and planning of training adapted to the audience needs;

- Production of targeted training material;
- Conduction of training;
- Preparation of forms for the evaluation of course and trainer and trainee evaluation questionnaires

Regarding external Training, in collaboration with the technical management of the project and upon requests or proposals from partners, special training events will be organised for the targeted groups (end-users, infrastructure administrators, development teams etc), exploiting instruments such as presentations, workshops, and round table discussions on emerging technologies and standards, as well as internal intermediate products of the project's work plan. As an example, INSPIRE will raise awareness and eventually train managers of repositories outside the project in the potential to grow their services by using the on-demand processing power of the Grid accessed through interoperability tools developed in this project.

Regarding internal training, special training events will be organised for diffusing into the gCube team the required expertise for realising the objectives of the project. Evolving technologies, scientific advances and low/mid level development topics will be targeted by such activities, utilising a large part of the instruments that target external trainee groups.

Deliverables

- *DNA3.1 Communication, Dissemination and Training plan* (M2-regularly updated) serves as the guidelines for dissemination, liaison and training activities; to define key messages, target audiences (scientific communities) and methods of communication and training. The Communication, Dissemination and Training plan will also contain metrics to measure dissemination and training success;
- *DNA3.2a-d Production of Printed and Multimedia Promotional Material* (M6, M12, M18, M24) supports the project dissemination and promotion activity;
- *DNA3.3a-b Communication, Dissemination and Training Activities Progress Reports* (M10, M22) docuemnts activities accomplished in the period and explains deviations from the dissemination plan, including figures about dissemination success (e.g., use of the D4Science web portal, Scientists linked to the D4Science e-Infrastructure, Institutes participating in the D4Science e-Infrastructure). It will define the types of training events to be performed during the period and report on achievements and lessons learned.

Work package number	NA4			**Start date or starting event:**				M1	
Work package title	E-Infrastructures Interoperability: Requirements and Solutions								
Activity Type	COORD								

Participant number	1	2	**3**	4	5	6	7	8	9	10
Participant short name	ERCIM	CNR	**NKUA**	CERN	ENG	BDM-USTRATH	FAO	FIN	4D SOFT	TERRADUE
Person-months per participant	2	12	16	7	8	13	7	5		8

Objectives

This work package is dedicated to design the best approaches for the realising interoperability between D4Science and other data e-Infrastructures and to identify the standards that may facilitate this process.

Description of work

Work package leader: NKUA

TNA4.1 Interoperability Development

Task leader: CNR; Participants: NKUA, CERN, ENG, BDM-USTRATH, FAO, FIN, TERRADUE;

- Identify and recommend best practices for knowledge modelling and representation;
- Develop methodologies leading to interoperability solutions;
- Organise the results of Tasks TNA4.3, TNA4.4, TNA4.5 and TNA4.6 in a coherent whole;

These activities are organised in two families of sub-tasks: the first dealing with interoperability between D4Science and each target infrastructure (TNA4.1.1-TNA4.1.5), the second dedicated to harmonise the identified solutions and promote their exploitation.

TNA4.1.1 D4Science-GENESI-DR Interoperability

Task leader: TERRADUE; Participants: ENG, BDM-USTRATH;

TNA4.1.2 D4Science-DRIVER Interoperability

Task leader: NKUA; Participants: ENG, BDM-USTRATH;

TNA4.1.3 D4Science-INSPIRE Interoperability

Task leader: CERN; Participants: ENG, BDM-USTRATH;

TNA4.1.4 D4Science-AquaMaps Interoperability

Task leader: FIN; Participants: ENG, BDM-USTRATH;

TNA4.1.5 D4Science-Other Repository Infrastructures Interoperability

Task leader: NKUA; Participants: CNR, ENG, BDM-USTRATH, FAO, TERRADUE;

TNA4.1.6 Interoperability Solutions Harmonisation

Task leader: CNR; Participants: NKUA, CERN, ENG, BDM-USTRATH, FAO, FIN, TERRADUE;

TNA4.2 Participation to Standardisation Bodies

Task leader: FAO; Participants: ERCIM, CNR, NKUA;

- Identify standardization bodies and participate in standardisation activities organised by them;
- Prepare contributions to the standardisation process (open standards as much as possible).

TNA4.3 Feasibility Studies

Task leader: NKUA; Participants: ENG, BDM-USTRATH, FAO, TERRADUE;

- Thoroughly study the applicable solutions identified as part of the Interoperability studies;
- Report on the need and relevance of the solutions identified.
- Report on their feasibility considering project time and resource constraints.

Advise project boards for the rejection of solutions upon the identification of in-feasible or risky cases, or for selection among closely competing candidate ones.

Deliverables

- *DNA4.1 Interoperability solutions* (M4 regularly updated) describes the interoperability solutions and approaches for developing a knowledge ecosystems of interoperable e-Infrastructures and data services;
- *DNA4.2a-b Approaches to Interoperability and Standards Activity Report* (M12, M24) documents the activities conducted in the context of this work package including the participation to standardisation bodies and the feasibility study results.

Work package number	NA5			**Start date or starting event:**			M1		
Work package title	Knowledge Ecosystem Usage: Requirements and Validation								
Activity Type	COORD								
Participant number	1	2	3	4	5	6	**7**	8	9	10
Participant short name	ERCIM	CNR	NKUA	CERN	ENG	BDM-USTRATH	**FAO**	FIN	4D SOFT	TERRA DUE
Person-months per participant		12	10	7			11	6		

Objectives

This work package manages the actions for the realisation of a long-term knowledge ecosystem based upon the implementation of the selected interoperability cases. This is a two-fold objective to:

- manage the requirements for the realisation of the knowledge ecosystem,
- validate the knowledge ecosystem as a whole.

Tools for the creation, collection, analysis, validation and tracking of requirements will be deployed at the start up of the project. Opportunities for cross-domain exploitation of resources and knowledge will be continuously monitored. Cross-community collaboration will be fostered by inter-community review of requirements and a deliverable analysing the opportunities and suggestions for their implementation.

In an iterative process, deployed infrastructures and the resultant VREs will be validated by the interested communities and the results will then be used to refine requirements for subsequent extensions/re-configurations of the infrastructures.

Description of work

Work package leader: FAO;

For Tasks TNA5.1 to TNA5.5 the same action plan applies although the focus for each corresponds to a different community / scenario (cf. Section 1.3.3). The actions planned are:

- Definition and tuning of scenario requirements and use cases;
- Exploitation and usage of the implemented VREs in a production environment;
- Validation (i.e. evaluation and recommendations) of delivered Virtual Research Environments;
- Exploring the potential of exporting the resources developed to support the scenario (e.g. specific tools) or as products of the scenarios activities (e.g. knowledge products) to other application domains within the knowledge ecosystem.

TNA5.1 INSPIRE Scenario

Task leader: CERN; Participants: CNR;

TNA5.2 DRIVER Scenario

Task leader: NKUA; Participants: N/A;

TNA5.3 AquaMaps Scenario

Task leader: FIN; Participants: CNR;

TNA5.4 FCPPS Scenario

Task leader: FAO; Participants: N/A;

TNA5.5 ICIS Scenario

Task leader: FAO; Participants: N/A;

Deliverables

- *DNA5.1 Communities Practices and Requirements* (M3 regularly updated) documents the analysis of existing practices, tools and workflows. As the deliverable is implemented through a web based tool for managing the requirements, the deliverable also reports on the creation, collection, analysis, validation and tracking of requirements;

- *DNA5.2a-b Communities VREs Validation* (M12, M24) documents the validation of deployed VREs, their functionality and conformance to requirements. Recommendations for future refinement/extension are also included.

1.4.3.4 Summary of staff effort

Participant number	Participant short name	NA1	NA2	NA3	NA4	NA5	Total person months
1	ERCIM	**15**	0	2	2	0	19
2	CNR	4	**11**	9	12	12	48
3	NKUA	1	1	8	**16**	10	36
4	CERN	3	1	3	7	7	21
5	ENG	0	0	0	8	0	8
6	BDM-USTRATH	0	0	3	13	0	16
7	FAO	1	1	**7**	7	**11**	27
8	FIN	0	0	0	5	6	11
9	4D SOFT	0	0	0	0	0	0
10	TERRADUE	0	0	1	8	0	9
Total		**24**	**14**	**33**	**78**	**46**	**195**

1.4.3.5 List of milestones

Milestone number	Milestone name	Work package(s) involved	Expected date	Means of verification
MNA1.1	Formation of the Quality Assurance Task Force (QATF)	NA1	M1	Web pages made available by the QATF for the monitoring of deliverables and milestones status (internal review and submission).
MNA2.1	Formation of Project Management Board	NA2	M1	Public schedule of anticipated meetings of the Project Management Board for the duration of the project.
MNA3.1a	Project Web Site	NA3	M1	The initial web site is available to the public.
MNA5.1	Community Cooperation Environment	NA5	M1	The environment supporting NA5 activities is set up. This environment mainly consists in services for sup-porting collaboration between the communities involved in the project by providing them with an organised view of the project' instruments (e.g. BSCW, Wikis, shared calendars) and outcomes (VREs).
MNA4.1	Approaches to Interoperability and Standards Cooperation Environment	NA4	M2	The environment supporting NA4 activities is set up. It consists in services for supporting collaboration between the work package participants. It promotes a fruitful exchange of information among them (e.g. BSCW, Wiki, shared calendars).
MNA4.2	Interoperability Solutions	NA4	M4	The identified approaches and solutions identified become part of the "Approaches to Interoperability and Standards Cooperation Environment".
MNA3.1b	Project Web Site	NA3	M8	An enhanced version of the initial web site set up at start up of the project. It offers full technical capabilities to the public. It hosts the training environment.
MNA1.2	*Outline for Action Plan for Exploitation and Sustainability*	NA1	M12	A detailed outline is made available serving as the structure for further investigation in the development of the *Action Plan for Exploitation and Sustainability*.

1.4.4 Pert diagram

The diagram below depicts the main relationships between the various tasks of the networking activities. In particular, it presents the coordination role played by NA2 with respect to the operation of the overall project. Moreover, it highlights the central role played by NA4 for realising the knowledge ecosystem and NA5 for exploiting it for implementing the target application scenario through a set of VREs. In fact, the diagram also presents *(i)* how NA5 provides SA and JRA with requirements and guidelines driving the implementation of the knowledge ecosystem and the expected VREs and *(ii)* how SA and, indirectly, JRA satisfy the needs of NA by providing these activities with the knowledge ecosystem and the VREs realising the requirements.

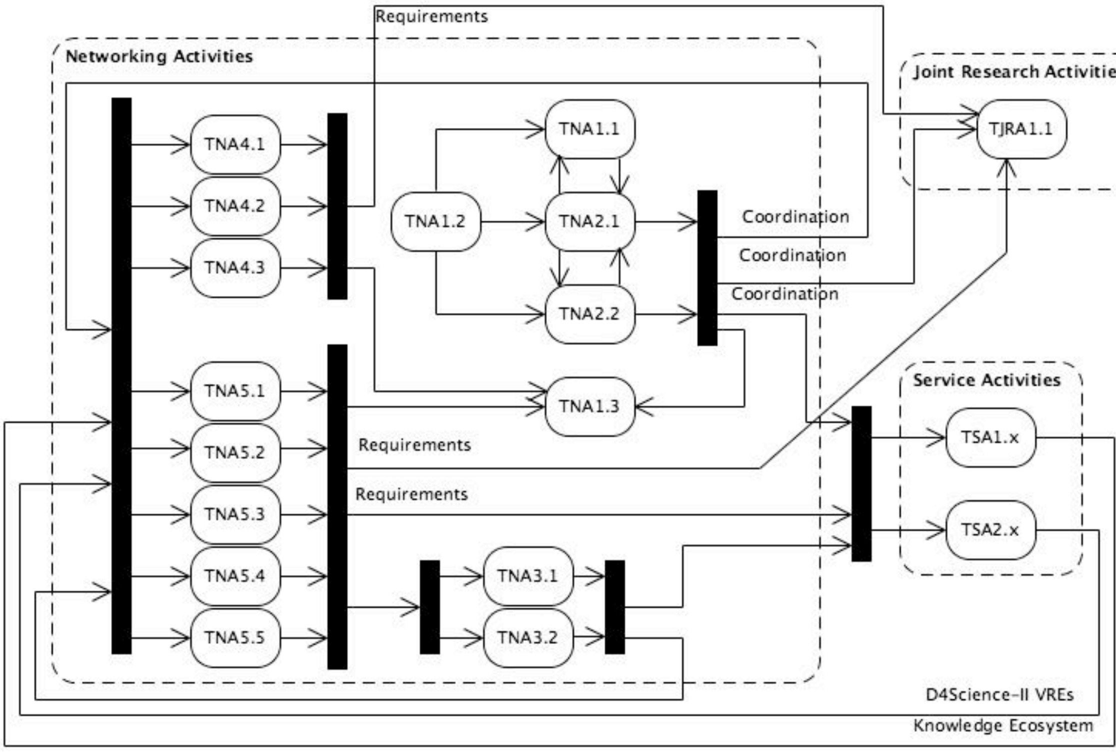


Figure 4. Networking Activities Pert Diagram

1.4.5 Risks Analysis and Contingency Plans

A risk breakdown structure for the NA activities is presented in the following table.

Table 1. Networking Activities Risk Analysis and Contingency Plan

Risk	Evaluation and Description[9]	Contingency Plans
Lack of organisational coherence	*Internal; Low; Medium impact* Could be caused by one or more specific underlying reasons ranging from communication difficulties to particular staffing difficulties at	Initially, the chances of this will be minimised by ensuring the implementation of a strong organizational structure in which the directors/managers are already

[9] Evaluation is expressed through keywords characterising the *provenance* (internal vs external), the *probability* (low, medium, high), and the *impact level* (low impact, medium impact, high impact) respectively.

	one or another of the member organisations.	identified within this proposal. Subsequently this risk can be minimised by ensuring: • Consensual values within the project, • Effective use of communications technology, and • Frequent planning control and review.
High staff turnover	*Internal; Low; Medium impact* Given the complexity of the environment in which D4Science-II operates, skilled staff may leave the project for longer-term and higher paying positions within industry. The virtual nature of the organisation may increase the probability of this risk occurring.	Project management should verify training plans for the younger researchers/developers to ensure that they will continuously evolve within the project. Review staffing reports at PMB meetings. Plan two or three occasions per year where all project staff can get together. Consider staff satisfaction review across the project.
Failure to hire staff on proposed time line	*Internal; Low; Medium impact*	Review hiring strategy, skills required, how and where advertised between consortium members and seek patterns and possible causes for failures and successes. ERCIM has a great deal of experience in hiring skilled staff via its consortium and fellowship programme.
Serious disputes between consortium members	*Internal; Low; Medium impact*	Aim to minimise the chances of disputes occurring by ensuring regular and clear communication between consortium members. Work package leaders should aim to follow an attitude of openness and trust, wherever possible. Where pre-dispute areas are suspected, offline discussions should be initiated. Where disputes become unavoidable, conflict resolutions procedures will be invoked as defined in Section 2.1.4.
Multi-disciplinary nature of the Consortium may lead to disciplines working in silos	*Internal; Medium impact* Lack of communication; limited understanding of needs; difficulties in testing/feedback	Prevent “silo” work packages. Sustain buy-in using face to face meetings and virtual interaction/discussions.
Lack of consensus with community on direction	*Internal; Low; High impact* The success of the project will depend heavily upon the level of community participation and acceptance. This will be illustrated by the fact that the deliverables reflect the needs of the community.	The process to organise and put in place the community will be started immediately and many efforts will be dedicated to the dissemination of results and to the aggregation of associate partners.

Loss of focus or departure from original aims	*Internal; Low; Medium impact*	Strong management structure, with regular planning reviews. Ensure that appropriate evaluation and QA procedures are in place.
Poor community take-up or engagement	*Internal; Medium; High impact*	Strong emphasis on promoting co-ordinating activities. The project will • Regularly assess the needs of the community, • Offer opportunities for the community to contribute to the definition of the co-ordination actions that should be provided, and • Ensure community involvement in the project through strong representation in project management.
Multi-disciplinary nature of the communities may lead to work separately	*Internal; Medium; High impact*	Strong emphasis on promoting co-ordinating activities. The project will • Ensure community involvement in the project through strong representation in project management.
Software fails to deliver expected functionalities or deliver limited ones to the communities	*Internal; Medium; High impact* The success of the project will depend heavily upon the delivery of expected functionalities.	This risk should be monitored and managed at project level. Strong management structure, with regular reviews. Define requirements early in project cycle, so that developers have adequate time to produce and adapt techniques Appropriate monitoring tools should be put in place in order to allow a constant control of the correct evolution of the infrastructure and a prompt reaction to delay.
Knowledge Ecosystem federated Infrastructure not available, supported or functional	*External; Medium; Medium impact;* Each of the federated infrastructures (cf. Section 1.3.2) has its own operational autonomy and lifecycle. In particular, some of them, i.e. GENESI-DR and DRIVER, are operated by European funded projects that will closed at the end of 2009.	In addition to the support that may be given by EU to future projects for supporting the maintenance and extension of these e-Infrastructure, there are other sustainability solutions that are emerging (each of such Infrastructures has its own sustainability plan). For instance, the DRIVER Confederation is working towards the sustainability of this infrastructure by involving National representatives and stakeholders in the area. In addition to that, the D4Science-II consortium includes partners largely involved in the development of each of

		the target e-Infrastructures (Terradue for GENESI-DR, CNR and NKUA for DRIVER, CERN for INSPIRE, FIN for AquaMaps). Technical solutions can be put in place to provide the knowledge ecosystem with the infrastructure expected resources, e.g. re-using the federative technology, implementing ad-hoc wrappers. These factors concur to minimise the risk.
Data/tools access rights difficulties	*External; Medium; Medium impact;* There could be unforeseen access rights issues to some data and/or tools.	Identify critical data early in project lifecycle and arrange formal access agreements.

1.5 Service Activities and associated work plan

1.5.1 Overall strategy

The objective of the Service Activities (SAs) is to deploy and ensure the correct, continuous and effective operation of the resources shared through the D4Science-II ecosystem while enabling their exploitation by the user communities.

In continuing the work of its predecessor, maintaining and enhancing the D4Science production infrastructure, the D4Science-II Service Activity will :

- Maintain the D4Science production sites in operation and extend the infrastructure computing resources;
- Maintain the existing D4Science VOs and VREs in accordance with the existing user communities needs;
- Extend the current set of supported VOs and VREs to the new ones selected in this proposal;
- Apply the integration process to the software released by JRA to incorporate new developments;
- Made available the necessary infrastructure portals for VOs/VREs administration, infrastructure support and VREs access when required;
- Update the infrastructure software repository with the new released functionality developed by JRA for deployment in the context of the supported VREs;
- Establish the procedures needed for the definition of agreements that clarifies relationships and responsibilities between the D4Science-II knowledge ecosystem and other data infrastructures sharing services, clients, and resources;
- Improve the infrastructure operations looking for more efficient (less amount of effort) and effective (better results) ways to deliver production-quality services by applying best practices procedures for services operation, by requiring JRA a higher level of automation and operability of the gCube software.

The Service Activities are planned and organised into three work packages:

- *SA1 – Knowledge Ecosystem Operation.* It coordinates the operation of the infrastructure ensuring continuous and reliable availability. It operates the underline computing infrastructure and makes sure that its resources are properly maintained. This activity also operates the VOs and VREs administration portal used by SA2, deploys the required monitoring tools and maintains the infrastructure software repository.
- *SA2 – VOs and VREs Definition and Operation.* It operates the VOs and VREs serving the targeted user communities, adapting to the evolution of their requirements. At this purpose, this activity is charged to manage VO users and resources, as well to deploy, configure and maintain VRE services. In addition, it provides assistance to other user communities for VO and VREs administration to access, use and exploit the shared resources.
- *SA3 – Software Integration, Testing and Distribution.* It releases reliable and usable software developed by JRA and exploited through the VREs. It coordinates and implement the process of integrating, testing and distributing documented software releases.

Figure 5 shows the interactions among SA work packages and the relationships of SA with the other project activities. The software developed by JRA is integrated, tested and released by SA3. The released software is made available into the infrastructure in the software repository by SA1. SA1 also deploy the infrastructure administration and monitoring portal for the infrastructure support and the VO/VREs administration. SA2 delivers the functionality made available in the software repository to the users by instantiating and orchestrating them into dynamic VREs. SA2 will monitor the progresses of SA3 and SA1 activities in order to plan the communities VREs deployment accordingly and provide SA1 with requirements driving the evolution of the production ecosystem. The users, registered in one of the recognised VOs, exploit the functionality and resources through the VREs. SA1 and SA2 monitor and support the correct operation of the system in collaboration with JRA and NA, making sure the D4Science-

II infrastructure provides a reliable service to the user communities. SA1 and SA2 will provide requirements related to the enhancement of the operation tools to JRA for implementation.

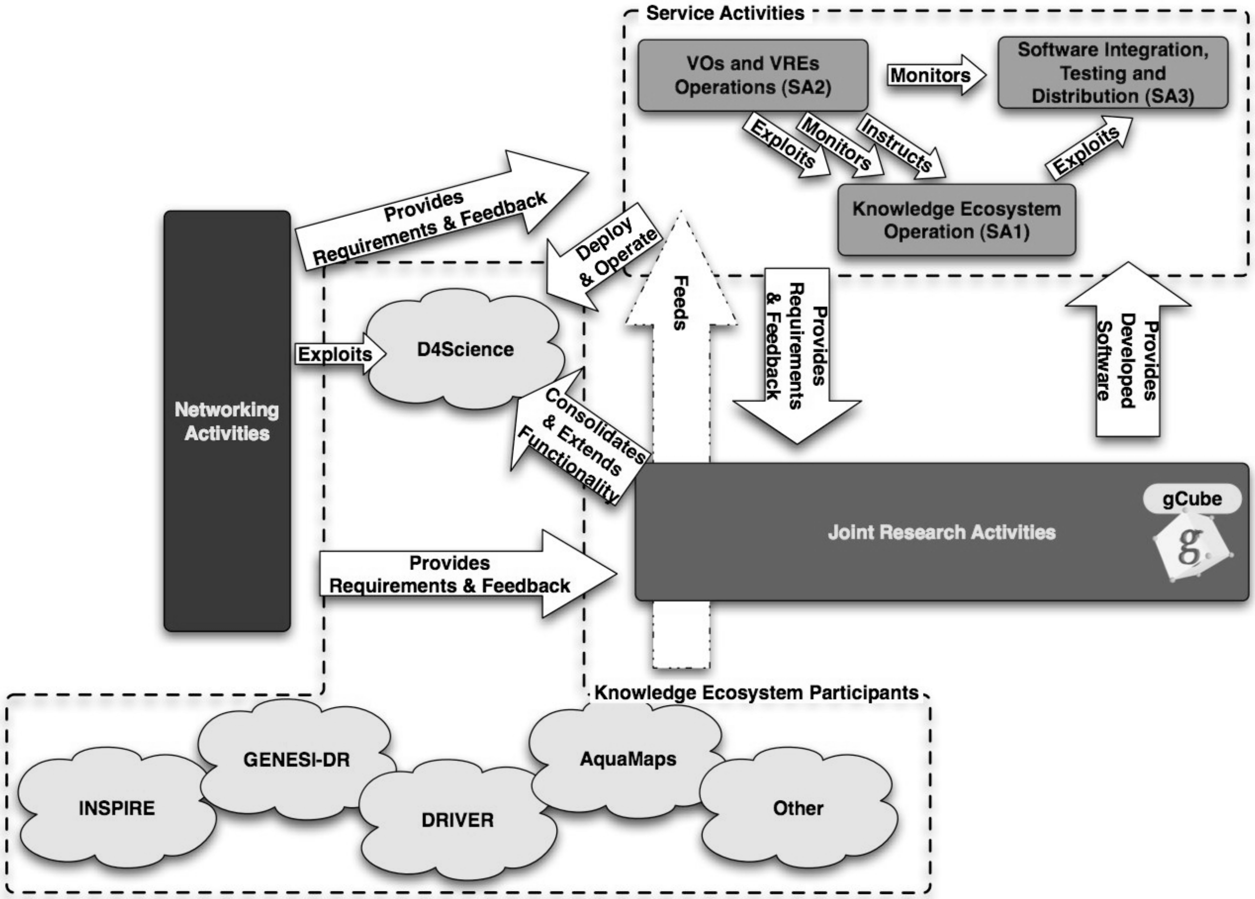
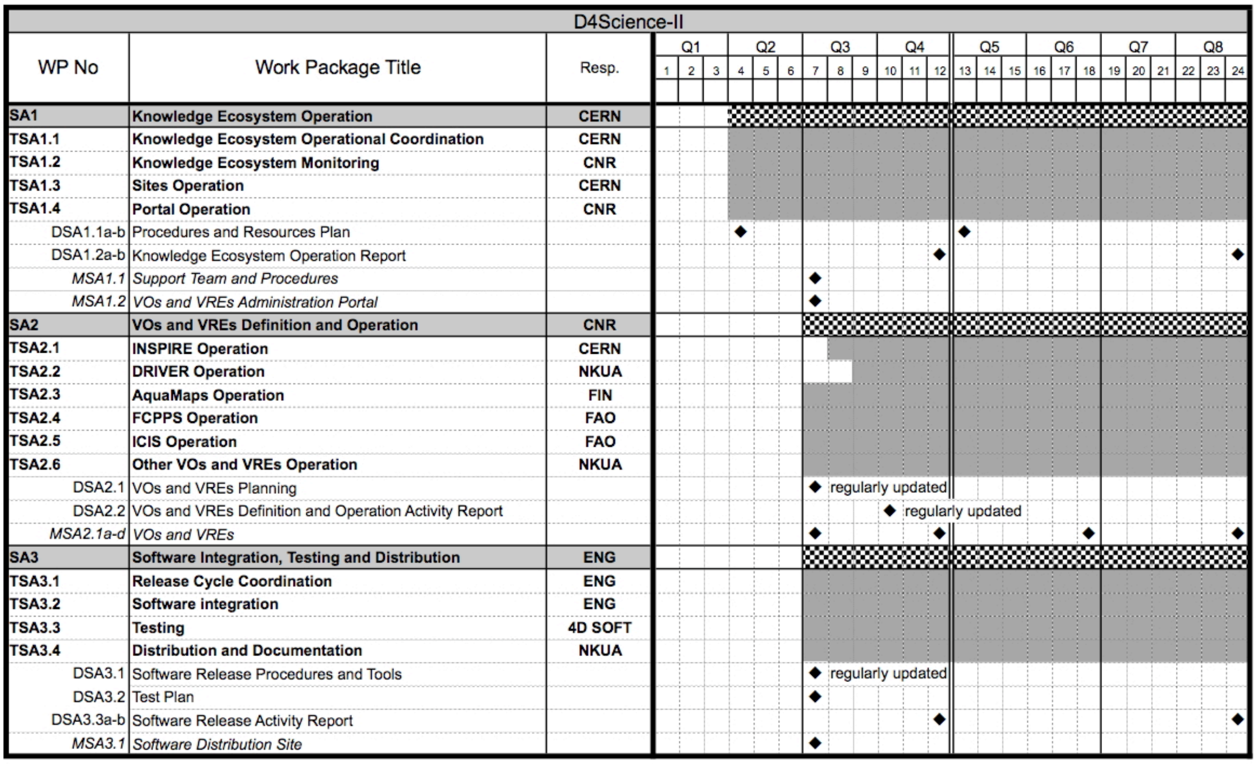


Figure 5. Service Activities Overall Strategy

The D4Science II infrastructure exploits the capability of the **EGEE grid infrastructure** to accomplish computational intensive tasks, for this SA will have to closely collaborate with the EGEE-III project.. D4Science II adopts the ETICS eInfrastructure for testing, integration and configuration of the software developed by the **ETICS 2** project.

1.5.2 GANTT Diagram



1.5.3 Detailed work description

1.5.3.1 Work package list

Work package No	Work package title	Type of activity	Lead participant No	Lead participant short name	Person-months	Start month	End month
SA1	Knowledge Ecosystem Operation	SVC	4	CERN	47	5	24
SA2	VOs and VREs Definition and Operation	SVC	2	CNR	54	7	24
SA3	Software Integration, Testing and Distribution	SVC	5	ENG	40	7	24
	TOTAL				141		

SA1 - Knowledge Ecosystem Operation

Objective of this work package is to effectively deploy and maintain the computing resources shared in the D4Science-II infrastructure and make available the software and tool (infrastructure administration portal) for SA2 to operate the VOs and VREs serving the user communities needs.

SA1 will *operate* the D4Science-II infrastructure in cooperation with the D4Science project during a few months at the beginning of the D4Science-II project lifetime. This will allow transition of know-how and skills between the two projects ensuring at the same time the continuity of the production service provided for the already supported user communities. The overlap between the two projects will also imply saving of some effort on this activity during the first months when the existing sites will be still operated by D4Science. SA1 will in fact maintain the D4Science production sites (gLite and gCube) in operation and extend the infrastructure computing resources progressively during the lifetime of the project. Links will be maintained with EGEE-III in order to negotiate and coordinate access to the EGEE grid infrastructure (see Letter of Support - Appendix B.3).

SA1 will be responsible for the update of the *infrastructure software repository* with the new functionality developed by JRA, released and certified by SA3, for deployment in the context of the supported VREs. SA1 will also enable the exploitation of the infrastructure by making available the *administration and monitoring portal* - the core management tool of the infrastructure - for VOs/VREs administration, infrastructure support and VREs access when required.

In addition, *feedback* will be provided to JRA for the enhancement of the operational tools. A higher level of automation and operability will be required to the software operated in the infrastructure, gCube, in compatibility with its evolution and with the objective to improve the service support activity with a more efficient first level support and incident management process.

SA1 will undertake a strong *coordination* role in the operation of the infrastructure by planning the deployment of the infrastructure and defining the procedures needed to ensure its proper functioning, including the operation and support tools and teams (DSA1.1a-b). SA1 will adopt and enhance the operational procedures and tools already operating the D4Science infrastructure. In this direction, SA1 will:

- Apply *best practices procedures* for services operation toward the improvement of the infrastructure operations looking for more efficient (less amount of effort) and effective (better results) ways to deliver production-quality services. Good operating practices will be applied to service delivery and service support, and may as well be investigated in other domains as the maturity and needs of the infrastructure evolve;
- Define specific *interoperability agreements*, i.e. policies and procedures governing its interoperation with collaborating scientific infrastructures, repositories and services. They will define which part of another infrastructure will be accessible by D4Science services (e.g. access policies, copyright policies) and how (e.g. only local elaborations of data, security policies, frequency of updates, etc.). In some

cases, knowledge of the VREs to be operated may be needed and will be taken into account in collaboration with SA2;

- Formalise a *resource sharing policy*, in line with the infrastructure sustainability strategy and the interoperability agreements, to regulate the overall sharing and usage of the interfaced community specific resources (in collaboration with SA2), as well as the access and allocation of the D4Science computing resources. This is needed in order to ensure a peaceful and satisfactory cooperation between the infrastructure user communities and the resource providers;
- Contribute to the improvement of the *infrastructure deployment strategy* in collaboration with SA2;
- Require *quality assessments* on the functioning and usage of the infrastructure to evaluate the level of maturity achieved and enable the infrastructure to evolve to the quality of services required.

SA1 will cooperate closely with SA2 to plan and optimise the infrastructure deployment, usage, and performances. Support to the infrastructure will be organised in coordination with SA2, JRA and NA in order to ensure effective resolution of the support requests. New functionality will be integrated in the infrastructure software repository as they are delivered by SA3.

DSA1.2a-b will periodically report on the status of the D4Science-II Ecosystem, while two milestones will be the checkpoint of major achievements. MSA1.1 establishes the time when new operational procedures are defined and implemented, while MSA1.2 is achieved when the VOs and VREs administration portal will be operational.

SA2 - VOs and VREs Definition and Operation

The objective of this work package is to enable the user communities to successfully exploit the functionality and resources offered by the D4Science-II ecosystem. This is achieved by ensuring availability, reliability and appropriate configuration of the supported VREs. In particular, this work package will take care (by relying on the facilities made available by the SA1 work package, namely the Portal supporting the VOs and VREs operation – MSA1.2) to manage:

- the *Virtual Organisations* needed to define the communities operational context, i.e. the set of users forming each community and the set of resources that can be used in the community activities;
- the *Virtual Research Environments* needed to define the community application context, i.e. the set of users partaking to a community specific VRE and the set of resources partaking to the specific application context;
- the *community specific resources*, i.e. the set of data sources, services and tools that are needed to operate the community specific VREs and that, according to their policies, can eventually be made available to the rest of the D4Science-II Ecosystem by promoting sharing and reuse across community boundaries.

The whole activity of this work package will be conducted by having the *user communities as primary actors*, i.e. representatives of such communities are expected to lead the activity needed to operate their VOs and VREs both by raising requirements and experiencing with the tools needed to operate them. The user community activity will be supported by the technical partners with the needed experience and expertises. The cooperation of these two teams guarantees on the one hand a *rapid and smooth operation* of the expected environment, on the other hand, it implies a *knowledge transfer activity* that will provide the community with the knowledge needed to operate the available instruments and ensure its future operational independence and sustainability.

The whole SA2 activity is organised in tasks dedicated to manage the operation of VOs and VREs in a specific scenario. All such tasks will be characterised by a continuous loop of *resources planning* - *deployment/configuration* - *allocation* - *monitoring/maintenance* to guarantee that the resources needed to meet the community expectations are available for usage. The VOs and VREs availability in the D4Science-II Ecosystem will be an important achievement for the project, recognised by MSA2.1a-d.

SA2 is planned to continue the activity and the service of the D4Science project, dealing with the operation of VOs and VREs scenarios supported by the ongoing D4Science activity and in accordance with the D4Science user communities needs. In particular, ImpECt [11] VOs and VREs will be maintained as part of TSA2.6. The overlap between the two projects will imply saving of some effort on this activity during the first months when the existing VOs and VREs will be still operated by D4Science. In parallel with that, new

VOs and VREs will be created, as well as existing one will be empowered as to take full advantage from the D4Science-II Ecosystem, e.g. cross-community usage of data or services.

In the VOs and VREs configuration processes, SA2 will make sure the *specific interoperability agreements* are respected and properly implemented, in collaboration with SA1. SA2 will also improve the VREs *deployment strategy* by further investigating on the VREs services deployment requirements to optimise the usage of the available resources and increase services reliability and performances (e.g. with service replication).

The overall work package activity will manifest, besides the VOs and VREs, in two deliverables *DSA2.1 VOs and VREs Planning* and *DSA2.2 VOs and VREs Operation Activity Report* providing an integrated and continuously updated view of the VOs and VREs in action in the D4Science-II Ecosystem.

SA2 will closely interact with other work packages belonging to both the SA area as well as the JRA and NA areas. With regard to the SA area this work package will monitor the progresses of SA3 and SA1 activities in order to plan the communities VREs deployment accordingly. In addition SA2 will deploy (*i*) the community specific resources contributing to populate and form the D4Science-II Ecosystem, e.g. data collections, application specific services, and (*ii*) the VOs and VREs addressing the community scenarios. With regard to the JRA area, SA2 will provide the technicians developing the D4Science-II technology with feedback on the infrastructure operation in the context of SA2, operational requirements will be formulated in relation with the VOs and VREs administration portal. This will reduce the technical support needed from the VRE managers and VO managers to operate the infrastructure. Finally, with regard to NA area, SA2 will mainly interact with NA5 to understand both the requirements and the feedback resulting from the actual exploitation of the deployed VOs and VREs in accomplishing the envisaged scenarios and consequently put in place the action needed to improve the quality of the service offered to the user communities.

SA3 - Software Integration, Testing and Distribution

This work package is devoted to the release of the software developed by JRA and distributed within the project lifetime in the D4Science-II infrastructure. The main objective is therefore to coordinate and implement the build, integration, testing and distribution of the code to be released in a deployable documented software distribution. This will be achieved through the integration, documentation and packaging activities; while software reliability and usability will be ensured by applying acceptance rules and testing.

At this purpose the project is willing to adopt and exploit the functionalities made available by the ETICS system, already used for the gCube integration activity under the context of the D4Science project. In particular, the project has been designed to operate the ETICS tool by its own. The strong collaboration and feedback provided by D4Science to ETICS guarantee that the latest developed version of this tool has adequate characteristics to serve the needs of the D4Science-II project. The few months overlap between D4Science and D4Science-II projects will allow transition of know-how for this activity and will imply saving of some effort being the gCube software still integrated in the context of the D4Scinece SA3 activity during the first months of D4Science-II.

Software integration and release is a continuous process during the lifetime of the project and will follow the structure and timing defined in the overall planning of SA3 in DSA3.1.

SA3 will be managed building on the experience gained during the DILIGENT and D4Science projects. The release cycle organisation, as the integration activity have proven to be effectively managed in the past. Nevertheless, enhanced software release procedures will be documented in DSA3.1. The testing activity will be improved in the direction of achieving better coverage and optimise the testing efforts and planned in the Test Plan document (DSA3.2). MSA3.1 will acknowledge the availability of the updated software distribution site.

1.5.3.2 List of Deliverables

Del. no.	Deliverable name	WP no.	Nature	Dissemi-nation level	Delivery date (proj. month)
DSA1.1a-b	Procedures and Resources Plan	SA1	O	PU	M5, M13
DSA2.1	VOs and VREs Planning	SA2	O	PU	M7 (regularly updated)
DSA3.1	Software Release Procedures and Tools	SA3	O	PU	M7 (regularly updated)
DSA3.2	Test Plan	SA3	O	PU	M7
DSA2.2	VOs and VREs Operation Activity Report	SA2	O	PU	M10 (regularly updated)
DSA1.2a-b	Knowledge Ecosystem Operation Report	SA1	O	PU	M12, M24
DSA3.3a-b	Software Release Activity Report	SA3	O	PU	M12, M24

1.5.3.3 Work package descriptions

Work package number	SA1			**Start date or starting event:**				M4	
Work package title	Knowledge Ecosystem Operation								
Activity Type	SVC								
Participant number	1	2	3	**4**	5	6	7	8	9	10
Participant short name	ERCIM	CNR	NKUA	**CERN**	ENG	BDM-USTRATH	FAO	FIN	4D SOFT	TERRA DUE
Person-months per participant		17	8	13	3		2	4		

Objectives

The objective of this work package is to ensure availability and reliability of the computing resources shared through the D4Science-II infrastructure; as well to enable the exploitation of the D4Science-II Ecosystem resources via the VOs and VREs administration portal.

Description of work

Work package leader: CERN;

TSA1.1: Knowledge Ecosystem Operational Coordination

Task leader: CERN; Participants: CNR;

- plan the overall deployment and evolution of the infrastructure;
- establish the operational procedures and teams;
- define interoperability agreements and resources sharing policy;
- provide operational requirements to JRA;
- negotiate and coordinate access to non-partners grid sites, through EGEE and related projects for gLite sites.

TSA1.2: Knowledge Ecosystem Monitoring

Task leader: CNR; Participants: CERN;

- infrastructure monitoring through the usage and enhancement of operational tools;
- opening, assignment and follow-up on support tasks;
- reporting on software bugs and operability, resources availability and usage.

TSA1.3: Sites Operation

Task leader: CERN; Participants: CNR, NKUA, ENG, FAO, FIN;

- sites deployment, certification, upgrade, operation (gLite nodes and gCube Hosting Nodes);
- support to new sites joining the infrastructure;
- deployment, upgrade, operation of gCube core services;
- update of the software repository.

TSA1.4: Portal operation

Task leader: CNR; Participants: NKUA;

- deployment and operation of the VOs and VREs administration portal;
- creation of new VOs.

Deliverables

- *DSA1.1a-b Procedures and resources plan* (M4, M13) defines the organisational structure, the procedures and the tools adopted to ensure the correct and effective operation, upgrade, extension of the infrastructure; as well as the overall deployment plan;

- *DSA1.2a-b Knowledge Ecosystem operation report* (M12, M24) reports on the status of the production infrastructure in terms of sites, middleware, quality of the service provided and usage.

Work package number	SA2			**Start date or starting event:**				M7		
Work package title	VOs and VREs Definition and Operation									
Activity Type	SVC									
Participant number	1	**2**	3	4	5	6	7	8	9	10
Participant short name	ERCIM	**CNR**	NKUA	CERN	ENG	BDM-USTRATH	FAO	FIN	4D SOFT	TERRADUE
Person-months per participant		17	10	7			4	10		6

Objectives

The objective of this work package is to enable the user communities to successfully exploit the functionality and resources offered by the D4Science-II Ecosystem. This is achieved by ensuring availability, reliability and appropriate configuration of VOs and VREs.

Description of work

Work package leader: CNR;

TSA2.1 INSPIRE Operation

Task leader: CERN; Participants: CNR;

TSA2.2 DRIVER Operation

Task leader: NKUA; Participants: N/A;

TSA2.3 AquaMaps Operation

Task leader: FIN; Participants: CNR, TERRADUE;

TSA2.4 FCPPS Operation

Task leader: FAO; Participants: CNR, TERRADUE;

TSA2.5 ICIS Operation

Task leader: FAO; Participants: CNR, TERRADUE;

TSA2.6 Other VOs and VREs Operation

Task leader: NKUA; Participants: xxx;

For Tasks 2.1 to 2.6 the same actions plan applies, although the focus of each corresponds to a different scenario (cf. Section 1.3).

- VO and VREs operation planning;
- VO and VREs deployment and configuration;
- VO and VREs monitoring and support.

Deliverables

- *DSA2.1 VOs and VREs Planning (M7 regularly updated)* describes the steps with the relative deadline needed to accomplish the VOs and VREs needed to meet the scenario expectations;
- *DSA2.2 VOs and VREs Operation Activity Report (M10 regularly updated)* summarises the evolution of the VOs and VREs running in the D4Science-II Ecosystem in terms of, e.g., applications created, resources deployed, user served;

Work package number	SA3			**Start date or starting event:**				M7	
Work package title	Software Integration, Testing and Distribution									
Activity Type	SVC									

Participant number	1	2	3	4	**5**	6	7	8	9	10
Participant short name	ERCIM	CNR	NKUA	CERN	**ENG**	BDM-USTRATH	FAO	FIN	4D SOFT	TERRA DUE
Person-months per participant		1	3	5	15				16	

Objectives

Objective of this work package is to release a quality D4Science software distribution, ensuring software reliability and usability applying testing and acceptance rules.

Description of work

Work package leader: ENG;

TSA3.1: Release Cycle Coordination

Task leader: ENG; Participants: CNR, CERN;

- overall coordination of the integration, testing and distribution activity;
- definition of software releases procedures and plans;
- definition of software acceptance conditions;

TSA3.2: Software Integration

Task leader: ENG; Participants: CERN;

- adoption of the ETICS build tool;
- software integration;

TSA3.3: Testing

Task leader: 4D SOFT; Participants: CERN;

- definition of a test plan;
- adoption of the test tool as plug-in of ETICS;
- set up and maintenance of the testing infrastructure, including the provision of resources;
- coordination and implementation of the testing activities;

TSA3.4: Documentation and Distribution

Task leader: NKUA; Participants: ENG;

- coordination of the packaging activity;
- software distribution and documentation;

Deliverables

- *DSA3.1 Software Release Procedures and Tools (M7 regularly updated)* describes the structure, the timing and the tools for documentation, integration, testing and distribution activities;
- *DSA3.2 Test Plan (M7)* describes the test strategy and structure, test priorities and guidelines;
- *DSA3.3a-b Software Release Activity Report (M12, M24)* reports on the outcome of the release activity performed in the periods.

1.5.3.4 Summary of staff effort

Participant number	Participant short name	SA1	SA2	SA3	Total person months
1	ERCIM	0	0	0	0
2	CNR	17	**17**	1	35
3	NKUA	8	10	3	21
4	CERN	**13**	7	5	25
5	ENG	3	0	**15**	18
6	BDM-USTRATH	0	0	0	0
7	FAO	2	4	0	6
8	FIN	4	10	0	14
9	4D SOFT	0	0	16	16
10	TERRADUE	0	6	0	6
Total		**47**	**54**	**40**	**141**

1.5.3.5 List of milestones

Milestone number	Milestone name	Work package(s) involved	Expected date	Means of verification
MSA1.1	Support Team and Procedures	SA1	M7	The D4Science-II support procedures are activated, the infrastructure is constantly operated and maintained. The reactions to the incidents occurring in the infrastructure conforming to the defined procedures will prove the achievement of this milestone.
MSA1.2	VOs and VREs Administration Portal	SA1	M7	Availability of the D4Science-II administration portal to SA2 for VOs and VREs operation.
MSA2.1a-d	VOs and VREs	SA2	M7, M12, M18, M24	The VOs and VREs are registered and can be checked through the D4Science-II monitoring tools. Four version of such a Milestone will be realised each resulting from the VOs and VREs deployed at the Milestone release time. The schedule according which VOs and VREs implementing the target scenarios will be defined during the project lifetime (DSA2.1).
MSA3.1	Software Distribution Site	SA3	M7	The gCube distribution site is published.

1.5.4 Pert diagram

The diagram below depicts the main relationships between the various tasks of the Service Activities. In particular, it shows how the three work packages cooperate to realise and operate the D4Science-II knowledge ecosystem and the expected VREs. TSA1.1 is the task called upon to orchestrate the operation of the ecosystem so as to satisfy the NA expectations. The result of the SA1 activities materialise in the knowledge ecosystem and in the portal through which this ecosystem can be used to create the VREs. This facility is then used by tasks TSA2.1-TSA2.5 to design and deploy the VREs needed to serve the target scenarios (cf. Section 1.3.3). SA activities are completed by the SA3 work package whose tasks receive the new versions of the software produced by JRA and provide SA1 with the set of integrated (TSA3.2), tested (TSA3.3), documented and packaged (TSA3.4) software artefacts that are needed to operate the knowledge ecosystem and the VREs. In addition to the relations among the SA tasks, the diagram depicts how the requirements driving the development of the knowledge ecosystem and the expected VREs flow from the NA to the SA as well as how the requirements and feedback resulting from the exploitation of the produced software within concrete scenario flow from SA to JRA to plan further enhancement activities.

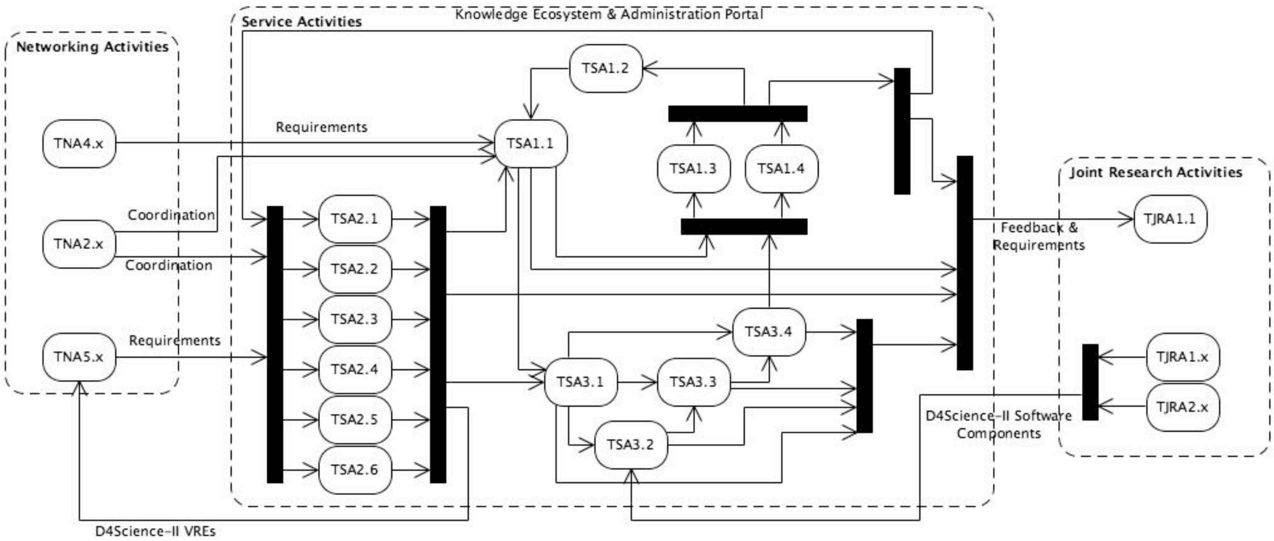


Figure 6. Service Activities Pert Diagram

1.5.5 Risks Analysis and Contingency Plans

A risk breakdown structure for the SA activities is presented in the following table.

Table 2. Service Activities Risk Analysis and Contingency Plans

Risk	Evaluation and Description[10]	Contingency Plans
Software not released on time	*Internal; Medium; High impact*; SA3 does not release the software on time or JRA fails in delivering on time. This will result in the new/enhanced functionality to be deployed late in the Knowledge Ecosystem and a potential delay for the NA activities.	This risk should be monitored and managed at project level. Careful planning, monitoring and structuring of the integration activity should contribute to minimise this risk. Moreover, the inheritance of D4Science technology is another factor concurring to minimise this risk. Fixing of critical

[10] Evaluation is expressed through three keywords characterising the *provenance* (internal vs external), the *probability* (low, medium, high), and the *impact level* (low impact, medium impact, high impact) respectively.

		bugs should be implemented and deploy with higher priority in order to ensure the stability of the Knowledge Ecosystem.
Knowledge Ecosystem not released on time	*Internal; Low; High impact*; If problems are encountered in delivery the envisaged ecosystem a delay is introduced in NA5 tasks thus hindering user communities in exploiting the expected facilities.	Appropriate monitoring tools should be put in place in order to allow a constant control of the correct evolution of the overall knowledge ecosystem and a prompt reaction to delay.
VOs and VREs administration portal not released on time	*Internal; Low; High impact*; If problems are encountered in delivering the administration portal a delay is introduced in NA5 tasks thus hindering to create the expected VREs.	An instance of the portal will be internally available since the beginning (it is a D4Science outcome). Its function will be continuously monitored and the new facilities will be internally and carefully tested to minimise service malfunctions.
Knowledge Ecosystem not available	*Internal; Low; High impact*; If problems are encountered with the exploitation of the knowledge ecosystem a delay is introduced in NA5 task thus hindering user communities in exploiting the expected facilities.	Monitoring tools should be put in place to control the functioning of the overall ecosystem, i.e. the D4Science infrastructure as well as the services implementing the bridges between the participating Infrastructures, and promptly react to malfunctions.
VOs and VREs administration portal not available	*Internal; Low; High impact*; If problems are encountered with the exploitation of the VOs and VREs administration portal a delay is introduced in NA5 task thus hindering user communities in experiencing with the expected VREs.	Monitoring tools should be put in place to control the functioning of the VOs and VREs administration portal and promptly react to malfunctions.
VRE not available	*Internal; Low; Medium impact*; If problems are encountered while operating a VRE a delay is introduced in NA5 task thus hindering user communities in experiencing with the expected VRE.	The technology supporting VREs has been developed in the context of the D4Science project. A "minimal" VRE can always be operated and this contributes to minimise the risk.

1.6 Joint Research Activities and associated work plan

1.6.1 Overall strategy

The objective of the Joint Research Activities (JRAs) is to develop and consolidate the technology needed *(i)* to support the implementation of the knowledge ecosystem D4Science-II has to operate (cf. Section 1.3) as well as *(ii)* to serve the target scenarios and the scientific communities directly benefiting of this ecosystem (cf. Section 1.3.3). Such a technology will largely be based on the existing gCube system [30], i.e., the system originally developed in DILIGENT and subsequently consolidated and extended during the D4Science project to build and operate e-Infrastructures supporting Virtual Research Environments. In the context of this project this enabling technology has to be reinforced to properly deal with the new and challenging application scenario, i.e. the implementation and operation of a knowledge ecosystem. One of the major expected results will be the technology needed to build a *common space of resources* from the integration of the constituent e-Infrastructures resources and to guarantee a seamless access to those resources. In addition to that, advanced services for data management benefiting of the envisaged ecosystem will be delivered. The second dimension of technology enhancement is the implementation of *innovative services* that will contribute to enlarge the pool of resources equipping the ecosystem and thus its usage scenarios. Such services appear to be required in specific application scenarios (cf. Section 1.3.3) but if properly implemented will also be used/usable in other contexts.

To achieve this objective a set of activities are planned and organised into a number of interacting work packages:

- *JRA1 – Knowledge Ecosystem Implementation.* It will organise and put in place the activities needed to implement the ecosystem enabling technology. These activities include: (*i*) the analysis of the requirements driving the ecosystem building and exploitation as well as the high-level design of the technological enhancements – requirements to be managed are related to how interoperability between each target e-Infrastructure and D4Science is dealt (NA4), to the scenarios' enhanced services (NA5) as well as to the experiences gained while operating the ecosystem and the VREs (SA1 and SA2); (*ii*) the implementation activities leading to a generic interoperability framework and its add-ons addressing single infrastructures' peculiarities; (*iii*) the implementation activities leading to an access interface for the consumption of gCube facilities; and (*iv*) the technology consolidation and enhancement activities needed to transform gCube into a system supporting the envisaged knowledge ecosystem.
- *JRA2 – Scenario Specific Implementation.* It will cover the needs for detailed design and implementation which will be raised by each of the scenarios supported by this proposal (cf. Section 1.3.3). It includes specific tasks devoted to the implementation of innovative services benefiting of their operating in the ecosystem as well as the adaptation of existing services to either enlarge their supported functionality or reinforce their operational behaviour. Examples of the technology enhancements that will be conducted in the context of this work package are: services performing innovative computational intensive bibliometric data mining (cf. Section 1.3.3.1), services implementing enhanced publication models (cf. Section 1.3.3.2), services providing access to more precise and complete species distribution maps (cf. Section 1.3.3.3), services enhancing the production of country-level reports (cf. Section 1.3.3.4), services supporting innovative fish stock monitoring and assessment activities by providing advanced views of catch data resulting from new predictive models (cf. Section 1.3.3.5). Despite such a technology will be developed to satisfy the needs arising in specific application contexts, effort will be spent to make it as much generic and extendible as possible in order to maximise its adaptation and usage in other scenarios served by the ecosystem. Each service will become a commodity contributing to populate the ecosystem.

Figure 7 shows the interactions among JRA work packages and the relationships of JRA with the other project activities.

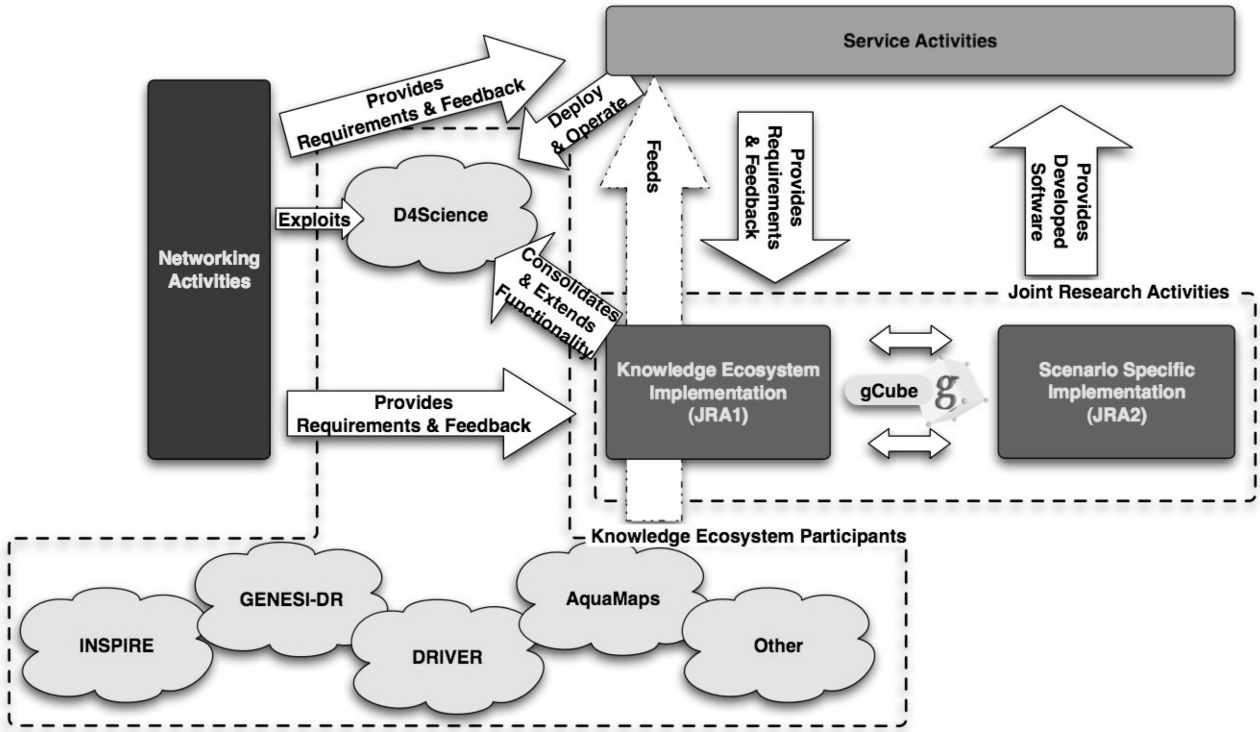
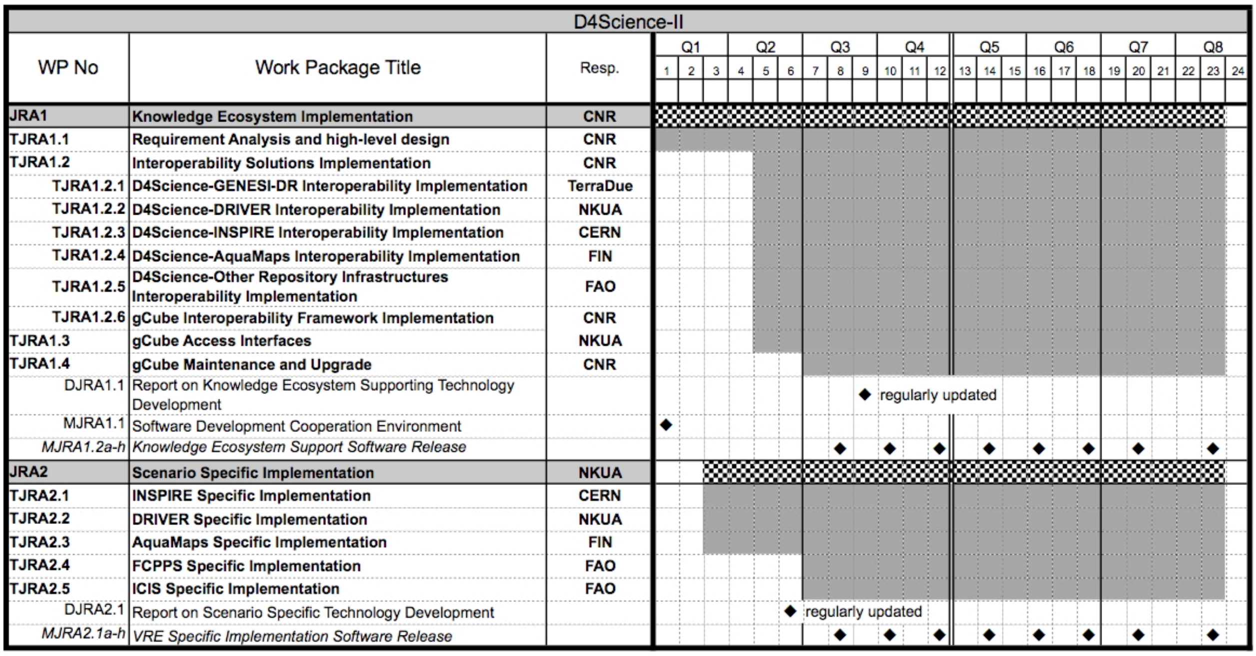


Figure 7. Joint Research Activities Overall Strategy

In this figure, it is shown how the requirements arising (*i*) in the context of the networking activities while defining and exploiting the envisaged ecosystem and the VREs supporting the D4Science-II scenarios, and (*ii*) in the context of the service activities while operating them, are handled by JRA1 and drive the consolidation and enhancement of existing technologies and the development of innovative software solutions. Moreover, the figure shows how the developed technology serves the communities through the SA activities that integrate, test and deploy the services implementing the ecosystem and the VREs.

1.6.2 GANTT Diagram



1.6.3 Detailed work description

1.6.3.1 Work package list

Work package No	Work package title	Type of activity	Lead participant No	Lead participant short name	Person-months	Start month	End month
JRA1	Knowledge Ecosystem Implementation	RTD	2	CNR	157	1	23
JRA2	Scenario Specific Implementation	RTD	3	NKUA	100	3	23
	TOTAL				257		

JRA1 - Knowledge Ecosystem Implementation

The objective of this work package is to coordinate and implement the development of the software system needed to build and operate the knowledge ecosystem objective of this proposal. To reach this objective, the work package puts in place four main lines of activities: (*i*) the *analysis of user requirements* and the identification of the technical solutions needed to satisfy them; (*ii*) the development of an innovative and powerful *interoperability framework*; (*iii*) the implementation of an *access interface* for programmatic consumption of ecosystem resources and facilities through gCube; and (*iv*) the *consolidation and enhancement of the gCube* technology.

The first line of activities is dedicated to collect and analyse the requirements arising in various contexts. Such requirements may include specific community scenario desiderata for innovative services promoting interoperability between the target e-Infrastructures (cf. Section 1.3.2) – actually between their software systems – as well as needs and feedback resulting from the exploitation of the gCube technology to operate the ecosystem or implement a VRE. This task will analyse the collected requests in order to elaborate a coordinated and coherent development plan assigning the implementation of the needed functionality or functionality enhancement to the proper system areas. Because of the central role of this line, all the technical partners participating in the development of gCube are involved in its activity so as to identify the better solutions for satisfying the exposed requirements. To support such an activity the software development cooperation environment that has been successfully exploited during D4Science (mainly consisting of an issue tracking system – TRAC, a Wiki, a shared work space – BSCW, and a version control system – SVN) will be adapted to satisfy the D4Science-II needs (MJRA1.1). By properly exploiting this tool the coordinated development plan will be designed and monitored. In fact, the proposal is based on the exploitation of an Agile-like software development methodology [8], i.e. software components will be developed in subsequent relatively short iterations, each containing the tasks needed to release new functionality or revise currently supported functionality in different ways – requirement analysis, planning, detailed design, implementation, integration, testing and documentation. This way the technical development will be prompt to quickly reply to expectations and adapt to changes with subsequent releases of the developed software (MJRA1.2a-h). The characteristics of each release will be carefully documented in a deliverable providing an ever update description of the system characteristics and capabilities (DJRA1.1).

The second line of activities is dedicated to implement the software solutions needed to make the target e-Infrastructures (cf. Section 1.3.2) interoperable with the D4Science one. This task will rely on the activity conducted in the context of NA4 in which key representatives of each e-Infrastructure will (*i*) identify and discuss the interoperability issues to be solved to make their systems interoperable and (*ii*) spent effort in analyzing existing standards and approaches to the identified issues. By relying on the knowledge resulted from NA4, D4Science-II software designers and engineers will design an *innovative and powerful gCube interoperability framework* that provides generic solutions and reference implementations for various interoperability issues. Since this framework is a fundamental element of the knowledge ecosystem enabling

technology, it must be truly extensible and easily adaptable to effectively solve the interoperability cases that have to be faced when federating e-Infrastructures and data repositories. The effectiveness of the framework will be tested in the context of this proposal itself, since the gCube Interoperability Framework is the technology that will be used to realize the interoperability with the target e-Infrastructures. The framework will be extended by deriving *specific implementations* and *plug-ins* needed to make each of the target e-Infrastructure interoperable with D4Science.
The third line of activities is dedicated to the other face of the interoperability issue, i.e. making the resources forming the ecosystem exploitable in a programmatic way through the gCube mediation. In fact, through the gCube Interoperability Framework this system becomes the aggregator of the resources from the federated e-Infrastructures and data services. A new *gCube Access Interface* will be implemented so as to give service providers a single access point to the gCube facilities. Thus providers will be freed from the difficulties intrinsic to the interaction with a service-oriented infrastructure consisting of thousands, highly dynamic and interoperating services as a gCube-based Infrastructure is.

The fourth line of activities is dedicated to *consolidate and enhance* the services currently forming the *gCube* system. During the D4Science project gCube services have been reinforced and reengineered so as to make this system capable to operate a production quality e-Infrastructure supporting Virtual Research Environments. In the context of D4Science-II the consolidation and enhancement activities will go in the direction of making gCube a system capable of operating a production quality knowledge ecosystem. The tasks related to the interoperability cover part of the needed reinforcement activity. In these tasks technological enhancements and solutions needed to deal with the novel resources resulting from the integration of heterogeneous e-Infrastructures (e.g. changes in the resource model supported, the resource sharing approaches, the data organization and retrieval services, the presentation services) will be implemented.

JRA2 - Scenario Specific Implementation

The starting points of the activity of this work package are the requirements. These depends on the respective scenarios, the technologies to support them and the technological environment surrounding the project. Requirements drive the materialization of each VRE implementing the scenario, in terms of case-specific analysis, software design and finally implementation.

Delving into the details of each scenario, it becomes evident that case-specific implementation can go into several directions and has to cope with completely different challenges, due to the fact that different resources and information are to be made available in both directions. More particularly the following activities will be part of the work plan so as to meet the objectives of the activity:

- *Implementation of new components that exploit the infrastructure offered by the Grid / gCube and capture case-specific needs and requirements*. The implementation of advanced algorithms to analyze author networks, citation networks, hybrid metrics, paper-trail of a scientific idea starting from a paper through all citations, as well as further work this will trigger, are examples of this type of development, being performed in the INSPIRE scenario. It is also what AquaMaps directly foresees in exploiting the infrastructure capacities for performing statistic analysis, while all cases have the opportunity to directly perform similar tasks (monitoring and assessment, prediction, knowledge management etc);
- *Porting of existing, community specific components to the D4Science e-Infrastructure created by the project*. Components are not restricted to functional software but also to components that will allow the migration of data and procedures to the new infrastructures. Porting ICIS/FCCPS/Biodiversity/EM cases analysis tools to the infrastructure, are solid examples of this class of implementations.
- *Exploitation of gCube's native facilities, as they are being offered by its core services via development of components that will operate upon the interoperability boundary*. Firing IR queries against gCube hosted repositories, accessing content placed on its storage space, exploiting metadata and content brokerage, are examples that will be among the major objectives of the DRIVER case but are also of common interest to all cases.
- *Exploitation of resources offered by other cases, be it data, physical system or services*. Exploiting the advanced analysis features of repositories developed by INSPIRE in DRIVER repositories is a typical case of such cross-domain fertilisation of project's work.

- *Exposure of case specific resources to the rest of the infrastructure stakeholders*. Bringing the massive DRIVER repositories to the rest of the infrastructure, as well as the cross-infrastructure delivery of data and information for the earth and life observation cases are solid examples of such implementations.

The AquaMaps, GENESI-DR, FCCPS and ICIS all have mutual interest for exploiting each others' resources, esp. high quality, curated, specialized data repositories (FishBase, OBIS, GBIF, UN-DATA, COMTRADE, ARGOS sensors data), which apply for the last two cases of the previous list.

The software solutions developed in this work package to implement the envisaged innovative services will be reported in DJRA2.1. This deliverable, implemented through a Wiki, will be initially, released at M8 and then regularly updated to facilitate the sharing and re-use of solutions.

Starting from M8, successive versions of the developed application specific services will be bimonthly released (MJRA1.2a-h).

1.6.3.2 List of Deliverables

Del. no.	Deliverable name	WP no.	Nature	Dissemi-nation level	Delivery date (proj. month)
DJRA2.1	Report on Scenario Specific Technology Development	JRA2	O	PU	M6 (regularly updated)
DJRA1.1	Report on Knowledge Ecosystem Supporting Technology Development	JRA1	O	PU	M9 (regularly updated)

1.6.3.3 Work package descriptions

Work package number	JRA1			**Start date or starting event:**				M1	
Work package title	Knowledge Ecosystem Implementation								
Activity Type	RTD								

Participant number	1	**2**	3	4	**5**	6	7	8	9	10
Participant short name	ERCIM	**CNR**	NKUA	CERN	**ENG**	BDM-USTRATH	FAO	FIN	4D SOFT	TERRADUE
Person-months per participant		38	37	14	12	25	4	20		7

Objectives

The goal of this work package is to develop the technology needed to implement and operate a knowledge ecosystem of interoperable e-Infrastructures. This goal will be achieved by carefully analysing the requirements resulting from this operational scenario – the interoperability issues arising while federating heterogeneous e-Infrastructure each having its own models, technologies and rules – and designing the software solutions and approaches needed to satisfy them. By analysing the identified needs and designing the relative solutions, the work package will coordinate the whole D4Science-II software development activity. It will also be responsible for the release of the Software Development Cooperation Environment (MJRA1.1) through which the whole activity will be organised and monitored.

Description of work

Work package leader: CNR

TJRA1.1 Requirement Analysis

Task leader: CNR; Participants: NKUA, CERN, BDM-USTRATH, TERRADUE;

- collection and analysis of knowledge ecosystem requirements;
- requirements refinement and interpretation;
- design and evaluation of alternative software solutions to identified issues and desiderata;
- configuration and operation of the Software Development Cooperation Environment;
- development and maintenance of a project-wide software development plan;

TJRA1.2 Interoperability Solutions Implementation

Task leader: CNR; Participants: NKUA, CERN, ENG, BDM-USTRATH, FAO, FIN, TERRADUE;

This task is organised in two sets of sub-tasks: one dedicated to the design and implementation of the *gCube Interoperability Framework* (TJRA1.2.6), another dedicated to the exploitation of such an innovative framework to design and implement the software components making each target e-Infrastructure interoperable with gCube (TJRA1.2.1, TJRA1.2.2, TJRA1.2.3, TJRA1.2.4, TJRA1.2.5).

For tasks TJRA1.2.1 to TJRA1.2.5 the following common set of activities is planned even if such activities are focused on the particular e-Infrastructure the task is dedicated to:

- design of approaches and methodologies for making the e-Infrastructure resources and services available to gCube;
- implementation of the envisaged solutions by relying on the gCube Interoperability Framework and its customisation and extension facilities;
- report on experiences, feedback, success stories, bugs and drawbacks on the gCube Interoperability Framework;

TJRA1.2.1 D4Science-GENESI-DR Interoperability Implementation

Task leader: TERRADUE; Participants: N/A;

TJRA1.2.2 D4Science-DRIVER Interoperability Implementation

Task leader: NKUA; Participants: N/A;

TJRA1.2.3 D4Science-INSPIRE Interoperability Implementation

Task leader: CERN; Participants: N/A;

TJRA1.2.4 D4Science-AquaMaps Interoperability Implementation

Task leader: FIN; Participants: N/A;

TJRA1.2.5 D4Science-Other Repository Infrastructures Interoperability Implementation

Task leader: FAO; Participants: TERRADUE;

TJRA1.2.6 gCube Interoperability Framework Implementation

Task leader: CNR; Participants: NKUA, ENG, BDM-USTRATH;

- design of a modular and extensible framework for solving common interoperability issues;
- implementation the mechanisms governing the framework and its openness;
- realisation of reference implementations equipping the framework with off-the-shelf solutions to various and most common interoperability issues;
- analysis of TJRA1.2.1-TJRA1.2.5 requirements and feedback on the released framework;
- analysis and potential consolidation and enhancement of TJRA1.2.1-TJRA1.2.5 software products to enlarge the set of gCube Interoperability Framework reference implementations.

TJRA1.3 D4Science Access Interfaces

Task leader: NKUA; Participants: CNR, ENG, BDM-USTRATH;

- consolidation and enhancement of the gCube ASL, the current implementation of the gCube Access Interface;
- analysis and refinement of gCube Access Interface specific requirements;
- design and implementation of advanced gCube resources consumption facilities;

TJRA1.4 gCube Maintenance and Upgrade

Task leader: CNR; Participants: NKUA, ENG, BDM-USTRATH;

- analysis and refinement of requirements pertinent to the gCube services;
- consolidation and enhancement of gCube services to face with novel needs;
- enhancement of the models and approaches underlying the gCube system to face with the knowledge ecosystem peculiarities;

Deliverables

- *DJRA1.1 Report on Knowledge Ecosystem Supporting Technology Development* (M9, regularly updated) documents the gCube technology by reporting on the overall system architecture, the per service features, the standards and protocols supported and any other technical detail needed to have a comprehensive understanding of the developed knowledge ecosystem enabling technology. The deliverable is implemented through a Wiki page to be easily and promptly adapted to reflect the actual status of the developed system.

Work package number	JRA2			**Start date or starting event:**					M3
Work package title	Scenario Specific Implementation								
Activity Type	RTD								
Participant number	1	2	3	4	**5**	6	7	8	9	10
Participant short name	ERCIM	CNR	NKUA	CERN	**ENG**	BDM-USTRATH	FAO	FIN	4D SOFT	TERRA DUE
Person-months per participant		10	8	24			10	42		6

Objectives

The objective of this work package is to cover the design and implementation case-specific needs for each particular interoperable infrastructure. Each case is analyzed and the system to implement it, on the side of the interoperating infrastructure, is designed in detail. Following the initial design, each case implements the specific components required so that it can exploit the facilities offered by the evolving (via JRA1) ecosystem core and it can export the identified interoperable features to the rest of the ecosystem cases.

Description of work

Work package leader: NKUA

For Tasks TJRA2.1 to TJRA2.5 the following actions apply, although the focus of each corresponds to a different scenario:

- Requirement analysis and detailed design of the software solution;
- Implementation/porting of advanced scenario specific services exploiting the knowledge ecosystem;

Additionally, each task will focus on scenario specific aspects.

TJRA2.1 INSPIRE Specific Implementation

Task leader: CERN; Participants: N/A;

- Complex bibliometric analysis of repository metadata and tuning of hybrid metrics to assess impact of scientific results
- Design of tools in INSPIRE to enable the processing of repository data through the ecosystem infrastructure for the execution of computationally intensive data-analysis tasks

TJRA2.2 DRIVER Specific Implementation

Task leader: NKUA; Participants: N/A;

- Implementation for exploiting advanced services and computational resources offered by the infrastructure, including but not limited to: advanced IR services offered by INSPIRE, ontology handling services, metadata and content brokerage, offered by gCube etc.

TJRA2.3 AquaMaps Specific Implementation

Task leader: FIN; Participants: N/A;

- Enhancement of the mapping modelling tools for exploiting environmental products available in the ecosystem (e.g. GENESI-DR);
- Porting of existing and development of new tools for exploitation of the computational resources of gCube powered infrastructures, in analyzing data sources.

TJRA2.4 FCPPS Specific Implementation

Task leader: FAO; Participants: CNR;

- Development of a dissemination service for knowledge produced by the intermediate steps of the country profiles production workflow.

TJRA2.5 ICIS Specific Implementation

Task leader: FAO; Participants: CNR, FIN, TERRADUE;

- Extending catches re-allocation tools in order to exploit fine-grained Fishing Vessel Monitoring System data and other data available in the ecosystem

Deliverables

- *DJRA2.1 Report on Scenario Specific Technology Development* (M6 regularly updated) describes the software solutions developed to implement the innovative services envisaged in the target scenario to benefit from the ecosystem. It reports on their architecture, the supported features, the standards and protocols supported and any other technical detail needed to have a comprehensive understanding of such technologies. The deliverable is implemented through a Wiki page to be easily and promptly adapted to reflect the actual status of the developed activities.

1.6.3.4 Summary of staff effort

Participant number	Participant short name	JRA1	JRA2	Total person months
1	ERCIM	0	0	0
2	CNR	38	10	48
3	NKUA	37	8	45
4	CERN	14	24	38
5	ENG	12	0	12
6	BDM-USTRATH	25	0	25
7	FAO	4	10	14
8	FIN	20	42	62
9	4D SOFT	0	0	0
10	TERRADUE	7	6	12
Total		**157**	**100**	**257**

1.6.3.5 List of milestones

Milestone number	Milestone name	Work package(s) involved	Expected date	Means of verification
MJRA1.1	Software Development Cooperation Environment	JRA1	M1	The environment supporting both JRA1 and JRA2 activities will be set up. This environment mainly consists in services for supporting collaboration between software engineers and developers involved in the JRA tasks. The web site hosting such services will be available.
MJRA1.2a-h	Knowledge Ecosystem Support Software Release	JRA1	M8, M10, M12, M14, M16, M18, M20, M23	The D4Science-II version control system (SVN) hosts the current version of the expected software components. The list of the expected software component is managed through the D4Science-II issue tracker (TRAC).
MJRA2.1a-h	VRE Specific Resources Implementation Software Release	JRA1	M8, M10, M12, M14, M16, M18, M20, M23	The D4Science-II version control system (SVN) hosts the current version of the expected software components. The list of the expected software component is managed through the D4Science-II issue tracker (TRAC).

1.6.4 Pert diagram

The diagram below depicts the main relationships between the tasks implementing the Joint Research Activities. In particular, it highlights the central role assigned to TJRA1.1 in managing requirements and feedback on the delivered technology and in orchestrating the rest of JRA design and development activities. The diagram shows the organisation of the overall development activity as organized into two main classes of activities – one dedicated to implement the technology needed to operate the knowledge ecosystem and another dedicated to develop innovative services benefiting from their operation in the ecosystem and designed to satisfy the needs arising in one of the target scenarios. The dependency of the second class of activities on the first one, i.e. the dependency of the scenario specific implementation (TJRA2.x) on the ecosystem implementation (TJRA1.x), is expected and reported. Because of the fundamental role played by TJRA1.2 – the task dedicated to implement the technology for making the knowledge ecosystem infrastructures interoperable with D4Science – its organisation in cooperating subtasks is reported. The sub-task dedicated to implement the interoperability framework (TJRA1.2.6) drives the activities called upon to solve the interoperability issue in the context of specific target infrastructures, namely, GENESI-DR (TJRA1.2.1), DRIVER (TJRA1.2.2), INSPIRE (TJRA1.2.3), AquaMaps (TJRA1.2.4) and other repository infrastructures (TJRA1.2.5). In addition to the relations among JRA tasks, the diagram depicts (*i*) how the requirements driving the development of the knowledge ecosystem and the expected VREs flow from the NA and SA – that experiment the developed technology in concrete scenarios – to JRA; and (*ii*) how the JRA produced technology is exploited by SA and JRA – to effectively build the knowledge ecosystem and operate the VREs realising the target scenarios.

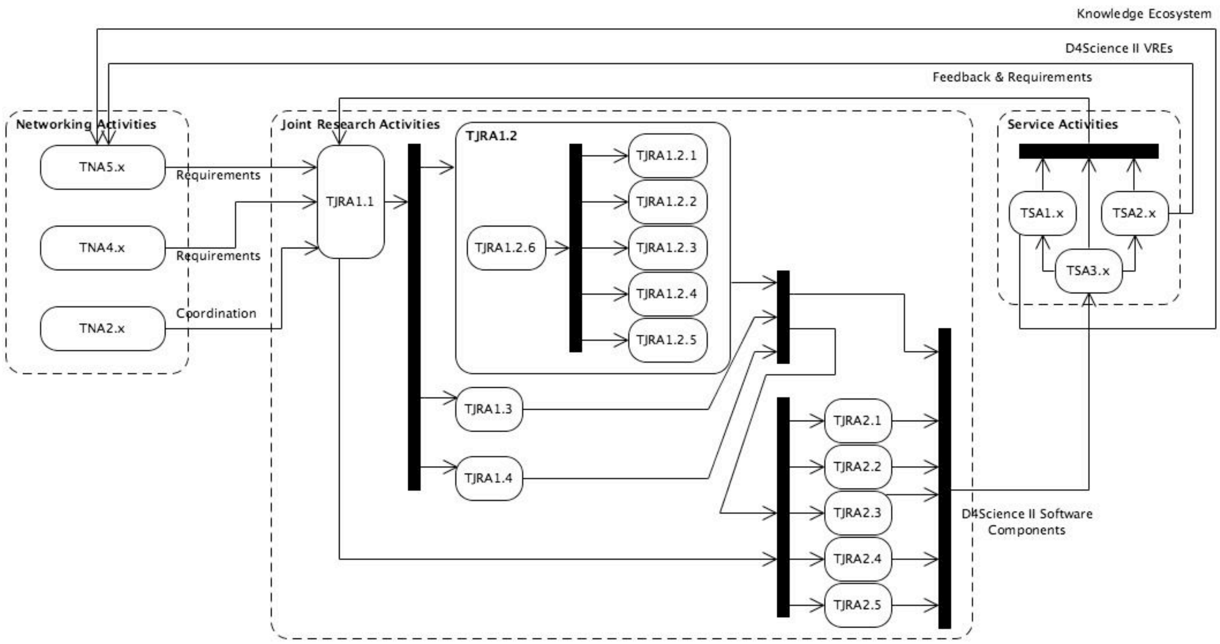


Figure 8. Joint Research Activities Pert Diagram

1.6.5 Risk Analysis and Contingency Plans

A risk breakdown structure for the JRA activities is presented in the following table.

Table 3. Joint Research Activities Risk Analysis and Contingency Plans

Risk	Evaluation and Description[11]	Contingency Plans

[11] Evaluation is expressed through keywords characterising the *provenance* (internal vs external), the *probability* (low, medium, high), and the *impact level* (low impact, medium impact, high impact) respectively.

Requirements lack of focus and stability	*Internal*; *Medium*; *Medium impact* To reach a common understanding on the communities' desiderata among user communities and software engineers is a general issue affecting any software development. Any delay is reaching such a common understanding will result in an unsatisfactory e-Infrastructure and will affect the whole project outcome.	The involvement of the user communities in all the phases concurring to the delivery of the knowledge ecosystem minimise such a risk. Moreover, the used Agile method combining top-down approach, functional design based on scenarios use cases, and the rapid prototyping in which users communities are continuously involved will guarantee that this risk be minimised.
Software not released on time	*Internal*; *Medium*; *High impact* This is a general issue to be taken into account for every JRA activity. Any delay in the release of the software will have a cascade effect on many of the other activities of the Project, starting at the Service Activities. The impact of the effect is variable and depends on several factors, such as the significant of the missing software, the existence of alternative (e.g. previous versions) solutions, etc.	Cross impact w.r.t. other WPs must be carefully identified in the overall planning. Continuous design and development activities monitoring, planning control and review must be performed in order to minimize this risk. Taking into account the status (i.e. new / existing / stable), the implementation complexity and the dependencies of software elements, both contingency and recovery plans will be elaborated in advance for critical elements. The existence of the software used to operate the D4Science production infrastructure and the top-down software design methodology adopted combined with the early prototyping development approach of new software components all contribute to reduce this risk.
Failure to timely develop appropriate technical solutions	*Internal*; *Low*; *High impact* The accomplishment of an appropriate technical solution that meets its requirements is affected by many factors ranging from technological issues to partner's expertises. Problems encountered in such a context will hinder the D4Science-II from fulfilling the requirements.	Mechanisms like an adequate resources allocation plan, tasks assignment, training plan of the staff, hiring of skilled resources, or external acquisition of know-how, will be put in place. The requirements' analysis, the project monitoring and management mechanisms and the internal reviewing procedures, all contribute to reduce the risk towards three directions: the appropriate allocation of resources, the commitment to the activities and the ensuring of results' quality. The expertises gained during DILIGENT and D4Science and the complementary nature of the partners constituting the consortium guarantee that this risk will be minimised.
Delays due to an unexpected complexity in implementing the	*Internal/External*; *Medium*; *High impact* The accomplishment of appropriate	The direct involvement of e-Infrastructure representatives in all the phases leading to the development of

interoperability between a e-Infrastructure and D4Science	technical solution for the specific interoperability issue is affected by many factors ranging from the characteristics of the specific infrastructure to partner's expertises. Problems encountered in such a context will reduce the scope of the delivered knowledge ecosystem.	each interoperability approach guarantees that this risk will be minimised. Implementing generic technological solution on the e-infrastructure sides minimises the risk in general. Confining this implementation to smaller areas - e.g. a subset of the "client" infrastructure's resources is made available through generic approaches, will mitigate the negative impact of the risk. Furthermore, the analysis of requirements and the feasibility of each Interoperability case, as performed by NA activities, will show potential high risks quite ahead of implementation so as to prevent unjustified project effort spending.
Delays due to an unexpected complexity in implementing the interoperability between D4Science and a "client" Infrastructure	*Internal/External*; *Medium*; *High impact* The accomplishment of appropriate technical solution for opening D4Science facilities to other infrastructures is affected by many factors ranging from the effectiveness of the Access Interface technology (TJRA1.2.3) to client infrastructure team expertise. Problems encountered in such a context will reduce the exploitation domain of the delivered knowledge ecosystem.	Adopting generic technological approach for the Access Interface minimizes the risk that its exploitation raises roadblocks to its clients. Even in the case where generic approaches might not reach the full potential of the backbone infrastructure, the negative impact of the risk is mitigated, e.g. a subset of the knowledge ecosystem resources/facilities will be made available through generic approaches. The deep knowledge of the involved RTD partners on the gCube system and the underlying technologies and surrounding specifications are guarantees for minimization of this class of risks.
Delays due to an unexpected complexity of developing scenario specific services	*Internal*; *Medium*; *High impact* In order to deploy VREs benefitting from the knowledge ecosystem, innovative services have to be implemented. Problems in accomplishing such an activity may result in a reduced capability of the delivered VREs to satisfy the concrete needs.	Mechanisms like an adequate resources allocation plan, tasks assignment, training plan of the staff, hiring of skilled resources, or external acquisition of know-how, will be put in place. However, strong community commitment and involvement in the phases of scenario design (NA5) and scenario specific technology development (JRA2) guarantee this risk be minimised.

Section 2: Implementation

2.1 Management structure and procedures

2.1.1 Governance vs. Management

The D4Science-II management structure distinguishes between the Governance and the Management of the project. Governance is administering to the project and exercising authority over the management decisions that are made therein.[12] Management operates within a hierarchy of delegated responsibility, with all management functions providing support to project activities and for the accomplishment of the goals of the consortium.

The D4Science-II management structure has been defined with distinctions between governance issues and management issues kept in mind, and will therefore have the following boards: Members General Assembly, Project Management Board, External Advisory Board and Technical Board. The **Members General Assembly** is the project's *governing board*, responsible for decision-making affecting consortium composition, resource allocation, implementation of the work plan, among other decisions having a direct legal or financial impact on project beneficiaries. The remaining three are *managerial boards*, each with a different set of responsibilities aiming at the accomplishment of common objectives. The **Project Management Board** is responsible for decision-making affecting project strategy, including risk management. A highly distinguished **External Advisory Board** will serve to provide valuable direction to the Project Management Board in the form of recommendations. In a complementary fashion, project operations will be represented by the **Technical Board**. The functions and members of the boards will be detailed in the sections to follow. Please find below a diagram depicting the interaction between Project Governance and Project Management as represented by the various boards.

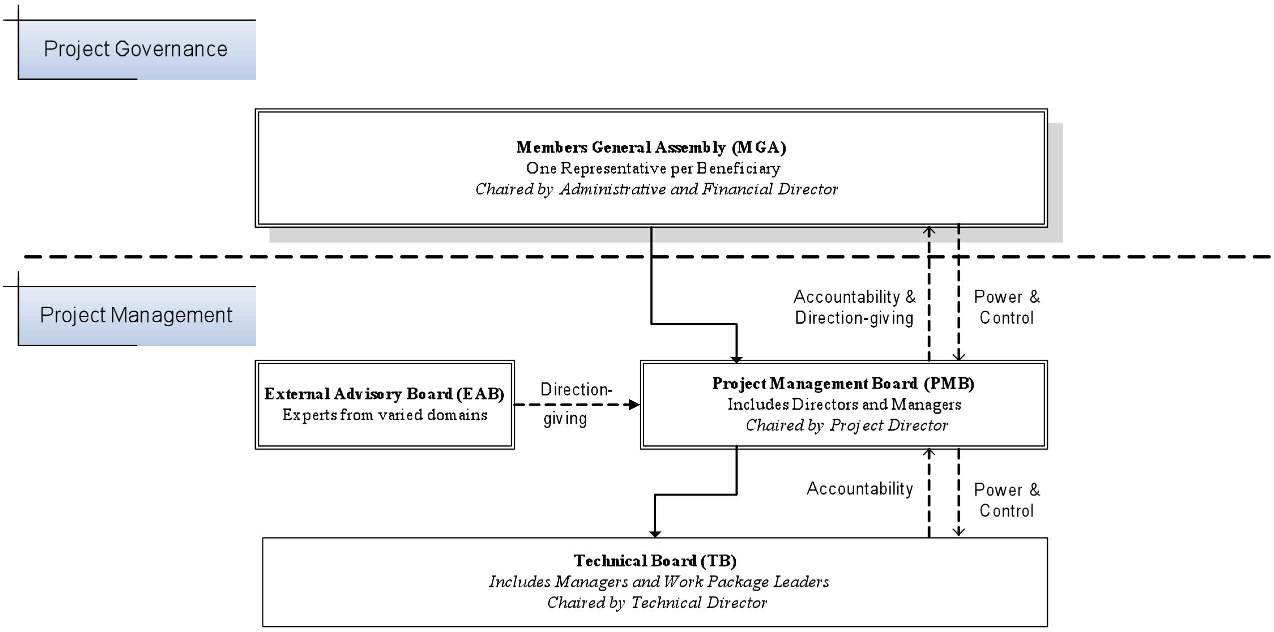


Figure 9. D4Science-II Governance vs. Management

[12] The "Consortium Agreement" that will be signed by all project beneficiaries prior to the launch of the project is primarily concerned with project governance mechanisms.

2.1.2 Project Governance

2.1.2.1 Members General Assembly (MGA)

The project's governing board will be named the **Members General Assembly (MGA).** The Members General Assembly is responsible for decision-making affecting consortium composition, resource allocation, implementation of the work plan, among other decisions having a direct legal or financial impact on consortium members. The MGA will ensure that the EU contract is properly executed and that the terms of the agreed Consortium Agreement are properly implemented. It will cover all aspects of the relations between partners, their responsibilities, liabilities, ownership of IPR, licensing, and exploitation issues. These types of governance decisions will require *two-thirds* of the votes, with *one representative per partner voting on behalf of his or her organization*.

The MGA encompasses one representative per partner and is chaired by the Administrative and Financial Director. The members of the board will meet in person at least *once per year*, with telephone conferences convened as relevant issues arise. Electronic voting outside the annual meetings of the MGA will be authorised.

2.1.2.2 Project Coordination

Project coordination will be shared between the **Administrative and Financial Director (i.e., Coordinator)** and a **Project Director**. The former directs project management and administration, while chairing the project's governing board (i.e., MGA) to hold the consortium accountable, financially and ethically, to the European Community for the use of funds to achieve the project objectives and provide a return on investment. The latter coordinates the overall scientific and technical activities of the project, leading inter- and intra-project interactions and chairing the project's strategic management board (i.e., Project Management Board). The complementarities of the roles allow each person to focus on achievement of the project work plan from different perspectives. The Administrative and Financial Director, Ms. Jessica Michel, and the Project Director, Dr. Donatella Castelli, have successfully employed this mechanism for over four years, working together as the management team of the DILIGENT and D4Science projects.

Overall project coordination is therefore divided between two work packages:

- NA1 Project Management, dedicated to overall project management and administration, including EU reporting and consortium management; and
- NA2 Scientific and Technical Coordination, supervising the project across all activities by coordinating interactions, monitoring the time schedule and recommending appropriate actions.

This clear separation between project *management* (NA1) and *coordination* (NA2) provides a structure in which the coordinators can provide: (*i*) efficient administrative and financial management; and (*ii*) thorough multi-disciplinary scientific and technical coordination. Furthermore, project coordination activities will be supported by a **Project Office**, including project assistance in both Sophia Antipolis (FR) and Pisa (IT). The Project Office includes public relations experts, experienced project managers to serve as consultants as needed, and project assistants working to ensure the timely delivery of financial statements and the organisation of events such as project meetings, workshops, review meetings, etc.

As Work Package Leaders and Chairs of the project's main decision making boards, the Administrative and Financial Director and the Project Director are the only two people within the D4Science organization to have cross-over responsibilities between project management and project governance.

The **Administrative and Financial Director** (**AFD**) is the recognized project Coordinator and serves as the official contact point for the European Commission. The AFD directs the administrative and financial management across work packages and reporting across partners. Legal issues such as the handling of Intellectual Property Rights and the maintenance of the consortium agreement fall into the realm of the AFD and her team. The AFD will manage from the Project Office in Sophia Antipolis where a dedicated staff will see to administrative and financial operations and provide general project support. The AFD leads NA1 and is represented by GEIE ERCIM.

The **Project Director** (**PD**) will lead the scientific and technical coordination of the project. The PD supervises the project across all activities and is responsible for creating the conditions necessary for successful and effective collaboration of the D4Science II team. The PD's high-level view permits her to also serve as the ambassador of the project, establishing meaningful cooperation with other projects and initiatives at the national and international level. The PD is also responsible for the monitoring of the time schedule and the timing of related activities in close collaboration with the Technical Director (functions described below). The PD leads NA2 and is represented by CNR-ISTI in Pisa.

2.1.3 Project Management

D4Science-II project managers are equipped with the necessary quantitative and analytical skills, tools, knowledge and decision-making capabilities to fulfil the functional responsibilities required for the achievement of the project work plan. To assist them in accomplishing this goal, separate boards for strategy and operations will be established. The Project Management Board will formulate and lead the implementation of the overarching D4Science strategy, including the creation of synergies with other initiatives and long-term sustainability issues on which an External Advisory Board will provide guidance. The Technical Board will lead the diverse technological activities encompassing the development and implementation of the D4Science e-Infrastructure, and consequently, the D4Science-II ecosystem. The following organisational chart represents the relationships between the project's strategic and operations managers; the role and composition of the managerial boards will be detailed below.

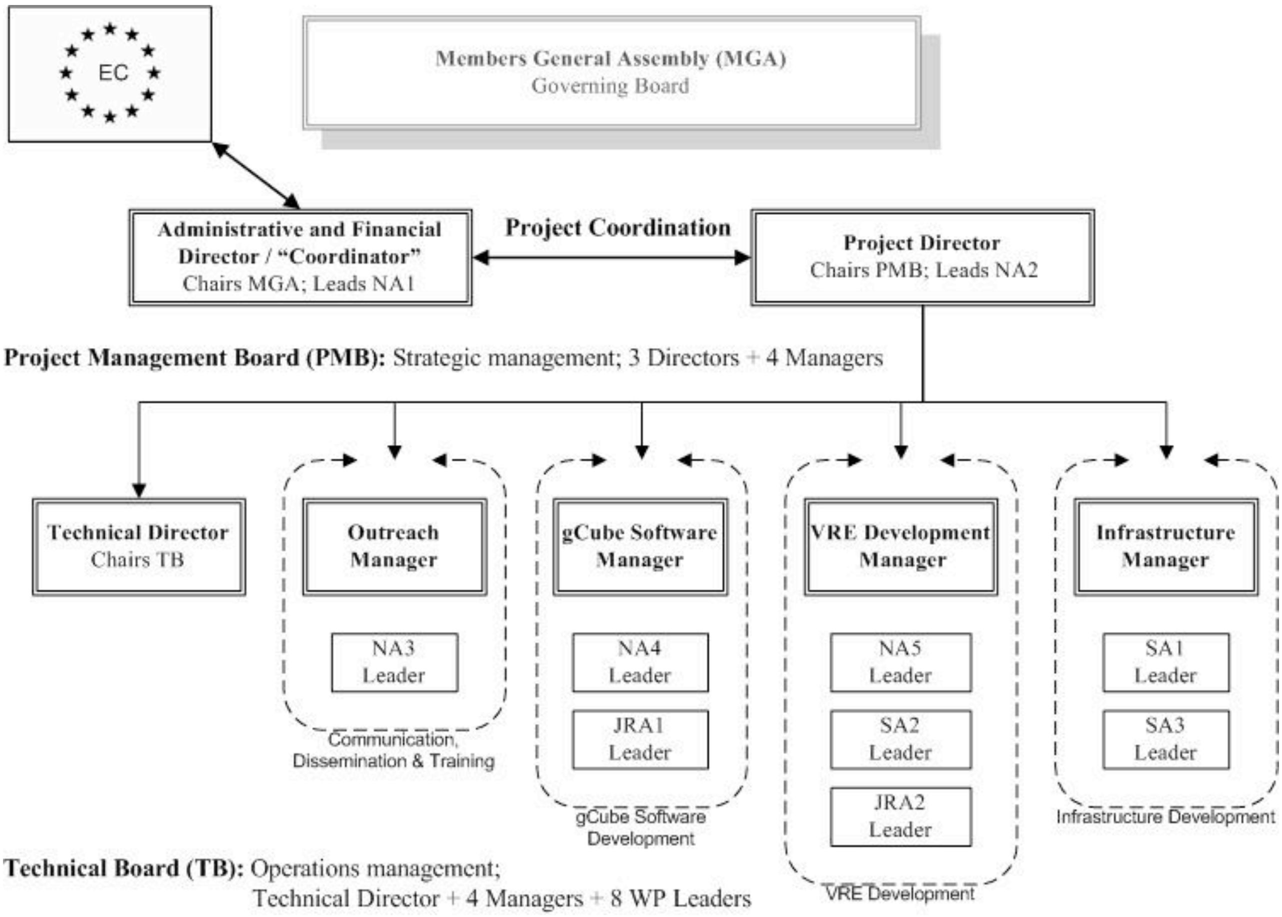


Figure 10. D4Science-II Organisational Chart

2.1.3.1 Strategic management

D4Science-II strategic management will be concerned with the creation of a sustainable and interoperable e-Infrastructure ecosystem to serve a large set of communities. To deal with the complex and multidisciplinary challenges presented by the communities and the technology, D4Science-II strategic management will encompass two boards. The first, the Project Management Board, must cover all of the projects' activity areas, and therefore includes the Coordinators, Managers who are responsible for well-defined "Development Areas", and a Technical Director whose responsibility is to coordinate cross-development integration activities in the creation of an ecosystem. The second board, the External Advisory Board, will provide support in strategy formulation and sustainability from experts with world-wide recognition for their talent.

2.1.3.1.1 Project Management Board (PMB)

The **Project Management Board** (**PMB**) is the supervisory body of the project. It is designed to promote continuous sharing of project knowledge across all areas of activity and according to the defined work plan. In this way, the PMB can make informed decisions affecting the project strategy, while proposing and rapidly implementing corrective measures concerning the work plan in the emergence of delays or deviations. Monitoring risk assessment and partner performance is an essential objective of the PMB and this will be performed through collaboration with the Quality Assurance Task Force (defined in Section 2.1.3.2.2). The PMB will initiate action if a partner's performance is deemed to be below par according to the tasks defined in the Description of Work.

The PMB is chaired by the Project Director for the duration of the project and includes all "Directors" and "Managers": the Administrative and Financial Director, Technical Director, Outreach Manager, gCube Software Manager, VRE Development Manager and Infrastructure Manager.

The Outreach Manager, gCube Software Manager, VRE Development Manager and Infrastructure Manager are responsible for each semantically meaningful development activity in which the project is organised. The operation and development of the D4Science-II knowledge ecosystem is dependant upon their communication and cooperation. The four managers have these general responsibilities:

- To be informed on the status of the work packages within his/her development area;
- To support the Technical Director in coordinating relationships across development areas;
- To advise the PMB on strategic decisions related to his/her activity area;
- To contribute to the tasks of the Technical Board (cf. Section 2.1.3.2)

Specifically, the **Outreach Manager** devises D4Science-II outreach strategy, while coordinating and supervising the work performed in the fields of Communication, Dissemination and Training (NA3). Training activities include tutorials for software engineers, administrative personnel and end users, from within the project's user communities as well as from other scientific communities. The role of Outreach Manager is assigned to Johannes Keizer of FAO.

The **gCube Software Manager** is responsible for gathering the requirements of the D4Science-II knowledge ecosystem (NA4), and the subsequent implementation of interoperable solutions (JRA1). The gCube Software Manager must be able to promote open exchange of resources among the knowledge ecosystem constituents, namely D4Science and the other main application-level infrastructures federated within, e.g., GENESI-DR, DRIVER, INSPIRE, AquaMaps. He must investigate and promote the usage of standards supporting several parts of the gCube system, from low-level communication protocols and encoding schemes, to data models to classification systems. Furthermore, the gCube Software Manager will contribute to the implementation of the project quality procedures by monitoring the related quality indicators and ensuring an appropriate management of related risks. The role of gCube Software Manager is assigned to George Kakaletris of NKUA.

The **VRE Development Manager** will coordinate the tasks undertaken in NA5 (Knowledge Ecosystem Usage: Requirements and Validation), SA2 (VOs and VREs Definition and Operation), and JRA2 (Scenario Specific Implementation). These activities include the involvement of the D4Science-II target communities in an iterative process for VRE configuration, testing and providing feedback; supporting and maintaining the resources for providing the VREs of the target scenarios; and providing the design and implementation of the case-specific needs for each scenario. The VRE Development Manager is that of an overall facilitator between the various communities, the developing technology and the infrastructure. Furthermore, the VRE Development Manager will contribute to the implementation of the project quality procedures by monitoring the related quality indicators and guaranteeing an appropriate management of related risks. The role of VRE Development Manager is assigned to Leonardo Candela of CNR.

The **Infrastructure Manager** will have a strong coordination role in the operation of the infrastructure by overseeing the planning of the deployment of the infrastructure and the definition and enhancement of the procedures needed to ensure its proper functioning. The work package Knowledge Ecosystem Operation activity (SA1) is closely related to the project's continuous process of software integration and release (SA3 Software Integration, Testing and Distribution) which releases the software packages delivered in the infrastructure by SA1. The Infrastructure Manager will coordinate cross-work packages relations and issues resolution and report to the PMB on the status of these activities. Furthermore, the Infrastructure Manager will contribute to the implementation of the project quality procedures monitoring the related quality indicators and ensuring an appropriate management of related risks. The role of Infrastructure Manager is assigned to Roberta Faggian Marque of CERN.

In addition to the four managerial functions described above, there will be a third "Director" present on the Project Management Board: the Technical Director. The **Technical Director** will coordinate technical aspects across all activities (i.e., software development, VRE development, infrastructure development and outreach where relevant), surveying to ensure that the project managers provides the proper level of support to all work packages. The engagement of the Technical Director in multiple heterogeneous technical activities makes him the best person to provide detailed resource allocation and scheduling, while monitoring the time schedule and the timing of related activities. Therefore, the Technical Director becomes a close advisor to the Project Director who employs a high-level approach to the supervision of the project across activities. The role of Technical Director is assigned to Pasquale Pagano of CNR-ISTI.

The PMB will strive to achieve consensus on issues concerning project strategy and the implementation of corrective measures. However, when deadlock occurs, votes will require a *two-thirds majority.* The PMB will meet in person at least *two times per year*, with telephone conferences once per month as a minimum. Electronic voting outside the bi-annual meetings of the PMB will be authorised.

Finally, this Board will apply recommendations from the External Advisory Board in order to improve or reorient project strategy.

2.1.3.1.2 External Advisory Board

The **External Advisory Board (EAB)** is a panel of experts from various scientific fields advising on project strategy and long-term sustainability issues. The D4Science-II EAB will be comprised of seven specialists from the following domains: digital libraries, grid, fisheries and aquaculture resources management, data analysis and visualisation, and data management in the context of biodiversity.

The following highly acclaimed scientists have already expressed their agreement to serve on the External Advisory Board of D4Science-II and have allowed us to publish their brief biographies below:

Christine Borgman is Professor & Presidential Chair in Information Studies at the University of California, Los Angeles (UCLA). She is the author of more than 150 publications in the fields of information studies, computer science, and communication. Prof. Borgman's research interests and teaching areas include information science, information retrieval, electronic publishing, information-seeking behavior, scientific data use and policy, scholarly communication, bibliometrics, and information technology policy. Her book, *From Gutenberg to the Global Information Infrastructure: Access to Information in a Networked World* (MIT Press, 2000), won the Best Information Science Book of the Year Award from the American Society for Information Science and Technology.

Fabrizio Gagliardi has over 30 years of experience in computing applied to physics. His professional experience includes international computing projects at CERN, in the CERN member states and in USA. From 1996 to 1999 he led the Data Management services of the Information Technology division, and from 1998 to 2000 he was in charge of the CERN participation to the EU project Eurostore. In July 1999 he was appointed responsible for the IT divisional industrial technology transfer programme. From January 2001 to March 2004, he led the EU DataGrid project. From April 2004 through October 2005, he served as the Project Director of the EGEE project. He is founding member of the International Advisory Committee of the Global Grid Forum and of other international and European committees. Fabrizio Gagliardi took leave from CERN on November 1st, 2005 to join Microsoft Corporation with the title of EMEA Director for the Technical Computing Initiative.

Erwin Laure is the Technical Director of the EU funded project "Enabling Grids for E-Science in Europe (EGEE)" working at the European Organization for Nuclear Research (CERN). After joining CERN in 2002 he worked on data management issues within the EU DataGrid (EDG) project, became the Technical Coordinator of EDG, and coordinated the middleware re-engineering activities in the first phase of EGEE. He holds a PhD in Business Administration and Computer Science from the University of Vienna, Austria. His research interests include grid computing with a focus on data management in grid environments as well as programming environments, languages, compilers and runtime systems for parallel and distributed computing. As of November 2008 he will be leading PDC-HPC, the supercomputing center at the Royal Technical University at Stockholm, Sweden.

Serge Michel Garcia is French and holds a Doctorate in Sciences (ScD) of the University of Aix-Marseille (France, 1976). Initially specialized in shrimp population dynamics and tropical fisheries management, he spent a decade working in Africa, on fisheries assessment and 3 decades working in FAO as an assessment and management specialist. He headed the national Oceanographic Research Centre (CRODT) of Dakar-Thiaroye (Senegal, 1976-79), the FAO Marine Resources Service (1984-90 and the FAO Fisheries Resources Division (then Fishery Management Division) before retiring in 2007. During these last two assignments, he has been monitoring the state of world fishery resources, tirelessly warning about un-sustainability of most current fishing regimes, promoting *inter alia* the use of open information systems, sustainability indicators, the precautionary and the ecosystem approach to fisheries and a more systemic approach to fisheries assessment and management. Among other responsibilities, he is a member of the Scientific Steering Committee of the CORE Census of Marine Life project (CoML) and member of the Governance Board of the Ocean Biodiversity Information System (OBIS).

Lucille Nowell is on assignment as a Program Director in the area of Data, Data Analysis and Visualization for the Office of Cyberinfrastructure at National Science Foundation (NSF), Lucy Nowell is a Chief Scientist from the Information Analytics group at Pacific Northwest National Laboratory (PNNL). She is also an alumna of the Virginia Tech Digital Libraries Research Laboratory, where she designed one of the first information visualization user interfaces for the Envision Project. Her research interests include long-term data preservation/archiving, user interaction with information in the context of massive data, usability engineering for information exploitation systems and digital electronic libraries, cognitive issues in user interface design, information visualization, intelligent user modelling and intelligent user interfaces, and information storage and retrieval. As Program Director for Data, Data Analysis and Visualization in NSF's Office of Cyberinfrastructure, her program responsibilities include: Sustainable Digital Data Preservation and Access Network Partners (DataNet); Community-based Data Interoperability Networks (INTEROP); Software Development for Cyberinfrastructure (SDCI) and, Strategic Technologies for Cyberinfrastructure (STCI).

Per Öster is the newly appointed Director of Application Service at Scientific Computing, Ltd. (CSC) in Espoo, Finland. Per Öster previously worked at the Center for Parallel Computers at KTH in Sweden and is internationally recognized as one of the Nordic region's top specialists in grids. He was actively involved with the EGEE II initiative funded by the European Commission as a member of the project's management board. Now, CSC is serving as one of the nine leading organizations in the EGI Design Study, with CSC's area of responsibility being communications and external relations. EGI is meant to continue from where the EGEE III project ends.

Edward Vanden Berghe is currently the Executive Director of the Ocean Biogeographic Information System (OBIS), based at Rutgers University, New Jersey. Before joining OBIS, he was the manager of the Flanders Marine Data- and Information Centre, based at the Flanders Marine Institute in Oostende, Belgium. He has been involved in several data integration exercises, and through these activities came to appreciate the important role of metadata, and of data repositories. He is member of several working groups on data and information management in the framework of the Intergovernmental Oceanographic Commission of UNESCO and of the International Council for the Exploration of the Seas.

2.1.3.2 Operations Management

D4Science-II operations management is primarily concerned with the technical objectives of the work plan, centred on providing scientists with Virtual Research Environments for unlimited generation and dissemination of scientific and technical knowledge. D4Science-II will transform the current operational D4Science e-Infrastructure into the lynchpin of an e-Infrastructure ecosystem, holding together and mediating between all other e-Infrastructures that may want to join the ecosystem.

The people responsible for the ensuring that the project's technical objectives are undertaken in an efficient and effective manner, while managing resources and services, are the Work Package Leaders, supported strongly by the members of the Project Management Board. Sub-project committees are also deemed necessary to support the Work Package Leaders in the undertaking of a important tasks that bridge several work packages or touch the entire project.

2.1.3.2.1 Technical Board (TB)

The operational project management is represented by the **Technical Board (TB)** and will be led by the Work Package Leaders; the Outreach, gCube Software, VRE Development, and Infrastructure Manager; and the Technical Director on a day-to-day basis. In addition to leading the numerous technological activities encompassed by the development and implementation of the D4Science e-Infrastructure (as centrepiece of the D4Science-II ecosystem), the TB members must be prepared to report frequently on the status of adherence to the established work plan.

Work Package Leaders will report to the appropriate development manager as indicated in Section 2.1.3.1.1, but are the most important contributors to the TB and will carry an equal voice. The main responsibilities of the Work Package Leaders are as follows:

- To be informed on the status of the tasks within his/her work package, and provide input on a quarterly basis to the Quality Assurance Task Force for the monitoring of project progress;
- To support the Outreach/gCube Software/VRE Development/Infrastructure Manager in coordinating cross-work package relationships within the appropriate activity area;
- To contribute to the tasks of the Technical Board.

The Technical Director will chair the TB, positioning him to recommend appropriate actions to the Project Management Board in the face of delay. The TB will meet in person *four times per year*, with telephone conferences held every two weeks. Votes will require *two-thirds majority*. Electronic voting outside the quarterly meeting of the TB will be authorised.

2.1.3.2.2 Sub-Project Committees

The project's managerial boards (i.e., PMB and TB) will have the possibility of proposing to the Members General Assembly the formation of sub-project committees as needs are identified. Sub-project committees are intended to assist in the efficient undertaking of the work plan by bringing together representatives from each consortium member on particular tasks, providing them with a forum for discussion and debate, with the mission of communicating the outcomes to the relevant audience. When approved, sub-project committees become informal bodies that define their own distinct rules for collaboration. At the present, the formation of the following D4Science-II sub-project committees is envisaged:

Quality Assurance Task Force (QATF): "Quality assurance" is a dedicated task under NA1 Project Management. The objective of the task is to ensure the monitoring and assessment quality indicators adopted by D4Science-II. In order to accomplish this task, a number of rigorous procedures have been defined within the context of the I3 D4Science and these same procedures will be exercised within the D4Science-II organisational structure: project reporting (e.g., deviations from the work plan, resources spent, deliverable quality), review preparation and post-review follow-up, activity-specific process, document management, etc., as well as the tools and metrics that will be applied to measure achievements. The Quality Assurance Task Force will obtain input from the Work Package Leaders in order to measure the project's progresses across development areas (i.e., outreach, gCube software, VRE development, infrastructure). An additional objective of the Quality Assurance task (TNA1.2) is to perform risk identification and analysis impacting the global I3 project. The leadership of the Quality Assurance Task Force is assigned to CERN and will include the participation of CNR and NKUA.

Technical Committee (T-Com): The Technical Director will lead cross-work package meetings for collaboration and alignment issues related to implementation of the knowledge ecosystem. These meetings between the participants of the project's technical work packages will serve to discuss about progress of work and to brainstorm, and to address integration issues by the use of a working group methodology. The T-Com should be seen as an extension of the Technical Board, and these meetings will be held in parallel with the quarterly meetings of the TB.

Technical Management Team: The Technical Management Team is simply comprised of the Technical Director, gCube Software Manager, VRE Development Manager and Infrastructure Manager, i.e., the people primarily responsible for the project's Service and Joint Research Activities. It is expected that these four individuals will meet by telephone on a weekly basis, promoting a very close collaboration between them and facilitating efficient knowledge exchange at all project levels. The Technical Management Team may invite any person to attend a weekly discussion, depending upon the topic(s) at hand.

Experience gained from the management of the DILIGENT and D4Science projects leads us to recognize the increased efficiency that these sub-project committees will bring to the realisation of a D4Science-II e-Infrastructure ecosystem, while the various boards ensure democracy and accountability at a global project level.

2.1.4 Conflict Resolution

Effective conflict resolution begins with understanding that the project managers (from Project Director to Work Package Leader) operate within a hierarchy of delegated responsibility and authority. As a general rule, project management will aim towards a goal of consensus building, promoting mediation over voting in order to ensure a maximum degree of cooperation within the consortium.

As a reminder, the Administrative and Financial Director and the Project Director are two members of the seven-person Project Management Board. The PMB includes all project Directors (3) and Managers (4), whereas the Members General Assembly (MGA) comprises one representative per partner.

When a conflict arises between partners, the following escalation process will be followed:

1. One of the disputing parties sends an email explaining the situation to the relevant Work Package Leader with the disputing party in copy (allowing the other party to reply if so desired).

   Action and outcome:

   + Work Package Leader successfully resolves conflict,

   - if no resolution achieved, go to step 2.
2. Work Package Leader sends an email to appropriate Development Manager (i.e., Outreach/gCube Software/VRE Development/Infrastructure) explaining the situation and the efforts already made toward conflict resolution, with both of the disputing parties in copy.

   Action and outcome:

   + Development Manager successfully resolves conflict through additional knowledge of the project, or via a teleconference discussion with the disputing parties,

   - if no resolution achieved, go to step 3.
3. Development Manager raises the issue with the PMB via a teleconference.

   Action and outcome:

   + If the conflict does not have a direct legal or financial impact on one of the partners, the PMB builds consensus or opts for a two-thirds majority vote if no consensus achieved.

   - if the conflict has a direct legal or financial impact on one of the partners, project management must formulate an action plan for resolution to be presented to the MGA and proceed to step 4.
4. Administrative and Financial Director communicates the situation to the MGA with an email summarising: the conflict; partners involved/impacted; actions already taken; and the action plan proposed by project management.

   Action and outcome:

   + The disputing parties accept the proposal for resolution,

   - The MGA votes to accept the proposed resolution which must be accepted by the disputing parties (by two-thirds majority vote; the partners directly involved/impacted cannot exercise veto rights).

2.2 Individual participants

2.2.1 GEIE ERCIM (ERCIM)

The European Research Consortium for Informatics and Mathematics will ensure the overall coordination of the project. ERCIM is a European Economic Interest Grouping (EEIG), comprised of a network of research institutes from twenty European countries, embodying more than 12,000 researchers and engineers. ERCIM's mission is to: foster collaborative work within the European research community in Information and Communication Technologies (ICT) and Applied Mathematics; advise the European Commission and national governments; and increase co-operation with European industry. As a consortium of members from many European countries, ERCIM creates a balance between European diversity and necessary homogeneity by building bridges between different cultures and facilitating the movement of technical ideas within academia and across borders.

GEIE ERCIM is the European host of the **World Wide Web Consortium (W3C)**. W3C is an international consortium where Member organizations, a full-time staff, and the public work together to develop Web standards. W3C's mission is to lead the World Wide Web to its full potential by developing protocols and guidelines that ensure long-term growth for the Web.

W3C primarily pursues its mission through the creation of Web standards and guidelines. W3C also engages in education and outreach, develops software, and serves as an open forum for discussion about the Web. In order for the Web to reach its full potential, the most fundamental Web technologies must be compatible with one another and allow any hardware and software used to access the Web to work together. W3C refers to this goal as "Web interoperability." By publishing open (non-proprietary) standards for Web languages and protocols, W3C seeks to avoid market fragmentation and thus Web fragmentation.

Through ERCIM, D4Science-II Coordinator, W3C will work with the consortium to identify methods for the open exchange of knowledge in the context of the work package NA4 dedicated to *Knowledge Ecosystem Building: Requirements and Solutions*.

The central ERCIM Office in Sophia Antipolis, France will coordinate the project. Since 1990, the ERCIM Office team has developed a proven expertise in managing research projects and large networks, and has been involved in over 30 projects funded by various European Community programmes. Community-funded projects under the Seventh Framework Programme include the Combination of Collaborative Project and Coordination and Support Action (I3) D4Science (No. 212488); and Coordination and Support Actions Digital World Forum (No. 216513) and MobiWeb2.0 (No. 212430). ERCIM is leading the work package for "Spreading Excellence" in the FP7 Network of Excellence VPH (No. 223920). Sixth Framework Programme projects funded by the EC and managed by ERCIM Office include the Specific Support Action GRID@ASIA (No.599413); Coordination Actions Beyond-The-Horizon (No.00662), Interlink (No.034051) and EchoGrid (No.045520); Specific Targeted Research Projects GridComp (No. 034442) and Net-WMS (No. 034691); Integrated Projects DILIGENT (No. 004260), ACGT (No. 026996), PALETTE (No. 028038) and VITALIS (No. 045389); and Networks of Excellence DELOS (No. 507618), MUSCLE (Contract No. 507752) and CoreGRID (No. 004265).

Jessica Michel will serve as the project **Coordinator**, hereby referred to as the **Administrative and Financial Director (AFD)**. Ms. Michel joined ERCIM in March 2004. Her project management experience includes the administrative and financial coordination of the D4Science I3; the World Wide Web Consortium (W3C) projects WAI-AGE, Digital World Forum and MobiWeb2.0; and a number of other EC-funded initiatives. Ms. Michel received a Master in Business Administration (MBA) from Solvay Business School of the Université Libre de Bruxelles (ULB), Belgium in 2001.

Philippe Rohou joined ERCIM in August 2006 and was named Manager of the ERCIM Projects group in May 2008. Mr. Rohou will provide overall administrative and financial support on the D4Science-II project, while shadowing Ms. Michel as Deputy Coordinator. His European project management expertise is derived from the coordination of the DELOS and CoreGRID Networks of Excellence, and as the current leader of "Spreading Excellence" for the VPH NoE. Mr. Rohou obtained his MSc in *Mathematics and Statistics* at Paris University and later spent 14 years with Digital Equipment Corporation managing

[13] The notation "No." indicates EC Contract Number.

projects, programmes, consulting groups and the company's European customer briefing centre. The next 5 years led him to manage a small events company, culminating with the creation of his own enterprise in 2004 offering corporate events and golf circuits on the French Riviera.

2.2.2 Consiglio Nazionale delle Ricerche, Institue of Information Science and Technologies – Pisa (CNR)

The **Institute of Information Science and Technologies (ISTI)** of the **Italian National Research Council (CNR)**, which is organised in 16 laboratories, is committed to producing scientific excellence and playing an active role in technology transfer. The team participating in this proposal belongs to the 'Multimedia Networked Information System Laboratory', which consists of 48 researchers and technicians conducting research and development activities on algorithms, techniques and methods for information modelling, access and handling, as well as new architectures and system services (P2P and Grid-based) supporting large networked multimedia information systems.

In the context of the project, CNR handles a number of key tasks including: *(a)* the overall scientific and technological coordination of the project activities; *(b)* the development of interoperability approaches toward the Ecosystem implementation; *(c)* the coordination of the activities leading to the definition and operation of the VREs; *(d)* the coordination of the development of the Ecosystem enabling technology, namely *gCube*; *(e)* the design, implementation, and management of the application framework – gCore – upon against which all other gCube services are implanted; *(f)* the design and implementation of two key functional areas of gCube, i.e. the Information System and the VRE Management, for which it has been responsible in the Diligent and D4Science projects.

The CNR-ISTI team has been involved in many EU-funded projects relevant to the topics addressed in this proposal, namely in the following FP6 projects: DELOS II NoE (No. 507618, Scientific Coordinator), DILIGENT (No. 004260, Scientific Coordinator), MultiMatch (No. 033104, Coordinator), BRICKS (No. 507457), BELIEF (No. 026500), CASPAR (No. 033572), DRIVER (No. 034047), SAPIR (No. 45128). In particular, in the DELOS NoE it has been responsible of the task dedicated to the production of the DELOS Digital Library Reference Model. It is currently involved in the 7th FP projects: DRIVER II (No. 212147), D4Science (No. 212488), TrebleClef (No. 215231) and BELIEF II (No. 223759). It is also the co-ordinator of the DL.org (No. 231551), a Coordination Action, currently under negotiation phase, whose main objective is stimulating the adoption of interoperability solutions among large digital libraries.

Donatella Castelli (donatella.castelli@isti.cnr.it) will act as Project Director of this I3 (cf. Section 2.1). She is a senior researcher at CNR-ISTI. She graduated in Computer Science at the University of Pisa and since 1987 she has worked at CNR-ISTI. Since then she has participated actively in several EU and nationally funded projects on Digital Libraries and Research Infrastructures. Among them are: ERCIM Digital Library-DELOS, DELOS, ECHO, SCHOLNET (Scientific Coordinator), CYCLADES, Open Archives Forum, and D-Lib Competence Center, ***DILIGENT*** (***Scientific Coordinator***), DRIVER and BELIEF. In the framework of the DELOS FP6 NoE she has led the activity dedicated to the production of the DELOS Reference Model for Digital Libraries. She is currently the ***scientific coordinator*** of the ***D4Science*** project and she is also involved in the DRIVER II and BELIEF II projects. Her research interests include digital libraries content and architecture modelling and interoperability.

Pasquale Pagano (pasquale.pagano@isti.cnr.it) will act as Technical Director of this I3 (cf. Section 2.1). Dr Pagano has a strong background on digital library distributed architectures. He has been one of the early developers of the ERCIM Technical Reference Digital Library (ETRDL) and has participated in the design of the most relevant DL systems developed by CNR, namely he has lead the design and development activity of the FP5 project SCHOLNET, designed the Virtual Library component in the FP5 project CYCLADES and participated in the design of the FP6 project DRIVER. Dr Pagano has served the ***DILIGENT*** project as ***Technical Support Manager***. He is currently the ***Technical Director*** of the ***D4Science*** project and he is also involved in the DRIVER II and BELIEF II projects.

Leonardo Candela (leonardo.candela@isti.cnr.it) will act as VREs Development Manager for this I3 (cf. Section 2.1). Dr Candela is a researcher at Networked Multimedia Information Systems (NMIS) Laboratory of the Institute of Information Science and Technologies - Italian National Research Council (ISTI - CNR). He graduated in Computer Science in 2001 at University of Pisa and completed a PhD in Information

Engineering in 2006 at University of Pisa. In 2001 he joined the NMIS Laboratory and was involved in the CYCLADES, Open Archives Forum, DELOS, ***DILIGENT***, DRIVER and ***D4Science*** projects. He is member of the OAI-ORE Liaison Group. His research interests include Digital Library [Management] Systems and Architectures, Digital Libraries Models, Distributed Information Retrieval, and Grid Computing.

2.2.3 National and Kapodistrian University of Athens, Department of Informatics and Telecommunications – Hellas (NKUA)

The National and Kapodistrian University of Athens (NKUA) is a major higher education public institution in Greece and South-East Europe. NKUA consists of 5 Faculties divided into 29 departments and its academic programs are attended by more than 50.000 matriculated students. The Department of Informatics and Telecommunications (I&T) is part of the Applied Sciences Faculty and consists of 40 academic staff, 20 technical-administrative staff, over 150 PhD candidates and research associates and approximately 1,500 undergraduate students; I&T is divided in three areas: Theoretical Computer Science, Computer Systems & Engineering and Communications & Signal Processing. The Laboratory for the Management of Data, Information and Knowledge counts more than 30 members, among which 3 are faculty, and has been involved in several research areas including Software Engineering, Databases and Knowledge Bases, WorkFlow Management Systems, Digital Libraries, User Interfacing, Personalization, Data Mining, and Distributed Systems. It has also successfully participated in relevant research and development projects.

In the context of D4Science-II, NKUA handles diverse tasks that fall within the expertise of its personnel including the: *(a)* design, implementation and deployment of information retrieval, distributed processing, presentation and personalisation systems and related services. These matters have been thematic areas for both research and development by the NKUA team throughout the last decade; *(b)* software production and site management tasks. These are areas of significant expertise as evidenced through project work undertaken within the context of the Greek Universities Network and national research initiatives; *(c)* coordination of training activities and participation to dissemination. As a research and teaching institution of higher education, NKUA is an apparent choice for the specific task. The list of related projects run during the last four years by the lab, includes D4Science (IST FP7), DRIVER II (IST FP7), DRIVER (FP6), DILIGENT IP (IST FP6), DIAS (eContent), DELOS Network of Excellence (IST FP6), BRICKS IP (IST FP6), Health-e-Child IP (IST FP6), Belief (IST FP6), Belief II (IST FP7), etc.

Yannis Ioannidis (yannis@di.uoa.gr) is a Professor at the Dept. of I&T of NKUA. He received his Ph.D. degree in Computer Science from the Univ. of California at Berkeley and he joined the faculty of the Computer Sciences Dept. of the Univ. of Wisconsin at Madison, where he became a Professor. His research work has been published in over 100 articles in leading journals and conferences and he holds two patents. He has received several awards for teaching excellence including the Presidential Young Investigator award (1991), and the 2003 VLDB 10-Year Best Paper Award. He has been a keynote or invited speaker in several conferences and a (co-)principal investigator in over twenty five research projects funded by various government agencies (USA, Europe, Greece) or private industry. He has been member of the program committees of over sixty conferences and currently he is an Associate Editor of several journals. He has served on the review board for the Lawrence Berkeley Laboratories in Berkeley and on the Science Council of the (now defunct) CESDIS Center for Excellence in Space Data and Information Sciences. Finally he has served as the Information Technology advisor to the Minister of Health of Greece.

Alex Delis (ad@di.uoa.gr) is a Professor at the Dept. of I&T of NKUA. He holds a Ph.D. in Computer Science from the Univ. of Maryland at College Park. He has taught at New York University-Polytechnic Institute, Queensland Univ. of Technology, Univ. of Patras, Catholic Univ. of Cordoba, and Univ. of Maryland. His research interests are in distributed systems and he has published in more than 80 journal and conference articles. His work has been supported among others by the US National Science Foundation, the Australian Research Council, the European Commission and the Greek Ministry of Education. He has served as PC member and organizer in more than 80 international conferences and has performed several seminars. Among the awards he received are the Best Paper Award in the 14th IEEE Int. Conf. on Distributed Computing Systems, the NSF CAREER Award and the Maurice V. Wilkes Medal for best paper in The Computer Journal. He has also been a Fulbright Fellow.

George Kakaletris (gkakas@di.uoa.gr) will act as gCube Software Manager for this I3 (cf. Section 2.1). He is Senior Software Engineer, holding a degree in Physics and an MSc on Informatics, both obtained by NKUA. His professional experience spans several years of employment by Greek IT companies both in S/W Engineering and R&D Dept. directorship/management positions. He has been a task / work package / project leader in numerous R&D projects. Currently, he works as an IT specialist for the Dept. of I&T of the NKUA. For the last 3 years he has also served as an ICT consultant to the Greek Parliament.

2.2.4 European Organization for Nuclear Research (CERN)

CERN, the European Organization for Nuclear Research, is funded by 20 European member states and has a yearly budget of approximately 1,000 MCHF. CERN has 2,500 permanent staff coming from the 20 member states and hosts about 9,000 High Energy Physics (HEP) scientists from all over the world

CERN is commissioning a new particle accelerator. The Large Hadron Collider (LHC) will be the world's most powerful accelerator providing data to four experiments each with up to 2,500 scientists and engineers coming from more than 250 institutes. These experiments will generate in the order of 15 Petabytes per year, over at least 10 years. These data will have to be shared with all the participating scientists looking for discoveries to understand the fundamental laws of nature. The computing capacity required to analyse the data far exceeds the capacity needs of any comparable physics experiments today and needs the combined resources of some 200 computer centres world-wide. CERN has chosen Grid technology to address the huge data storage and analysis challenge of LHC.

In D4Science-II, CERN will play a twofold role. The CERN IT team will coordinate the operation of the D4Science infrastructure benefiting from its pluri-annual experience in operating the EGEE production grid and its background knowledge as partner in DILIGENT and D4Science projects. CERN will also coordinate the quality assurance activity. The CERN Library, in cooperation with CERN IT, will coordinate and work on the study and implementation of interoperability solutions for the INSPIRE repository and the exploitation of the INSPIRE VRE to develop innovative services benefitting from it.

CERN, "where the Web was born", has been at the forefront of computing for many years and now leads the world's largest Grid project Enabling Grids for E-SciencE (EGEE). CERN also has a long tradition of collaborating with IT industry, including via EU-supported research programmes. The IT Department currently has 280 staff, predominantly engineers, who operate one of Europe's largest Computer Centres supporting over 10,000 users. The IT Department has prominently contributed, and contributes today in the context of the EC Framework Programme 7 (FP7), to several EC co-funded grid projects[14] aiming at extending the EGEE production grid infrastructure to new geographical areas, to serve new applications domains and to support the Grid community. In FP7 CERN coordinates the EGEE-III project, that operates the largest multi-disciplinary grid infrastructure in the world exploited by D4Science through interoperation with the gLite middleware, and ETICS 2, an e-Infrastructure for testing, integration and configuration of software that is currently providing the build and test system for the EGEE and D4Science software. The CERN Library is contributing to the FP7 PARSE.Insight project, aimed to shed light on issues of preservation of data in the light of a pan-European e-infrastructure for scientific data.

The CERN charter, over half a century ago, enshrined that "… the results of its experimental and theoretical work shall be published or otherwise made generally available" and this has inspired the CERN Library to play a leading role in both European and worldwide Open Access movements, aiming to provide anyone with immediate and free access to the results of scientific research. In partnership with other leading HEP laboratories worldwide, CERN is playing a crucial role in building INSPIRE: an emerging digital library

[14] In FP7: EGEE-III Coordinator (No.222667), SEE-GRID-SCI (No.211338), BalticGrid-II (No.223807), D4Science (No.212488), ETICS 2 Coordinator (No.223782), GridTalk (No.223534), EGI_DS (No.211693). In FP6: EELA (No.026409), EuChinaGrid (No.026634), EuMedGrid (No.026024), BalticGrid (No.026715), SEE-GRID (No.002356) and SEEGRID-2 (No. 031775), Health-e-Child (No.027749), DILIGENT (No.004260), ICEAGE (No.026637), OMII-Europe (No.031844), ISSeG Coordinator (No.026745), ETICS Coordinator (No.026753), EGEE Coordinator (No.508833) and EGEE-II Coordinator (No.031688). In FP5: GRACE (No.38100), DATAGRID Coordinator (No.25182)

system for HEP, which will be the one-stop shop for all information needs in the field. INSPIRE is built on the CERN Invenio open source digital library solution.

Frédéric Hemmer (CERN IT) studied Electrical and Mechanical Engineering (and Computing) in Brussels. He joined CERN in 1984 where he served as Systems Engineer in Databases, Real-Time Systems and more generally Distributed Computing. In the 1990's he became the software architect of the ComputerWorld Honors awarded CERN SHIFT project aiming at moving High-Energy Physics applications from Mainframes to Distributed RISC/Unix systems, later migrated to PC/Linux systems. He has been the initial author of the RFIO remote file access protocol, still in use today. From 1994 he took the responsibility of operating the Physics Data Processing Services at CERN (100's of machines, Terabytes's of data, Gigabit/second interconnections). As of 1998 he took responsibility of CERN Windows service (>5,000 computers) and later Mail and Web Services. In 2004 he joined the EGEE (Enabling Grid for E-sciencE) project where he served as Middleware Reengineering Manager (coordinating Grid Middleware development of 80 people across 8 countries). In 2005 he was appointed as CERN Deputy IT Department Head (280 Staff, 200 visitors, 50 MCHF budget in 2006), He served as EGEE related projects (EELA, SEE-GRID, EUMEDGRID, EUCHINAGRID, DILIGENT, ICEAGE, ETICS, Health-e-Child, etc.) Management Boards formal representative of CERN and holds the same role for the current EGEE-III related projects (BalticGrid-2, D4Science, ETICS-2, GridTalk, SEEGRID-SCI). He acts as a deputy to the CERN representative of the EGEE and EGI_DS PMBs. He is a member of IEEE and ACM.

Salvatore Mele (CERN Library) holds a PhD in Physics and worked for 15 years at the CERN LEP accelerator, where he led teams that measured fundamental physics constants, hunted for the Higgs boson and searched for hints of extra dimensions. He is now project leader of the CERN Open Access activities. In this capacity he is the interim project manager for the emerging SCOAP3 consortium, aiming to convert to Open Access the entire HEP literature. He is also one of the architects of INSPIRE, the future repository and information system for the entire discipline, as a liaison between the research, IT and library communities involved in its inception. He manages the CERN participation in the PARSE.Insight FP7 project and is the coordinator of the SOAP FP7 project.

Tim Smith (CERN IT) holds a PhD in Physics and worked for 10 years at the CERN LEP accelerator, where he executed technical projects in online and offline computing and led analysis teams performing precision measurements. Subsequently he has worked for 12 years in the IT department leading teams in computing farm management, physics data management, and now document management. He was a work-package manager of the EU DataGrid project, the forerunner of EGEE. For the past several years he has been heavily involved in initiatives to drive digital archives at the institutional and subject level and to populate them with content of a broad range of media types.

Jens Vigen (CERN Library) is CERN Head Librarian. Over the ten last years, he has been deeply involved in designing infrastructures for Open Access in Scholarly Communication and advocating Open Access to research results in all HEP sub-communities. He is collaborating with various partners in the information industry, ranging from database producers to electronic books retailers, on integration of third party services in a digital library setup. Currently he participates to the definition of the INSPIRE requirements, mainly focusing on computing intensive processes such as hybrid metrics for research evaluation, text mining of large data sets and the issues related to creating clusters of related material.

Roberta Faggian Marque (CERN IT) will act as Infrastructure Manager of this I3 (cf. Section 2.1). Since 2006, coordinator of the EGEE Related Projects Liaison Office (in EGEE-II and EGEE-III) she acts as the main link between EGEE and the collaborating projects to ensure they progress in a mutually beneficial way. In 2008 she was appointed the Service Manager of the D4Science production infrastructure to oversee the projects operation activities. She studied Computer Science at the University of Venice, she works at CERN since the year 2000. She collaborated to the implementation of the CERN library information system, coordinated the implementation of the electronic publication of the CERN bulletin, designed and implemented several dissemination Web sites. From year 2003 Roberta Faggian Marque has been involved in the Grid related activities at CERN, inside the IT department. Section Leader for the support of the EGEE collaborating projects, she has been involved and coordinated CERN's contribution to a number of EU co-funded projects linked to Grid technology, ensuring the necessary transfer of know-how and aligning their developments to EGEE various activities. Between them GRACE, SEE-GRID, DILIGENT and D4Science.

2.2.5 Engineering Ingegneria Informatica SpA (ENG)

Engineering Group is Italy's largest systems integration group and a leader in the provision of complete IT services and consultancy. Engineering is present with 6500 employees and 35 branch offices, throughout Italy, has a direct commercial presence in the EU, in Ireland and Belgium, and outside the EU in Brazil and Latin America. The Group has a global production capacity in 30 different countries, managing IT initiatives linked to the development of projects in the industrial and telecommunications sectors. It has strengthened its presence in the managed operations sector thanks to the European data centre at Pont St. Martin which handles 100 clients nearshore. Consolidated revenues of 457.1 million euro for the year 2007, a 7.4% increase over 2006; a pro-forma value of production of 709 million euro in 2007. The holding company, Engineering Ingegneria Informatica, has been listed on the Milan Stock Exchange since December 2000 and ordinary shares are traded in the AllStars segment. Of the total 12.5 million shares, 33% are traded on the market and 67% are held by company founders.

The Engineering Group operates through 7 business units: Finance, Central Government, Local Government and Healthcare, Oil Transportation and Services, Utility, Industry and Telecom, supported by an SAP transverse skills centre and by its Central Office for Research & Innovation, with researchers active in Italian and international research projects and an investment budget of 50 million euro for the current three-year period 2006-2008.

Engineering was one of the first Italian companies to adopt the ISO 9001 standard for quality management systems in the early 1990s. Since 1996 the company has adopted NATO standard AQAP 2110/160 certification. In the software development sector, the group holds level 3 certification in accordance with the CMMI® model (Capability Maturity Model Integration, version 1.2).

The Engineering service centre at Pont St. Martin is one of Europe's most advanced, managing 40,000 workplaces, 1000 remote connections, 10,000 electronic mail boxes and about 7000 SAP users. With 100 outsourcing contracts for over 3500 servers, Engineering provides the following services in close proximity to its clients and through different types of contract:

- application management of multi-year information system management contracts;
- business process outsourcing of entire vertical components of the client's information system; with its specialist staff, the centre guarantees the performance not only of the application component but of all process aspects (administration, accountancy, report management, billing and settlements) underlying business activities;
- facility management;
- workplace management services (Fleet Management, Service Desk);
- centralized management and distributed management services in a variety of contexts (Enterprise Monitoring Center, Server & Storage management, Network management);
- solutions and services for the management of complex technological upgrade projects, optimization of IT resources, implementation of specific solutions (Server Consolidation & Virtualization – Disaster Recovery);
- professional technology, design and methodology consultancy services.

ENG R&D Laboratories: The R&D Department, driving the R&D initiatives for the whole Group is organised to work in strict cooperation with business divisions in order to facilitate knowledge and technology transfer processes as well as the direct involvement of production staff within research initiative, allocating an annual budget of around 5% of the global turnover.

R&D facts:

- Participation in over 50 international research projects;
- Steering committee chairman and founding member of ETP (European Technology Platform) **NESSI** – Networked European Software and Services Initiative;
- Steering committee member of ETP **NEM** - Networked Electronic Media;
- 120 researchers.

Grid R&D Unit: this R&D unit, is made by twenty researchers with strong links toward Engineering senior consultants and production directors for technology transfer and requirements and challenges gathering. Currently this unit is involved in the following grid-related initiatives:

- **D4SCIENCE** (Digital Library Infrastructure test-bed on Grid-enabled Technology) – FP7-2007-RI212488
- **ETICS2** (eInfrastructure for Testing, Integration and Configuration of Software) – FP7-2008-RI223782
- **GriFin** (Grid For Finance) – Italian Minister of Research and Education – DM28488 ex Law297/1999
- **ECHOGRID** (EC-China strategic GRID Roadmap) – FP6 –2006-IST-045520
- **NESSIGRID** (Networked European Software and Services Initiative-Grid) – FP6-2005-IST-033638
- **ERINA** (Recommendations for Exploiting Research INfrastructure Potential in key areas of the Information Society (e-Government, e-Health, e-Learning) – SMART- 2006/0046 – 20CE-0094783/00-02

In the context of the proposed project, ENG will mostly concentrate on the same areas for which it has been responsible in the Diligent and D4Science projects, namely: *(a)* the design and implementation of security-related solutions, focusing on interoperability aspects and *(b)* the overall coordination of the integration, testing and distribution activity with a specific involvement in the integration task.

Andrea Manieri (andrea.manieri@eng.it) graduated in Computer Science with Prof. C. Boehm at the University of Rome "La Sapienza", he joined Engineering Ingegneria Informatica R&D in 1998. In 2000 he has been involved as Technical Leader in the ECOLNET[15] project. In 2002 he was appointed as RTD activities coordinated in Engisanità Spa, an Engineering Group company on HealthCare market. On April 2003 he joined again the Engineering labs as responsible of development of new business on Grid Technology and recently (2007) he's been appointed as head of Grid and service-based infrastructure R&D unit.

Paolo Roccetti (paolo.roccetti@eng.it) received his University Degree in Computer Science at University of Bologna (Italy) with a thesis on "System engineering on a Grid Computing environment: integration with a coordination infrastructure". In March 2005 he joins the R&D laboratory as responsible of grid security developments. Since March 2007 he took the responsibility of the grid security in GriFin technical development. His interests are mainly related to Grid Computing, focusing on security of distributed and heterogeneous environments.

Paolo Fabriani (paolo.fabriani@eng.it), received his University Degree in Computer Science in July 2000 at University of Rome "La Sapienza". After a one-year collaboration with the same University and CNR, he joined the R&D Laboratory of Engineering Ingegneria Informatica S.p.A. Since 2001 he has been involved in different research projects (TrainME, ArchWare, ECOLNET, DILIGENT, ETICS) and other initiatives (design and development of JDuck, a set of graphic libraries for web application development). Currently he is responsible of the D4Science integration and build activity. His main research interests include distributed systems, web technologies and human-computer interaction.

2.2.6 University of Strathclyde (DBM-USTRATH)

With its 5 faculties and 44 departments, the University of Strathclyde is the second largest in Glasgow, Scotland, holding a long-standing worldwide reputation for excellence in academic research and innovation. The Department of Computer and Information Sciences in the faculty of Science undertakes world-class research in Information Science, Artificial Intelligence, and Software Engineering, and prides itself for its external research collaborations in Scotland, the rest of the UK, and abroad. The I-Lab Group is concerned with the design and evaluation of advanced access technology to multimedia information. Over the past four years, BDM-USTRATH has been involved in a number of projects directly or indirectly related to the aim of the proposed project, including, PENG (FP6), REVEAL-THIS (FP6), DILIGENT (FP6), D4Science (FP7).

[15] (European Collaboration Network) – FP5 - IST-11061-1999

In the context of the proposed project, BDM-USTRATH will mostly concentrate on the same areas for which it has been responsible in the Diligent and D4Science projects, namely: *(a)* the design, implementation, and operation of back-end annotation services and associated client-libraries, *(b)* the design, implementation, and operation of the services for content-based source description, selection, and fusion in the context for distributed search subsystem, *(c)* the design, implementation, and management of the application framework upon against which all other gCube services are implanted.

Richard Connor (richard.connor@cis.strath.ac.uk) has been working in computer systems research since 1985, when he became part of Persistent Programming Research Group at St Andrews. He moved to Glasgow University in 1997 as an EPSRC Advanced Research Fellow, and formed a research group investigating the use of persistent language paradigms within the context of Internet and global computation systems. In 1999 he moved to a chair of Computer Science at the University of Strathclyde, and the primary focus of his group's work there is in effective and efficient query systems for semistructured, globally distributed information. He has an extensive publication record in conferences (OOPSLA, EDBT, DBPL and POS) and journals (ACM TOPLAS, The Computer Journal, Journal of Software Technology, and IEEE Internet Computing), has co-chaired international workshops in both the Persistent Object Systems and Database Programming Language series (both part of the SIGMOD anthology), and has served on the programme committees of VLDB, POS, DBPL, BNCOD, ICDE, HICSS, WAIM and ACM/SAC. In the last ten years, he has jointly held over a million pounds of research funding, including awards from: SERC, EPSRC, BBSRC, SHEFC, Esprit (Frameworks 2, 3, 4 and 5), and ICL; he has been the PI for five of these awards. Professor Connor is a member of the EPSRC peer review college, and has served on panels for computer science, software technologies, health informatics, advanced and senior fellowships, and JREI. He has chaired panels in software technologies and CS advanced fellowships. Since 1999 he has worked as a consultant to the Research and Standards group at Reuters, notably with respect to mediated information architectures in addressing the problems of real-time global messaging systems. He is a member of the advisory board of the Kelvin Institute, and is a Senior Consultant to Enigmatec Corporation Ltd, a software technology company funded by Amadeus Capital Partners, Pentech Ventures, and Intel.

Fabio Simeoni (fabio.simeoni@cis.strath.ac.uk) is a graduate from the Department of Mathematics of the University 'La Sapienza' in Rome and has obtained his Msc. Degree in Advanced Information Systems from Glasgow University, in Scotland. Since 1999, he has been working as a Senior Researcher at the Department of Computer and Information Sciences of the University of Strathclyde, also in Glasgow, investigating programming models and architectural principles for widely distributed and loosely-coupled systems, particularly in relation to information retrieval and aggregation services within the practice of Digital Libraries services. Since 2004, he has been involved in the Diligent and D4Science projects, where he has led the design and development of the gCube Core Framework (gCF) as well as service-based frameworks for Distributed Information Retrieval and Annotation Management.

2.2.7 The Food and Agriculture Organisation of the United Nations (FAO)

The Food and Agriculture Organization of the United Nations, the largest autonomous agency within the United Nations system and one of the main players in the agricultural information community, was founded in 1945 with a mandate to raise levels of nutrition and standards of living, to improve agricultural productivity, and to better the condition of rural populations. FAO is active in land and water development, plant and animal production, forestry, fisheries, economic and social policy, investment, nutrition, food standards and commodities and trade. Its specific priority is encouraging sustainable agriculture and rural development, a long-term strategy for the conservation and management of natural resources.

The Knowledge Exchange and Capacity Building Division (KCEW) is in charge of information and knowledge management within the Organization with experience in developing web applications based on semantic and knowledge systems. As a means for knowledge organization, FAO provides AGROVOC, the world's largest multilingual thesaurus for agricultural sciences and related subjects. Extensive work, including feasibility studies and application development at the prototype stage, has already been carried out in-house. Currently, FAO is collaborating with international partners in the design and implementation of a distributed architecture for the management and maintenance of domain ontologies.

The Fisheries and Aquaculture Information and Statistics Service (FIES) is responsible for the collection, compilation, validation, analysis and dissemination of reliable and up-to-date information on all aspects of world fisheries and aquaculture. The Service coordinates and provides advice assistance and information to FAO Members on the development of international standards for fishery and aquaculture statistics. It supports partnership arrangements for the collation and sharing of information, provides access to global fishery and aquaculture information via corporate library resources and internet-based information systems, provides editorial support and facilitates communication and preparation of publications and web pages according to corporate standards.

In the context of the proposal, the FAO team handles a number of tasks that fall within its expertise which include (a) communication, dissemination and training; (b) identification and promotion of appropriate international standards; (c) organization of user communities and their requirements for the realisation of a knowledge ecosystem, including VRE definition, cross community exploitation and VRE validation.

Johannes Keizer (Johannes.Keizer@fao.org) (Information Systems Officer) holds an MSc in Biochemistry and a PhD in Biology. He has been involved in the development of information systems for more than 10 years. He has been working with FAO since 1998 and is currently responsible for FAO's Documentation Catalogue, an international network of documentation centres (AGRIS) and the multilingual agricultural thesaurus, AGROVOC. His Team consists of about 18 experts, working in documentation, ontology and thesaurus development and metadata standards. Initiatives such as the AgMES and the AOS have been launched by the Team under his leadership. He is currently the Outreach manager of the D4Science project.

Marc Taconet (Marc.Taconet@fao.org) holds a Master's degree in Fisheries sciences. He has been working with FAO since 1987 in various African countries starting with fishery bio-statistics, evolving progressively to data bases, GIS and spatial approaches to fisheries monitoring, at national, regional and finally global scale. Since 1999, M. Taconet has lead at FAO Fisheries and Aquaculture Department in Rome the development of the Fisheries Global Information System (FIGIS), which now provides the web based infrastructure to the Department's integrated information resources, and the backbone to global information partnerships for the monitoring of world fishery resources. M. Taconet leads a team of developers and information management and communication specialists of about 15 persons. He is the Fisheries and Aquaculture Resources Management (FARM) Communities Manager for the D4Science project.

Yves Jaques (Yves.Jaques@fao.org) holds an MSc in Software development. He has been working with FAO Fisheries and Aquaculture since 1999, developing fisheries software including statistical stand-alone and web-based systems. A strong proponent of user-driven development he has worked extensively with fisheries scientific user communities in requirements gathering, verification and validation activities. A keen supporter of federated data systems, he has worked consistently with colleagues in other FAO departments to consolidate standards, bridge data islands and produce synthetic information products that merge previously fragmented data into new value-added products.

2.2.8 Fishbase Information & Research Group Inc. (FIN)

The FishBase Information and Research Group Inc (FIN) is a national non-stock, non profit, and tax-free non-governmental organization established in the Philippines in 2003. FIN has been created and continuously being supported by the FishBase Consortium, a group of nine international institutes and organizations that is providing long-term support for FishBase. Members of the FishBase Consortium are: WorldFish Center (Penang, Malaysia); Food and Agricultural Organization of the United Nations, (FAO, Rome, Italy); Institute of Marine Research (Kiel, Germany); Fisheries Center, University of British Columbia (Vancouver, Canada); Museum National d'Historie Naturelle (Paris, France); Royal Museum for Central Africa (Tervuren, Belgium); Swedish Museum of Natural History (Stockholm, Sweden); Aristotle University of Thessaloniki (Thessaloniki , Greece); and Chinese Academy of Fishery Sciences (Qiangdao, China).

FIN's mandate is to support the growth of FishBase, a public domain information system dedicated to enhance understanding, conservation and management of fishes worldwide; to support interaction of the FishBase team with international experts on taxonomy, ichthyology, biodiversity and fisheries; to support participation in cutting-edge research in these fields; and to support the interaction with FishBase users and partners worldwide to make sure their needs and contributions are accommodated. FishBase is currently the largest global services information system with about 1.3 million visitors per month and over 1000 citations in scientific literature. FIN, with strong links with the FishBase team of the WorldFish Center, have worked with them in more than 10 projects involving fisheries databases and research through funding support from international and national donors. Specifically, for AquaMaps we have produced range maps for marine and freshwater species, and have developed tools for cross species analyses.

In the context of the proposal, the FIN team will ensure the technical and organisational support needed to make the AquaMaps data e-Infrastructure interoperable with D4Science as well as lead the development of innovative services to primarily serve the AquaMaps scenario.

Rainer Froese (rfroese@ifm-geomar.de) will act as the Scientific Supervisor in his capacity as Chair of the FIN Board. Dr. Froese is a senior scientist at IfM-GEOMAR. He has a proven track record of successfully leading large international groups (such as the FishBase team and Consortium, and the INCOFISH project) and delivering products such as web portals (www.fishbase.org), CD-ROMS (e.g., Catalogue of Life), and multi-authored books (e.g., FishBase 2000) on time. He was funding member of international biodiversity initiatives such as Species 2000 and OBIS, and a member of GBIF Governing Board and as good personal relationships with most of the international players in the area of biodiversity. He is the principle architect of FishBase. He specializes in fisheries and aquatic biology issues, and has authored or co-authored over 100 scientific publications with over 600 citations in ISI Web of Science. He is one of few Europeans who received the prestigious Pew Fellowship Award in Marine Conservation.

Mary Ann Bimbao (mpbimbao@yahoo.com) will serve as the Administrator for this initiative. Dr. Bimbao, a funding member of FIN and its Board Treasurer, is the Executive Director of FIN. She has more than 20 years of experience in various projects involving peoples organizations, non-government organizations, research organizations and academic institution in the developing world. She specializes in the fields of fisheries economics and gender issues. She has created databases on socio-economics particularly, having significantly contributed to the development of RESTORE (Research Tool for Natural Resources Management, Monitoring and Evaluation), a tool that comprises a set of participatory research procedures and computer-based analysis for sustainability while she worked with the International Center for Living Aquatic Resources Management (formerly ICLARM, now WorldFish Center). Also while at ICLARM for over 13 years she served as the Secretary of the Network of Tropical Aquaculture Scientists and Assistant Editor of its newsletter Aquabyte, managing the information exchanges of 500 members in 100 countries. She was the Administrative Officer of WP1 INCOFISH EU-funded project (2005-2008). She will also serve as the Administrator of an upcoming, accepted EU-funded project under FP7, on Highland Aquatic Resources Conservation and the Sustainable Development of Aquaculture.

2.2.9 4D SOFT Software Development Ltd. (4D SOFT)

4D SOFT Software Development Ltd. is a privately held SME with 30+ employees, founded in 1990. The company’s activity area covers the whole development process from the consulting to creation of a business model, from software design to introduction of the system, and especially software maintenance and testing. Within the area of software engineering 4D SOFT deals with, test outsourcing, test planning, independent validation and verification and development of testing tools.

Based on its professionals’ theoretical and practical knowledge 4D SOFT improves the testing processes at different leading national and international companies. Apart from the business field 4D SOFT regularly get in touch with different research organizations and universities to share, maintain and evolve our state-of-the-art knowledge in this profession.

Finishing an EU IST project Component+ 4D SOFT introduced a new unit testing method called EBIT (Extended Built-In Testing). In this way units with difficult internal data structure can also be tested. It applied an improved category-partition testing for test planning. 4D SOFT also deals with test automation, where test cases are generated from atomic test cases.

4D SOFT participates in the research and implementation of new testing tools such as General Java Tester that is unit testing tool for Java and a unique static analyser tool called Deeptest. This tool is capable to deeply analyse java, it key features are: impact analysis, revealing null pointer exceptions and throw catch pairs.

EU IST project experience: Component+ (FP5), DILIGENT, ETICS, PLASTIC (FP6), D4Sience, ETICS2 (FP7).

The clients of 4D SOFT Ltd. come from various sizes and business fields such as commerce, patent/trademark, telecommunication, banking, insurance and so on. As a small company it focus on the SME market in which it has already built a good reputation and business network. The company continuously and consciously manage and emphasize these relationships in order to gain a better understanding of the changing market demands.

In the context of the proposal, the 4D SOFT team will (*i*) define the test plan, (*ii*) set up and maintain the testing infrastructure by applying ETICS and (*iii*) coordinate and implement of the testing activities.

István Forgács (forgacs@4dsoft.hu) received his PhD in computer Science in 1993. His research interests include software testing, data flow analysis, slicing and program maintenance. István Forgács has 30+ papers in leading journals and conference proceedings. He was the PC member in numerous conferences such as ISSTA, ESEC-FSE, etc. Currently he is the managing director of 4D SOFT Kft. He led several significant Hungarian and more international projects including EU IST project Component+, DILIGENT, ETICS, ETICS2, D4Science and the development of Panorama Analyser for COBOL, Deeptest for Java and CREG++ for C++ that are a static regression testing tool uniquely capable of the impact analysis of large code. He was the inventor and project leader of Y2K.O. that was the only dynamic Y2K tool capable of safely investigating over 300,000 LOC/day/person.

2.2.10 Terradue S.r.l. (TERRADUE)

Terradue S.r.l. is a private company created in February 2006 and supported by the European Space Agency (ESA) and the Lazio's Business Innovation Centre (BIC) program of pre-incubation and incubation. The core elements of the company worked for nearly four years in the development of a GRID infrastructure in the European Space Research Institute (ESRIN) of ESA to support Earth Observation operational and scientific applications and services.

Terradue works to exploit and strengthen best practices in distributed data processing, archiving and discovery putting the emphasis on the delivery of robust operational systems while keeping a concrete roadmap to build the next generation data processing and storage systems. Terradue focuses its activities on the use of Web Services, GRID and peer-to-peer technologies to support distributed spatial data management, and high performance computing applications in collaborative digital environments. Terradue has a strong knowledge in the management and distribution of very large spatial data sets carried out collaboratively with best-known open source projects.

Terradue is a technological partner of GENESI-DR [32], an EC FP7 project whose challenge is establishing open Earth Science Digital Repository access for European and worldwide science users allowing these to seamlessly access and share all data, information, products and knowledge originating from space, airborne and in-situ sensors. Terradue participated in the EC FP6 project Business Experiments in GRID (BEinGRID) whose mission is establishing effective routes to foster the adoption of Grid technologies. The result of the participation is the successful integration of ESA Data User Element (DUE) GlobAerosol processor into a GRID infrastructure as an example of a class of services requiring large and near real time data access and complex processing. Terradue maintains and develops the ESRIN GRID Web Services infrastructure for Earth Observation and is responsible of the integration of the ESA Principal Investigators scientific processor on the ESRIN GRID infrastructure.

In the context of the proposal, the Terradue team leads and participates in tasks aiming to provide interoperability solutions between GENESI-DR and D4Science-II and Community Specific Resources Implementation needs for the interoperability. Terradue is in a strong position to guarantee the success of D4Science-II as it has a deep knowledge of the technologies adopted within GENESI-DR and a strong background in EO services development and operations.

Dr Pedro Gonçalves (pedro.goncalves@terradue.com) is an Environmental Engineer with a PhD on 'Scale Invariance of Forest Fires Spatial Patterns – Environmental Modelling in a Global Distributed Data Access Architecture'. Pedro is one of the founders of Terradue where he acts as the Technical Director of Terradue activities and leads the Research and Development activities. He was the Project Manager of the Earth Observation GRID Processing on-Demand (G-POD) project for its infrastructure definition and conception. He was POC for the GRID integration with other national space agencies in the context of CEOS-GRID project.

Fabrice Brito (fabrice.brito@terradue.com) is a Forestry Engineer. He founded Terradue with Pedro Gonçalves where he is the project manager and responsible for the ESRIN GRID Infrastructure maintenance project lead. While in ESA where he participated in the development of several G-POD components. He designed and integrated in G-POD EO data processing services used for the production of MERIS Level 3 products (e.g. MERIS MGVI products in cooperation with the EC-Joint Research Centre) and for communication purposes (e.g. global MERIS RGB composite products) and all scientific projects supported by ESA-ESRIN Grid Processing on-Demand Infrastructure.

2.3 Consortium as a whole

D4Science-II consortium has been formed in a way that effectively addresses the following objectives:

- To establish strong links with all the task forces or entities that the foreseen work depends upon;
- to build a strong scientific and technological basis for realising the project;
- to maximise the impact and the sustainability of the project's outcome; and
- to create a strong project leadership that will manage to ensure the feasibility of the work plan and the delivery of high quality output.

Managed by ERCIM, all partners already participating in the ongoing D4Science effort including CNR, NKUA, BDM-USTRATH, CERN, ENG, FAO and 4D SOFT have established well defined roles. By distributing tasks and responsibilities according to the technical competencies of individual participants the consortium has already formed a sustained group of players capable to satisfy the demanding and dynamic nature of a research undertaking to the scale of D4Science-II.

CNR has already acquired significant experience through its technical coordination of D4Science and DILIGENT and is recognized by all partners as the lead for pertinent technical activities. With its experience in developing fundamental components, i.e., Core Services, of the gCube system, CNR is now the consortium expert in successfully addressing crucial factors and vital issues when it comes to defining highly distributed dynamic VREs. Furthermore the organisation exposes an expertise profile in the area of Information Organisation and Management which is essential for the realisation of several aspects of the gCube system dedicated to the data management. It is also currently involved, together with NKUA, in activities dedicated to identify and promote interoperability among large digital libraries. This will be beneficial both in the definition of solutions for interoperability and in liaising with data services that might become components of the ecosystem. These facts, along with the breadth and depth of research activities performed by the organisation, favours CNR's assignment to the role of Scientific and Technical Leadership of the consortium.

NKUA provides expertise in the areas of distributed search and optimization and it is expected to further enhance the already well-established corresponding gCube functionalities. During the last few years, NKUA has designed and developed core search, query processing methods, and process optimisation and execution (especially under SOA environments) functionalities. These are all crucial aspects for the development of a distributed VRE. NKUA has disseminated its project findings to scientific communities. During the last year, NKUA in cooperation with BDM-USTRATH has invested in low level aspects of information retrieval, such as indexing of diverse types of data including geospatial/temporal, textual and even high-dimensional feature vectors, data transport and data processing flows, and finally creating a comprehensive approach for serving data e-Infrastructures. With its established international recognition in distributed information retrieval, the BDM-USTRATH group adds new dimensions to the abilities of the infrastructure. The combined teams of NKUA and BDM-USTRATH have served as project experts in distributed search and information retrieval technologies over grid environments. Together, they have contributed to the implementation of the advanced gCube Information Retrieval Services components.

Any community has its security requirements and any grid infrastructure has its own policies. The strong competence of ENG is just that; ENG is, the partner who designed, assisted with the development and supervised the security-related activity in DILIGENT and D4Science. ENG has a particular know-how on web security, and its presence in national projects on grid security (e.g., GRIFIN, Grid for Finance), will assure the adoption of proper solutions. Therefore, ENG will also lead the security area of D4Science-II as it maintains expertise in providing solutions for secure applications over e-Infrastructures and will extend its solutions in the challenging case of e-Infrastructure interoperability. ENG has established cooperation with

EGEE (as in the Middleware Security Working Groups) and actively participates in OGF and other initiatives on Authentication, Authorization and Accounting (AAA). Being a major Italian ITC industry, ENG has served as the commercially-minded partner and has helped stir the project towards the definition of a solid development environment and the needed procedures to achieve the quality level for software products. ENG is certified ISO9001 and has recently achieved the level 3 on CMMi for all its production departments.

gCube system exploits the international grid infrastructure that has been established by the EGEE project [24] led by the CERN IT Department. This particular department of CERN has played a major role in DILIGENT and D4Science by contributing in the component validation and the seamless integration of gCube with gLite [35]. It is thus a key partner for the effective utilisation of EGEE's infrastructure and for communicating technical and non-technical requirements to EGEE. With its vast experience of serving users through e-Infrastructures, CERN's team participation to the project is essential. Apart from managing the relevant service activities, the team from CERN will establish and monitor the rules for providing high quality services to the e-Infrastructure consumers. Additionally, with the contribution of other partners, the team is expected to validate the components developed by the project and will evaluate the integration of the newly produced packages in the existing gCube system.

The complexity of the produced software necessitates well-defined and formal testing procedures. 4D SOFT with its expertise in software testing complements the consortium by contributing knowledge, experience and its domain leadership. Having gCube tested during the development and deployment of D4Science-II is a major challenge towards furnishing quality, production-level software. Testing a complex platform such as gCube requires that testers be fully aware of not only the platform components but also of how these elements are combined in order for the tests to be meaningful. Having followed the entire life cycle of the gCube system development, 4D SOFT maintains two advantages: (*i*) it is an external and consequently an independent player into the development and (*ii*) it has a complete and clear view of the entire architecture and its interdependencies.

The front-end of gCube's capabilities are the various portals that have to be developed according to the specific requirements of involved user communities. NKUA and CNR have had an already lengthy collaboration in the context of DILIGENT and D4Science and collaboratively produced the front end for several VREs. The continuation of this effort is mandatory for the completeness of the solutions to be provided so that the proposed ecosystem becomes a reality.

Finally, ERCIM will carry out the administration of the project as the organization in question has shown time and again proven expertise in the field. ERCIM has very successfully coordinated all the financial and administrative aspects of the D4Science and DILIGENT projects. ERCIM offers D4Science partners logistical support in ascertaining needs, ensures the fulfilment of financial requirements as defined by the European Commission, and addresses specific organisational challenges that might affect the execution of the work plan as a whole. More importantly, ERCIM has coordinated several project consortia in FP6 and FP7, efficiently managing over 61 M € in European Community contribution to date, establishing the expertise and trust required to address the inherent complexity of an I3 project.

In addition to project management experience, ERCIM will facilitate the introduction of standardisation activity in the D4Science ecosystem. On the one hand, ERCIM hosts the European headquarters of the World Wide Web Consortium (W3C), thereby strengthening relationships throughout Europe to better support the development of Web technology and to jointly share the results of their collaboration. Then in 2007, ERCIM and the European Telecommunications Standards Institute (ETSI), both based in Sophia Antipolis, France, signed a landmark Memorandum of Understanding to launch a close collaboration between the two organisations. The cooperation with ETSI allows researchers to accelerate the innovation chain by overlapping research, testing, standardisation and exploitation, thus creating a leading position for the ICT industry within Europe. Concretely, the "Infinity Initiative", supported by the European Commission, was established envisaging a number of joint events which include advanced workshops, seminars and meetings to strengthen the links between standardisation and research. Through ERCIM's tight collaboration with W3C and ETSI, both standardisation bodies will be engaged in the open exchange of knowledge for promoting the adoption and dissemination of standards in the D4Science-II ecosystem.

The consortium of the D4Science project has evolved in moving to D4Science-II. New partners with specific expertise on the data e-Infrastructures enabling technologies and on the scenarios' application domains addressed by the project have been included.

Facilitating and securing the long-term sustainable development and utilisation of the world's fisheries and aquaculture is one of the missions of FAO. Working teams from FAO's Fisheries and Aquaculture Information and Statistics Service and from the Knowledge Exchange and Capacity Building department (engaged in creating global shared standards for knowledge exchange in agricultural science and technology) are strongly involved in meeting this objective. They will both participate in D4Science-II by bringing their requirements, technical expertise and contacts in this area. The FCCPS and ICIS scenarios are part of concrete activities currently performed within these departments. A similar role is played by FIN. FIN works towards improving fisheries through developing fisheries information systems that are being used as references by scientists, policymakers, researchers and other stakeholders in their initiatives to increase fish production, manage fisheries resources and aquatic ecosystems. In particular, the FIN team is responsible for the development and maintenance of the AquaMaps service. As such it can offer the best support for the integration of this data service in the ecosystem.

TERRADUE, being one of the main technological partners in the GENESI-DR project, brings necessary technological knowledge to achieve interoperability with the GENESI-DR repository e-Infrastructure. Moreover, with TERRADUE's experience in the use of Web Services; grid (they have actively participated in the development of the ESA operational grid infrastructure) and peer-to-peer technologies to support distributed spatial data management; and high performance computing applications in collaborative digital environments, they will also support the development of D4Science-II scenarios, such as AquaMaps.

The CERN Library Department is responsible for the organisation, design, development and maintenance of the INSPIRE repository. CERN Library also has extensive expertise in the development of repository software. In particular, they have developed Invenio, the open source digital library system that is used to support INSPIRE. CERN is thus expected not only to substantially contribute to the integration of the INSPIRE service in the ecosystem, but also to the definition of general approaches to interoperability.

Finally, NKUA and CNR, as leading partners of the DRIVER I & II Projects and the main contributors to the DRIVER e-Infrastructure, will bring into the consortium the required know-how for implementing the planned interoperability and for achieving the DRIVER scenario.

The use of the D4Science-II production infrastructure will serve to demonstrate the advantages of the emerging ecosystem approach to the scientific communities represented in the consortium. By taking advantage of the strong networks of international projects and initiatives presented by the D4Science partners, which include International Organizations participating in challenges having a global reach, D4Science intends to extend to many other fields of science, thereby progressively extending interoperability links and clientele far beyond the current consortium. As evidenced in Appendix B, numerous European and international initiatives have expressed a desire to establish synergies and exploit the capabilities of the D4Science-II production e-Infrastructure, demonstrating faith in the capabilities of the current consortium and optimism for sustainability prospects.

2.4 Resources to be committed

Form A3.1: Consolidated budgets

Participant Short Name	Method of calculating indirect costs	Estimated eligible costs and requested EC contribution		RTD	Coordination	Support	Management	Other	Total
ERCIM	Real Indirect Costs	Eligible costs	Personnel	0	22 372	0	83 895	0	106 267
			Subcontracting			40 000			40 000
			Other direct costs	0	2 105	22 000	15 895	0	40 000
			Indirect costs	0	25 712	1 540	96 420	0	123 672
			Total budget	0	50 189	63 540	196 210	0	309 939
		Requested EC contribution		0	26 190	63 540	196 210	0	285 940
CNR	Simplified Method	Eligible costs	Personnel	227 040	208 120	0	18 920	165 550	619 630
			Subcontracting				1 000		1 000
			Other direct costs	16 489	15 115	0	1 374	12 023	45 001
			Indirect costs	197 434	180 981	0	16 453	143 962	538 830
			Total budget	440 963	404 216	0	37 747	321 535	1 204 461
		Requested EC contribution		330 722	238 861	0	37 747	321 535	928 865
NKUA	Transitional Flat Rate	Eligible costs	Personnel	243 000	189 000	0	5 400	113 400	550 800
			Subcontracting				1 000		1 000
			Other direct costs	40 000	0	0	0	0	40 000
			Indirect costs	169 800	13 230	0	3 240	68 040	254 310
			Total budget	452 800	202 230	0	9 640	181 440	846 110
		Requested EC contribution		339 600	202 230	0	9 640	181 440	732 910
CERN	Transitional Flat Rate	Eligible costs	Personnel	258 997	126 007	0	24 806	212 280	622 090
			Subcontracting						0
			Other direct costs	10 360	5 263		1 388	13 161	30 172
			Indirect costs	161 614	9 189	0	15 716	135 265	321 784
			Total budget	430 971	140 459	0	41 910	360 706	974 046
		Requested EC contribution		323 228	140 459	0	41 910	360 706	866 303
ENG	Real Indirect Costs	Eligible costs	Personnel	78 750	52 500	0	0	118 125	249 375
			Subcontracting						0
			Other direct costs	4 737	3 158	0	0	7 105	15 000
			Indirect costs	37 850	25 234	0	0	56 776	119 860
			Total budget	121 337	80 892	0	0	182 006	384 235
		Requested EC contribution		60 669	59 554	0	0	182 006	302 229
BDM-USTRATH	Transitional Flat Rate	Eligible costs	Personnel	132 788	84 984	0	0	0	217 772
			Subcontracting						0
			Other direct costs	15 000	0	0	0	0	15 000
			Indirect costs	88 673	5 949	0	0	0	94 622
			Total budget	236 461	90 933	0	0	0	327 394
		Requested EC contribution		177 346	90 933	0	0	0	268 279
FAO	Transitional Flat Rate	Eligible costs	Personnel	94 178	174 902	0	6 727	40 362	316 169
			Subcontracting						0
			Other direct costs	7 447	23 830	0	532	3 191	35 000
			Indirect costs	60 975	13 911	0	4 355	26 132	105 373
			Total budget	162 600	212 643	0	11 614	69 685	456 542
		Requested EC contribution		121 950	212 643	0	11 614	69 685	415 892
FIN	Transitional Flat Rate	Eligible costs	Personnel	93 000	16 500	0	0	21 000	130 500
			Subcontracting						0
			Other direct costs	20 000	0	0	0	0	20 000
			Indirect costs	67 800	1 155	0	0	12 600	81 555
			Total budget	180 800	17 655	0	0	33 600	232 055
		Requested EC contribution		135 600	17 655	0	0	33 600	186 855
4D SOFT	Transitional Flat Rate	Eligible costs	Personnel	0	0	0	0	61 440	61 440
			Subcontracting						0
			Other direct costs	0	0	0	0	10 000	10 000
			Indirect costs	0	0	0	0	42 864	42 864
			Total budget	0	0	0	0	114 304	114 304
		Requested EC contribution		0	0	0	0	114 304	114 304
Terradue	Transitional Flat Rate	Eligible costs	Personnel	81 900	56 700	0	0	37 800	176 400
			Subcontracting						0
			Other direct costs	10 000	0	0	0	0	10 000
			Indirect costs	55 140	3 969	0	0	22 680	81 789
			Total budget	147 040	60 669	0	0	60 480	268 189
		Requested EC contribution		110 280	60 669	0	0	60 480	231 429
Total Budget				2 172 972	1 259 886	63 540	297 121	1 323 756	5 117 275
Total Requested EC contribution				1 599 395	1 049 194	63 540	297 121	1 323 756	4 333 006

To achieve its ambitious goals and objectives, the D4Science-II consortium has developed a careful financial planning covering all foreseen activities of the project. The table above summarises the costs related to the project per partner and per activity, respectively. Please note that the ***real*** and ***actual*** average person-month rates per partner were used for the calculations concerning "personnel". The indirect costs were calculated according to the partners' individually selected methods for determining indirect costs, with the exception of the Coordination activities, in which indirect costs are limited to 7% of direct eligible costs:

RTD: The personnel costs within this category are directly related to the creation of new knowledge and new technology as described within Joint Research Activities. A total of 257 person-months have been dedicated to the RTD activities of the proposed project. In the "other direct costs" sub-category for beneficiaries CNR, NKUA, CERN, ENG, BDM-USTRATH, FAO, FIN and Terradue, a travel allocation appears for participation in project meetings or events related to undertaking the Joint Research Activities.

Coordination: The personnel costs within this category are directly related to fostering a culture of co-operation within and beyond the D4Science-II consortium, allowing for the identification of interoperability needs of the repositories and infrastructures participating in the project, and by extension the scientific communities that use them, as described within the Networking Activities' work packages. A total of 171 person-months have been dedicated to the proposed project's coordination activities. Coordination activities include the scientific and technical coordination of D4Science; dissemination, communication and training; requirements gathering and solutions, including standardization; and supporting adoption by the various scientific user communities. In the "other direct costs" sub-category, ERCIM (by supporting W3C), CNR, NKUA, CERN, ENG, FAO and Terradue show a proportion of their travel allocation for the participation in project meetings or events related to the undertaking of the coordination activities (i.e., Networking Activities excluding project management).

It is important to note that ERCIM, CNR and ENG use either the *real indirect cost* or *simplified* method for calculating indirect costs. Thus, the "indirect costs" category for these partners show the real indirect costs that will be applied as additional resources to the project. However, only 7% of the direct eligible costs may be requested as Community contribution within this category. This 7% of the direct eligible costs is the amount shown as "indirect costs" for partners participating in Networking Activities while using the *transitional flat rate* method for calculating indirect costs.

Support: There are no personnel costs allocated to this category. The costs related to supporting the participation of External Advisory Board (EAB) members to annual meetings with the Project Management Board will be managed by ERCIM (shown in the sub-category "other direct costs"). The costs include an allocation of 2.500 € per year in travel and subsistence for members who will be travelling from outside of the EU (i.e., Christine Borgman, Lucille Nowell, Edward Vanden Berghe) and 1.500 € per year for the travel and subsistence of those members domiciled in Europe (i.e., Fabrizio Gagliardi, Erwin Laure, Serge Michel Garcia, Per Öster). The total allocation for the External Advisory Board is 22.000 € for the duration of the project. Overhead is charged at 7% for the administration of the reimbursement requests.

Note that 40.000 € in **subcontracting** costs have been requested as some of the scenarios envisaged as part of the ecosystem will require that the systems of a few participating organizations be upgraded in order to meet infrastructure standard specifications (e.g., WSDL compliant web services). Other types of system developments (e.g., development of data models, implementation of standards) will be necessary for the implementation of user interfaces of specific products, such as the ICIS query and reporting system. It is difficult at this stage to anticipate which scenarios will require these particular development activities, so a financial provision is being made at the coordinating unit level (i.e., ERCIM) in order to support in a flexible way the scenarios according to actual needs.

Management: The personnel costs within this category are directly concerned with the overall legal, ethical, financial and administrative management of D4Science-II; quality assurance; and exploitation and sustainability, as described in NA1 Project Management within Networking Activities. A total of 24 person-months have been dedicated to the proposed project's management activities. ERCIM, as Coordinator, will consume 12 of these person months for project administration alone. CNR will consume 1 person-month for project administration, related to the scientific contributions required in the periodic and final reports. 4 person-months in this category are attributed to Quality Assurance activities: CNR-ISTI, CERN and NKUA representatives compose the Quality Assurance Task Force for monitoring the project's quality procedures and metrics. The remaining 7 person-months are dedicated to the development of an Action Plan for exploitation and sustainability, an activity led by the Coordinator.

Under "other direct costs" appears a proportion of the travel allocation for ERCIM, CNR, NKUA and CERN. Additionally, ERCIM has been allocated 8.000 € for administering a travel budget that will provide for the participation of members of the Earth Observation (EO) community to D4Science-II meetings and events. The ImpECt VRE developed in the context of D4Science will be completely operational and used by the EO community, so it is intended that the ImpECt VRE will be maintained throughout the duration of this project and beyond.

According to the rules on reporting and payments outlined in Section 11.4 of the Annex II General Conditions to the FP7 Grant Agreement, only partners requesting Community contribution equal or superior to 375.000 € will be required to submit a certificate on the financial statement for claims on final payments in a project with a duration of 2 years or less. In the case of D4Science-II, this means that CNR, NKUA, CERN and FAO will be required to submit a final certificate on the financial statement. 1.000 € have been assigned for this purpose to CNR and NKUA as sub-contracting costs. CERN and FAO have certified internal auditors and do not need any funding for the production of the certificate.

Note that all management activities, including quality assurance and sustainability, account for **6,9%** of the total requested EC contribution.

Other: The personnel costs within this category are directly related to the deployment and provision of the continuous and effective operation of the resources shared through the D4Science-II Ecosystem while enabling their exploitation by the user communities, as described within Service Activities. A total of 141 person-months have been dedicated to the service activities of the proposed project. In the "other direct costs" sub-category, CNR, CERN, ENG and FAO show a proportion of their travel allocation for the participation in project meetings or events related to service activities, while 4D SOFT shows the totality of their travel allocation as this partner has their entire person-month allocation in this category.

Additional resources: CERN, as project beneficiary and through INSPIRE, will contribute additional unfunded resources to D4Science-II: Approximately 6 person-months of "Senior" staff (i.e., Salvatore Mele, Tim Smith and Jens Vigen; please see short biographies in Section 2.2.); and approximately 12 person-months of technical development. It is also expected that US collaborators involved in INSPIRE (Stanford, California and Fermilab, Illinois) will also contribute approximately 2-3 unfunded person-months for consulting on the specifications of the API and bibliometric algorithms that will be developed and run on the grid in the INSPIRE scenario, through the D4Science-II infrastructure.

During the lifetime of the project the D4Science-II e-Infrastructure will be established and extended with new sites. Please see Appendix A for a detailed description of the resources that will be federated by the diverse communities and the D4Science-II consortium, and maintained for the duration of the project.

Subcontracting of minor activities not directly associated with the work undertaken by any one beneficiary, yet having a direct impact on the demonstrability of project results is currently foreseen, as described under the "Support" category above. The reserved amount equals 40.000 € and will be managed by ERCIM.

Additionally, the production of certificates on the financial statements (estimated at 1.000 € each for CNR and UOA) is scheduled for financial closure of the project.

Section 3: Impact

3.1 Expected impacts listed in the work programme

3.1.1 Expected Impact

3.1.1.1 Technology

The core technological objective of D4Science-II is the realisation of an *open*, *versatile and extensible knowledge ecosystem* based on solid specifications and proven good development practices. Once fully operational, this ecosystem is expected to strengthen salient interoperation concepts that would lead to a unified scientific data and knowledge e-Infrastructure. D4Science-II will provide the means to empower such a unifying approach in the emerging technological and scientific landscape. Our emphasis on utilizing WS-data-related specifications provides fertile ground for creating a standard substrate for the flow and processing of information among heterogeneous stakeholders and infrastructures – all done without compromising their autonomy nor their technological investment. Through its technological vision as well as its targeted and scenario-specific activities, D4Science-II shall generate strong bindings between stakeholders, foster long-term relationships, and ultimately seek to bring together independent entities into a interoperable consortium of e-Infrastructures. In addition, such interoperable consortium will offer clear benefits to the stakeholders, further attracting other e-Infrastructures serving a large spectrum of scientific communities.

The main D4Science-II objective of building a common layer operationally-placed at the boundaries of independent e-Infrastructures along with bringing new value-added features renders the project attractive to a number of players as evidenced by the integration scenarios discussed in this proposal. In this regard, D4Science-II will enable previously *un-related infrastructures to "consume" new data and information sources, as well as new algorithms and process to peruse both*. Subsequently, D4Science-II will *reach out to a very large aggregate audience*. Catering for each piece of this audience individually would be simply impossible and would imply unsustainable implementation costs. Furthermore, members of the D4Science-II *knowledge ecosystem may provide new services to consumers by exploiting the storage and computational capacities brought forward and made available by the collective power of all e-Infrastructures involved.*

The proposal builds on both explicit and implicit synergies that exist among e-Infrastructure domain stakeholders and its impact will be multiplied due to the following facts:

- E-Infrastructures are evidently strengthened and their effect is substantially enhanced by the inclusion of more physical resources, repositories, and standardised services.
- Additional resources lead to improved data and resource availability, reduction of costs through the creation of economy of scale, and multiple opportunities for innovation that in combination attract communities and individuals for conducting their scientific work.
- As users become aware of both the utility and potential of the ecosystem through the scenarios used, the produced technology will inevitably enjoy faster and wider adoption. In turn, this will help spur richer services using more diversified and better interoperating data resources.

D4Science-II aspires to not only boost the *adoption and use of large-scale, federated, interoperable digital repositories* through the elimination of interoperability barriers but also to *open new horizons to the scientific communities by allowing new exploitation perspectives for data/information at various stages of their existence*, i.e. from unprocessed data up to derived knowledge. The introduced capacity for making available and processing massive data acquired from diverse scientific and technological domains gives unprecedented *opportunities for exploitation of information* in ways that previously was either impossible or even not considered.

Building on the long term vision of the Grid, the widely adopted standards of OGSA, and the vast experience in processing data at large by its predecessors (DILIGENT, D4Science and other collaborating projects including INSPIRE, ETICS 2, DRIVER and Health-e-Child), D4Science-II foresees the creation of a robust knowledge ecosystem with tools and procedures that safeguard and self-manage all its salient assets: *data, physical/soft resources, knowledge* and *users*. A mature set of specifications, spanning from validation rules to access policies and from protocols to applied technologies is at the disposal of the Infrastructure users and it is envisaged that they will collectively yield production-level quality of service(s).

By increasing the volume and enhancing diversity of its data, the proposed knowledge ecosystem shall benefit the quality of the information and the knowledge produced. We also expect the enhanced quality of the rendered services will become the springboard for shaking off old procedures of inquiry and establish new and effective workflows for conducting scientific analyses. To this end, it is anticipated that users rapidly gain confidence for their results and outcomes obtained while working atop the VREs partaking in the proposed ecosystem. Along these lines, a critical side-effect of our proposal is that raw and/or "second-level" data will ultimately come closer to their information/knowledge products; this closer association may influence high-level by-products of inquiry including publications and even perhaps modelling of human behaviour. Therefore, a more comprehensive, provenance-based, reliable and productive scientific environment will be furnished where disambiguation can be readily resolved, access to data for validation will be permanently within-reach and necessary curation steps can be taken over time.

The vast processing, storage and network capacity brought forward along with the withdrawal of technological barriers in terms of standard-based data/information exchange and reference will offer significant opportunities for R&D activities in the new fields of data curation and domain preservation. Data can be stored and moved around in large volumes and potentially in several flavours and stages of their lifetime. The latter are all identified as critical aspects for long term preservation. The offered enhanced availability of "source" or "derived" scientific data, the capability to resolve access control issues in a federated manner and the provision of integrated processing constitute fundamental blocks for the realization of a knowledge ecosystem that is both curation- and rich-service-enabled.

Bringing together resources from several communities, in a multidisciplinary landscape, *impacts the operation of the involved groups* both directly and indirectly as it is evidenced by the adopted interoperability scenarios:

- The large number of resources garnered inside and around the constituent e-Infrastructures of the ecosystem form a vast pool of reusable or collaboratively exploitable entities, be it physical (processors, storage, network) or conceptual ones (software, services, data etc.). This economy of scale, along with the adoption of open specifications, reduces the cost of ownership for the involved parties due to the following reasons:
  - Standalone physical resources tend to be severely underutilized; actually this fact is what drives the modern hardware virtualisation trend and the Internet paradigm today;
  - The software and service reuse economy model of implementing-once-using-many-times can be readily adopted.
  - The fact that robust open and standardised specifications reduce the cost of interoperability across domain boundaries does apply.
  - Unified access to data and knowledge through federation and reuse reduces both direct and indirect costs of operations as it increases stability, helps with training and attains overall better management.
  - Offering the possibility of doing more science and bringing more novel services to any single community that otherwise would be financially impossible.
- Bridges between multidisciplinary communities are formed by bringing a common denominator into the scene and allowing cross-fertilisation of technology and scientific results in a number of directions:
  - Harnessing data under new perspectives, in a cross-domain joint-processing model, where new knowledge can be derived after taking into consideration previously unexploited data.
  - Porting existing data and service interoperability techniques that have been successfully applied in various domains to new scientific fields.
  - Implementing new techniques of data exploitation that are derived either from the availability of new sources or by the creation of new cross-domain aggregations of data.

- Visibility of communities as well as exploitation of their work and produced artefacts, are enhanced by having their information sources, logical resources and concepts "join" the ecosystem. Both visibility and work exploitation shall impact positively a number of communities.

Safeguarding data by policies, procedures and the appropriate technology does help avert any potential negative impact. By and large, this has been the experience in D4Science so far.

Bringing a large number of multidisciplinary communities under the concept of interoperability raises major challenges for specifications and technology choices that have to be made. Evolution of concepts, technologies and policies is pushed by several aspects of the D4Science-II vision that among others include:

- Interoperation of adopted technologies;
- Interoperability and control of data;
- Provision of domain boundary crossings for both data and software;
- End-user access and resource exploitation policies;
- Change and configuration management while the e-Infrastructure is in operation;
- Adoption of new workflows and procedures for organizations functioning in the D4Science-II ecosystem.

Lastly, the anticipated evolution and adoption of new scientific ethics and policies which will have to be put in place as a response to an initiative such as D4Science-II is expected to be crucial. The exchange of unprocessed (raw) data, the added capabilities to validate the flow of information and provenance at various levels (starting from raw data and going all the way up to the level of derived knowledge) and the abolishment of barriers that have been instrumental in the formation of isolated scientific cores create a emerging scientific frontline to which involved communities have to react and very likely adapt.

3.1.1.2 Community / Infrastructure Specific Impact

DRIVER

Enabling interoperability in D4Science-II with the DRIVER storage substrate will *permit the integration of the Grid with one of the largest under-construction federated digital repositories infrastructures* in Europe. This will have a bidirectional impact: firstly, the scientific community will have immediate access to the large DRIVER information space and at the same time will be capable of exploiting DRIVER artefacts either as first-level product(s) or as federation component(s). Secondly, the DRIVER community will have unhindered access to vast Grid computational, network and storage resources and by doing so it may be empowered to furnish new capabilities and services to its users. Data can be processed on-demand for building "higher-level" products (derivatives). More importantly the emerging concept of the *Enhanced Publication* – a document outlining research performed along with associated links and pointers to data-sets utilized, software packages used and pertinent detailed scripts and procedures followed to obtain discussed results – shall be boosted by the D4Science-II ability to rely on artefacts that can be long-term preserved, curated, and provenance-enabled. DRIVER may receive an infusion of data from other federated collections such as the citations of INSIRE and develop new personalization options to its users. The quality of services offered can be dramatically improve at minimal costs since the overheads of managing this computation / storage substrate comes at essentially no essentially cost. Furthermore, the implied synergy for building common models on top of existing specifications for referencing and data access empowers the development of models for handling of complex documents and positively impacts the creation of enhanced publications.

INSPIRE

INSPIRE aims to achieve computationally intensive bibliometric analysis of scientific publications in the field of High-Energy Physics (HEP) beyond the capacity usually available to repositories. Through its interoperation with D4Science-II, INSPIRE will be able to offer advanced data mining services to its end-user. These applications will allow the identification of network of authors, the suggestion of experts in particular subfield through the analysis of citation history and novel approaches to the assessment of the impact of a scientific result, group or individual through hybrid metrics. It is worth pointing out that these applications are not limited to the content of the INSPIRE repository, High-Energy Physics. The approach in question can be clearly exported to any other repository as a form of added value built atop pertinent records. This will positively influence other D4Science-II participating e-Infrastructures as their service-quality will be improved at no cost, should common standards for interoperability be adopted. DRIVER is to also benefit a great deal as it can adopt INSPIRE-based data mining algorithms. To this effect, DRIVER shall further advance the adoption of digital repositories by authors, researchers, policy-makers as well as institutions and organizations.

AQUAMAPS

The main objective of AquaMaps is to provide state-of-the-art visualization-based models for marine species. To this effect, the effort allows the biodiversity community to establish baselines for species geographic distributions and predict either growth or decline of populations within regional ecological envelopes according to transparent algorithms directly used by aquaculture experts. Currently, the biodiversity community is heavily using maps that are constrained in terms of the physico-chemical parameters they use. As a D4Science-II e-Infrastructure, AquaMaps will harness expertise and information currently found in other e-Infrastructures to take the much needed leap forward in enhancing its produced maps. AquaMaps will initially *draw more voluminous and rich data* from interoperable sources; such data are furnished by services offered by the GENESI-DR, OBIS, NEON and GBIF projects. Moreover, AquaMaps will exploit the computational capability of the Grid to generate *more precise, accurate, synthetic and predictive maps.* The ability to synthesize earth observation data and other environmental sources along with new data assimilation and access techniques will yield tremendous opportunities of enhancing species maps. By joining D4Science-II, AquaMaps will facilitate enhanced monitoring of climate change and its impact on aquatic populations distribution and abundance, and on aquatic ecosystems. Improved aquatic cartography will also benefit ICIS in better re-allocating catches. Last but not least, the added interoperability features will make it much easier for other e-Infrastructures to take advantage of the predictive data generated by the AquaMaps VRE which is, of course, not an option at this time.

FCPPS

Fisheries Country Profiles are one of the most widely solicited data products from FAO Fisheries department and regional players, consistently ranking at the top of delivered pages in FAO's portal. They are also one of the most cumbersome products to generate as they involve a multitude of actors at national, regional and global levels and include a wide array of poorly connected data sources. The outcome of this process is reduced coverage and infrequent updates. In turn, this results to Fisheries department being at times unable to accommodate the requests by decision-makers for high-quality, accurate, frequently-updated reports with good coverage. The extension of FCPPS to a more complete set of needed data sources that are found in the D4Science-II ecosystem, together with strong semantics and an organized publishing pipeline will enable FAO to meet the needs of its users for one of its most sought-after products. In addition, the current system suffers from problems of interoperability with both corporate level departments and other UN organizations that need to include fisheries data in higher-level multi-domain profiles. D4ScienceII's emphasis on interoperability and shared standards will bring these data streams within much easier reach of its partners.

ICIS

By connecting and harmonising regional fish catch statistics and enhancing them using techniques of catch re-allocation based on species occurrence and physical ocean data, FAO will be able to report to the UN General Assembly, the Council on Fisheries (COFI) and the Coordinating Working Parties (CWP) that their requests for better Fish stocks monitoring and assessment tools are being addressed and backed with transparent products and operational mechanisms. By connecting this achievement with other "by-products" such as the enhanced ability to monitor the impacts of climate change using species distribution modelling software (AquaMaps), D4Science-II will demonstrate its sustained value among the software development and user community.

The improvements targeted during this second phase of implementation will respond to the D4Science partners' wishes expressed during the June 2008 ICIS requirements assembly and address issues expressed by the broader CWP community during its July 2008 meeting. This should make potential fishery agency contributors comfortable with the concept of collaborative VREs and warrant their full participation. It is even expected that by touching upon the benefits of improved workflow for science, these agencies will fully adopt D4Science as a framework for data sharing, analysis and output. For example, the critical scientific revision of official fishery statistics preceding regional resources assessment working groups (some 50 different agencies) would be a net beneficiary of the tools developed through the project. Limited interoperability is an important point often raised by D4Science partners as they are interested in exploiting data not just within a specific VRE, but also from outside and often from within their own existing legacy systems.

The integration of Vessel Monitoring System (VMS) catch data in the infrastructure will accompany and provide incentive for speeding-up the development of information standards for data exchange among Regional Fisheries Management Organisations (RFMOs) and national agencies, based on accepted data security mechanisms. Such data sharing efforts are anticipated to have enormous potential repercussions on the improvement of assessment models which will be able to take into account finer time-space scale levels. The derived assessment models will address the need for jointly taking into account ecological behaviour of aquatic resources and fisheries monitoring while relying on near-real time data offering a more dynamic and timely view of fishing areas.

Ultimately, the infrastructure and associated toolkit(s) would become available for use by fisheries ministries at national level in order to consolidate various sources of monitoring information for both analysis and policy decision-making.

GENESI-DR

GENESI-DR currently operates the integrated access to Earth Science Digital Repository allowing access to data, information, products and knowledge originating from space, airborne and in-situ sensors. By establishing interoperability bridges with D4Science-II, it will be possible to share the access to the GENESI-DR federated resources (catalogue functions, data access, processing services etc.) that adhere to a common set of standards, data policies and interfaces. D4Science-II participating infrastructures will be provided a virtual collection of digital Earth Science data, irrespectively of their location in the various single federated repositories framed by the GENESI-DR data policy. GENESI-DR builds upon the existing, operational and focused Earth Observation (EO) European infrastructure. Because it involves key Earth Science centres responsible for operational data acquisition, processing, archiving and distribution it harmonises operations at key Earth Science data repositories limiting fragmentation of solutions. It also provides guaranteed, reliable, easy, effective, and operational access to a variety of data sources, and demonstrates how the same approach can be extended to provide access to all Earth Science data.

GENESI-DR has also the objective to validate the effective capabilities required to access distributed repositories for new communities and to integrate new scientific and technological derived paradigms in operational infrastructures in responds to the latest Earth Science requirements. As such, the GENESI-DR digital repositories and the Earth Science community can take advantage of the amplified D4Science-II capability of data personalization from various other federated collections. Together with the ability to long-term preservation, not only information formerly unavailable will become rapidly available for processing but also the opportunity for not-disclosed-before research will become viable. In this context, human-activity data meshed with statistics that do not emanate from Earth Observation sources can be combined for modelling complex ecosystems, for better prediction and planning.

3.1.2 Steps to the Impact

Achieving the D4Science-II foreseen impact is by itself challenging and spans several domains, starting from policies and ending with technology. To attain this, we intend to pursue a stepwise approach as follows:

Act on well established cases with strong technological footprint

The selection of the interoperability scenarios of the project's workplan has been carried out while adopting the long term vision and needs of the respective stakeholders. It is important to clarify that the selected scenarios do demonstrate both e-Infrastructures and major scientific communities behind these e-Infrastructures.

The DRIVER infrastructure builds a long term federated information space that continuously expands its vision, and attracts the interest of a large number of scientific repository holders around Europe. INSPIRE offers a voluminous repository with high quality information and sophisticated data mining services in the area of High-Energy Physics (HEP). Managed by the CERN Library INSPIRE aims to offer sophisticated data mining processing on its well established repository content. Improved AquaMaps in conjunction with ICIS and FCCPS scenarios will serve FAO, FIN, regional/local fish and aquaculture communities counting thousands of scientists, researchers and decision makers; they will also help with the introduction of a large set of tools. All the above e-Infrastructures constitute self-sustainable "systems" on their own. Finally, the GENESIS-DR serves the Earth Observation Sciences communities that are technology savvy and up to date.

The involvement with scenarios featuring strong technological footprints assists in many ways as:

- Domain expertise is the pre-requisite for carrying out the proposed task in a solid manner and staying focused on the specific objectives of each scenario.
- Direct exploitation of the artefacts of the partners will have an amplifying effect at all levels (i.e., the specific scenario, the entire project, the e-Infrastructure used, and finally the science).
- Concrete and continuous validation of the vision and the progress of work can be attained.

Identify the core problems at all dimensions

It is well established – partially by our own experience with D4Science – that the problem of interoperability cannot be addressed exclusively through resolution of technology factors only. Domain boundaries have to be crossed and operational aspects of the infrastructures have to be gracefully resolved so that new potential can be unleashed. Thus capacities/capabilities, policies and perspective(s) of each scenario have to be well understood in conjunction with the technological dimension, which includes, data, resources and specific assets by underlying D4Science-II-partaking infrastructures.

Build on standards

Exploiting well-established specifications for enabling interoperability among disjoint domains and/or infrastructures is the *a priori* best known practice to be followed. Nevertheless complex cases, such as the ones the proposal focuses on are usually not well captured by one and only specification. Composition and in some cases extensions are required and there is a continuous need for monitoring and validation so that the adoptions/extensions remain within the commonly acceptable limits. Hence, a great deal of success for D4Science-II will emanate from the successful adoption and application of existing and/or emerging standards. In this regard, the partners have already established contacts with a number of international bodies that play major roles in the effort of developing standards including W3C, ETSI and OGF.

Build strong synergies and on strong, proven technology

In order to cross the boundaries of domains and infrastructures, it is required that an initial substantial effort is put forward. This is essential so that strong synergies ensue at both policy and technology level. More specifically, we consider creating policy synergies as a key enabler to the success of D4Science-II.

Aiming to provide a production-level interoperable knowledge ecosystem that must support and not be simply compromised of functioning together systems requires the adoption of proven technology.

Ensure the long term existence of resources

Interoperability at such scale can only be beneficiary with a long term vision of maintenance and evolution. This depends not only on the long term vision of the Grid already in place and maintained by national initiatives around Europe through the EGI-DS and NeOn initiatives. More importantly, the international communities the proposal already engages along with other potential trans-national and regional scientific players attracted to our ecosystem will help maintain the system and the entire e-Infrastructure over time.

Face scientific challenges

As D4Science-II focuses on the domain of data/information/knowledge management and interoperability, it exposes a number of scientific challenges which have to be effectively dealt with. Both the strong scientific team of the project and the very strong synergies and relationships with communities acting as domain experts ensure the flow of information, use of best practices and the cross-validation of scientific artefacts in compliance with the stated project objectives.

Address technological challenges

Similarly a number of technological challenges have to be addressed, several of which cross the domain of infrastructures and disciplines. Our substantial and well-versed team of designers, implementers and testers who jointly have a number of years of collaborative experience is well qualified to address the above challenges.

Build knowledge poles, train and promote

As a last but not least important step of the project impact is the dissemination of the artefacts of our joint scientific and technological work. Scientific knowledge has to be disseminated, beneficiaries have to be trained so as to best exploit the products, and the constructs, software, services, resources, have to be promoted outside the boundaries for reuse and exploitation. The latter is expected to reach communities far beyond the scope of those found in the D4Science-II scenarios and so, it will contribute to the sustainability of the ecosystem.

3.1.3 Pan-European Scale Activities

There are a number of reasons suggesting that D4Science-II has to display a Europe-wide penetration rather than remain in a more localized/national level.

The endeavour of creating interoperable infrastructures has a real impact only when carried out at international level. By their nature, local models tend to be more restricted as far as the challenges they present. Existing common and/or national policies (i.e. stemming from legal or ethical background) or other "hidden" rules that indirectly affect the interoperability even technologically, can inadvertently limit the applicability and the implementation scope of the project.

It is also essential that the benefits of interoperable infrastructures *reach multiple players across the continent* for the project to fully materialize its impact. At an international level, more practices are brought forward and a larger pool of resources becomes immediately exploitable. Especially for the scientific e-Infrastructure, *international interoperability is a fundamental requirement* as the communities themselves are international. Within D4Science-II, all the scenarios examined span several stakeholders and countries, even if they are multi or single community cases. For instance, INSPIRE already has an major US base in its operation and DRIVER is an infrastructure effort that involves storage sites from across the continent. Bringing a wide breadth and diversity of entities emanating from *several domains of business, operational environments and countries across the globe and forming a large ecosystem, we inherently increase the possibilities of achieving the impact of the project*. Actually, the impact will cross European boundaries, helping to achieve excellence worldwide.

Finally, it is important to indicate that there are numerous challenges in several technological and scientific areas for bringing forward such a research and development agenda. The aim to excellence requires that scientific and technological resources have to be garnered beyond those found within local and national boundaries so that both the necessary capabilities for carrying out the project tasks and the critical mass for building a powerful Task Force are summoned.

3.1.4 Assumptions, Risks and Dependencies

The D4Science-II workplan makes several assumptions including reliance on external inputs which, if not verified or available on time may jeopardize the course of the project. In order to anticipate and timely react to potential risks including the most likely ones identified below the consortium will establish a strong Risk Management strategy as part of the project definition.

D4Science-II heavily depends on the artefacts being produced by the D4Science project. At the time when D4Science-II is expected to go under way, D4Science will have delivered a fully robust software platform with new features; in all likelihood, these new characteristics will be also influenced by our early D4Science-II work. Taking into account that D4Science has reached a mature stage and advanced features have been already implemented (expanding the D4Science initial objectives) along with the strong commitment of partners, it is now very likely that D4Science-II will have even more tools implemented and at the disposal of users (compared with the set of tools available today). Regardless of concurrent activities, D4Science proceeds at this time with use of interoperability standards and creates the basis for realizing at least portions of the vision of this proposal.

In addition to D4Science deliverables, D4Science-II also assumes the availability of a number of additional infrastructures to demonstrate selected interoperability test scenarios. For some of those infrastructures, the current EU funding for the development and/or operations might end before the completion of D4Science-II. Such situations could leave the project without the expected partner motivation and more importantly the funding for having infrastructures interoperate with others within the envisaged ecosystem. We expect to overcome such adversities as the involved scientific communities have already bought into the concept, strong groups or organisational units are already in place, and they actively exploit most of the participating infrastructures. With proper information and very limited monetary support, these communities shall be still able to be part of D4Science-II and benefit from the additional tools D4Science-II is set to produce.

The scientific feasibility of endeavour is another critical factor. There are several limitations to achieve full interoperability of data in a domain-agnostic manner. Yet, new approaches under development combined with a step-by-step approach will lead the project to the maximum possible level of e-Infrastructure interoperability.

The D4Science-II technological feasibility is warranted by the demonstrated excellence of partners on the domain and through their strong synergies and relationships with pertinent task forces. The commitment and size of partners can cover up for additional needs for technological development that might occur as result of scientific developments or other unanticipated factors. It is essential to highlight that this commitment to the objectives is clearly marked by a long term investment on the fields of training, scientific development and technological progress undertaken by nearly every partner. Should potential failure of dependent technologies occur either during delivery timelines or in quality, the consortium has the ability to reroute implementation plans with alternative solutions or even undertake new activities as a countermeasure.

3.2 Dissemination and/or exploitation of project results, and management of intellectual property

3.2.1 Dissemination Measures

D4Science-II success greatly depends on how fruitful the project will be on the dissemination front. Interoperability does need to be achieved not only at the scientific / technological level but also at the asset and user level. Project outcomes will be widely publicized to gain adoption from players around the world. In turn, dissemination will increase the number of e-Infrastructures operating in the ecosystem and help substantially augment the volume of respective data/information.

The numerous challenges to be addressed, will lead to decisions that have to be validated by a multitude of diverse audiences, in order to have D4Science-II weaknesses revealed and be able to gain wide acceptance.

The importance of the project artefacts calls for very intense dissemination efforts to be undertaken. Pertinent efforts that we intend to take include:

- All project outcomes will be intensively cross-validated by domain experts, task forces, and scientists. The active engagement of our project members into designated domain-specific working groups is a dissemination activity of paramount important and a key collaboration element for the consortium.
- Technology products, specifications and scientific results will be made available through several channels so as to attract new communities and have their correctness and applicability challenged.
- Other infrastructures will be sought for examining interoperability with D4Science-II and maximizing the number of affected communities and resource pools upon the completion of the core D4Science-II concepts.
- As the interoperability scenarios to be materialized maintain long-term plans, we expect that they will attract numerous users and visitors from multiple new disciplines. The latter in turn will create an amplifying effect for the inclusion of new e-Infrastructures and their resources into the D4Science-II ecosystem. At this stage, we foresee collaborative efforts with OBIS, Regional Fishery Centers and a number of actors in the biodiversity domain.
- Due to the size and nature of several participants in the project, internal dissemination is also very critical for the success of the project. FAO, CNR, CERN, ERCIM, FIN, BDM-USTRATH and NKUA are organizations that involve hundreds (some even thousands) of individuals some of whom are working on state-of-art ICT technologies. This activity is especially critical for participating organizations that interface with user-communities as they can attract more groups to the D4Science-II knowledge ecosystem.

The consortium members will present and expose the project outcomes to the maximum extent by using traditional means such as presentations, archival journal articles, conference publications, interviews, etc., or modern avenues such as web pages, blogs, and RSS feeds. Another dissemination activity involves EGEE itself and its quarterly conferences and meetings. EGEE can also attract new communities to the infrastructure by demonstrating D4Science-II to its communities.

3.2.2 Exploitation of Results

The long term sustainability of the project artefacts and the full achievement of their anticipated impact on a wide number of communities working in High-Energy Physics, Earth Observation Sciences, Aquaculture & Marine Engineering, Management of Digital and Geographic Content as well as many more potential players, necessitates that the exploitation of both intermediate and final results is a fundamental objective.

User communities directly involved in the project maintain very promising exploitation plans as they hope to:

- Solve every day problems, share experiences and build on success stories in order to proceed with their work.
- Improve the quality and the scope of their current work and/or provided services by exploiting new sources of data and new ways for processing them.
- Obtain access to large sets of data for more sophisticated experimentation while pursuing scientific excellence.

In the course of the proposed activities, technological partners will have opportunities to benefit through:

- Experimentation with emerging technologies.
- Production level design and implementation practices and policies in a large multidisciplinary setting will benefit the group's collective expertise.
- Several project outcomes that can be exploited in cases outside the scope of the core objectives, applying only minor code re-factoring.

Educational partners, from the higher education sector, foresee direct exploitation of the project work plan:

- Training of students in applied ICT technologies and collaborative environments.
- Obtaining the means and know how for experimentation on emerging technologies, specifications and concepts in e-Infrastructures and their interoperability.
- Scientific inquiry is of major relevance to those in research institutions. Publications stemming from work carried out during the project are very promising due to the complexity of pertinent constructs and concepts. Raising awareness of the capacities of the institution in the scientific field is identified as one of the objectives of all educational organisations.

Finally a business model can be created around the provision of services on top of the infrastructure.

In the context of the new ecosystem, Europe will emerge as a winner on two fronts:

- European scientists in the aforementioned communities will be the first to take advantage of the new environment, proceed with the necessary paradigm shift in their methodologies, and reap the fruits of combining diverse information in innovative ways and of applying novel approaches to their research endeavours.
- Scientists in the rest of the world will be attracted and possibly compelled to use the suggested pan-European e-Infrastructure for their own work, thereby increasing and reinforcing the European leadership in the relevant technologies as well as the impact and global relevance of the proposed ecosystem as a unique instrument for supporting science.

3.2.3 Knowledge and Products IPR Management

Opportunities for dissemination and exploitation in D4Science-II are substantial as few restrictions apply in several domains. Although the proposed knowledge ecosystem will have the responsibility in applying policies over the data managed by it, it is evident that there are several data collections gathered and/or produced by communities and infrastructures with which Memoranda of Understanding (MOUs) will have to be signed. To enumerate some:

- A large series of rich sources related to aquatic species: NeOn Lifecycle for Networked Ontologies, Census of Marine Life (CoML), Ocean Biogeographic Information System Database, World Register of Marine Species, FishBase, ReefBase;
- Large fisheries' related resources such as Global Ocean Observing System oceanographic datasets, Vessel Monitoring System (VMS) data repositories of Regional Fisheries Management Organizations (RFMOs), the global database of world marine protected areas (MPA Global), UN statistics, including the UN Commodity Trade Statistics Database (UN Comtrade), FAOSTAT, ASFA Abstracts, AGROVOC;
- The GENESI-DR digital repositories for Earth scientists, gathered by several operation Earth Observation Centers;
- The in-excess of one hundred DRIVER-hosted repositories;
- INSPIRE High-Energy Physics Metadata Repository and its sources of metadata and full-text articles.

In the Environmental Monitoring domain the situation might be a challenging one as access to data might be allowed only under strict policies and full compliance with complex conditions. Although core data might not be directly available to users for a number of reasons, the knowledge produced by their processing may becomes public. These policies have to be materiliazed by D4Science-II technology.

In another direction, the knowledge and technology developed as part of the D4Science-II project flows freely among the partners that implement and exploit the system as part of their contractual obligations. No fees or other kinds of restrictions are applied within the project's boundaries.

Furthermore, following the paradigm of predecessors, we anticipate that D4Science-II will deliver products free of IPRs which nevertheless do not violate existing IPRs. An IPR-liberal approach will be applied when it comes to products and/or project software constructs unless they explicitly infringe IPRs and/or violate other explicit/implicit IPR ownerships. In the foreseen dissemination plan, possible IPRs will be handled in most cases in manner similar to treating cited research work.

A peculiar aspect in the straightforward IPR handling approach of D4Science-II, is the safeguarding of data and software due to the technological nature of the infrastructure. Service-Oriented Architectures and Dynamic Deployment over a highly distributed system could potentially end up delivering assets beyond the scope of access policies. Thus D4Science-II will continue to cautiously safeguard the framework's payload (software and content) on a per-case basis. The tools and specifications will be provided so that this provision is handled in a formal manner.

Finally as far as the rest of the resources are concerned, which include services, computational elements, and storage elements, their delivery is to be performed without any use-restriction to the communities targeted by the project. Communities attracted in the future are likely to be supported by the project, if adequacy of resources is in place. At the same time, the option for communities to bring in their own resources is a fallback solution.

Nevertheless, data and software licensing, in such a complex system that involves and reuses several components, aggregates multi-disciplinary sources and brings together several Infrastructures, presents in itself a challenge.

Section 4: Ethical Issues

The Coordinator confirms that no ethical issues arise in this proposal, including, but not limited to issues of informed consent, research on human embryo/foetus, privacy, research on animals, research involving developing countries, dual use and /or ICT implants.

Table 4. Ethical Issues Table

	YES	PAGE
Informed Consent		
• Does the proposal involve children?		
• Does the proposal involve patients or persons not able to give consent?		
• Does the proposal involve adult healthy volunteers?		
• Does the proposal involve Human Genetic Material?		
• Does the proposal involve Human biological samples?		
• Does the proposal involve Human data collection?		
Research on Human embryo/foetus		
• Does the proposal involve Human Embryos?		
• Does the proposal involve Human Foetal Tissue / Cells?		
• Does the proposal involve Human Embryonic Stem Cells?		
Privacy		
• Does the proposal involve processing of genetic information or personal data (eg. health, sexual lifestyle, ethnicity, political opinion, religious or philosophical conviction)		
• Does the proposal involve tracking the location or observation of people?		
Research on Animals		
• Does the proposal involve research on animals?		
• Are those animals transgenic small laboratory animals?		
• Are those animals transgenic farm animals?		
• Are those animals cloning farm animals?		
• Are those animals non-human primates?		
Research Involving Developing Countries		
• Use of local resources (genetic, animal, plant etc)		
• Benefit to local community (capacity building ie access to healthcare, education etc)		
Dual Use		
• Research having potential military / terrorist application		
ICT Implants		
• Does the proposal involve clinical trials of ICT implants?		
I CONFIRM THAT NONE OF THE ABOVE ISSUES APPLY TO MY PROPOSAL	X	

Appendix A Federated Resources

A.1 INSPIRE

The High-Energy Physics community has been building information resources for decades, giving light to several innovative solution which later spread to other fields of science. INSPIRE, the new generation of such systems, whose live beta release is planned for 2009, will replace the current SPIRES which will be switched off when the former will become fully operational. It will integrate the content of present HEP repositories and databases to host the entire body of metadata and the full-text of all open access publications in the area, past and future, including conference material. It will contain almost 1 Million records (articles, preprints, reports, conference proceedings) including references and authors and at least 1/2 Million Open Access full-text articles. In particular, it will contain the collections described below.

Table 5. High-Energy Physics Information systems and resources

CERN
Name: **INSPIRE** *Type*: Curated metadata and full-text of scientific publications. MARC21, OAI-PMH, XML for metadata and PDF, PS and other graphical format for publications. *URL*: http://hep-inspire.net *Description*: Around 1 million records comprising the last 30 years of publications in High-Energy Physics. It includes metadata, links to full-text of publications, citation as well as a knowledge base of author names, and author affiliations. It is in its alpha version and it will soon include 400'000 Open Access full-text documents with full-text search capabilities.
Other Providers
Name: **arXiv** *Type*: Metadata and full-text of scientific publications. Metadata available in OAI-PMH, full text in latex, postscript or PDF formats. *URL*: http://arXiv.org *Description*: Over 400'000 Open Access preprints in High-Energy Physics and beyond, submitted by authors since 1991.
Name: **SPIRES** *Type*: Metadata, in propertory format. *URL*: http://slac.stanford.edu/spires *Description*: Curated metadata for the last 30 years of High-Energy Physics publications, including resolving ambiguity on author names, applying authority files for institution names, extracting and cross-linking references, interlinking of publications and conference, attribution of keywords.
Name: **Commercial systems** *Type*: Metadata and full-text of scientific publications *URL*: N/A *Description*: Several paying-access databases for metadata. Publisher's websites for the access to journal articles.

A.2 GENESI-DR

In line with long-term initiatives for Earth Observation (EO) in Europe, GENESIS-DR is an infrastructure for the integration, validation, preservation, and uniform dissemination of data which originates from space, airborne, and in-situ sensors and is stored – along with related tools, models, and knowledge artifacts – in digital repositories maintained at a number of Earth Centers dispersed all over Europe. The table below shows the subset of datasets of particular interest for the development of the D4Science-II scenarios.

GENESI-DR Collections
Name: **ENVISAT Meris Level 1** *Type*: N1 binary files *URL*: http://www.genesi-dr.eu *Description*: MERIS is a programmable, medium-spectral resolution, imaging spectrometer operating in the solar reflective spectral range. The Level 1B products contain radiance measurements at top of atmosphere for the 15 MERIS bands, re-ordered, calibrated, geo-located, annotated with product confidence data, calibration data, classification flags, and environment parameters. The average file size is 540 Mbytes per orbit and over 34,000 orbits at beginning of September 2008 with an increase of 14.5 orbits per day. This represents over 18 Tbytes with a daily increase of 7.6 Gbytes.
Name: **ENVISAT MERIS Level 2** *Type*: N1 binary files *URL*: http://www.genesi-dr.eu *Description:* MERIS is a programmable, medium-spectral resolution, imaging spectrometer operating in the solar reflective spectral range. Level 2 derives geo-physical parameters, depending on the nature of the observed surface, from the MERIS Level 1B products. The Level 2 products contain different types of mixed geo-physical information according to the type of each pixel. Water pixels contain normalised water-leaving reflectance, algal pigment index I, algal pigment index II, yellow substance, suspended sediment, photosynthetically active radiation, aerosol epsilon, aerosol optical thickness and total water vapour column. Land pixels contain rayleigh-corrected reflectance, top of atmosphere vegetation index, rayleigh-corrected vegetation index, aerosol epsilon, aerosols optical thickness, total water vapour column and surface pressure. Finally, the cloud pixels contain TOA reflectance corrected for stratospheric aerosol, cloud top pressure, cloud optical thickness, cloud albedo, cloud type and total water vapour column. The average file size is 621 Mbytes per orbit and over 34,000 orbits at beginning of September 2008 with an increase of 14.5 orbits per day. This represents over 20 Tbytes with a daily increase of 9 Gbytes.
Name: **ENVISAT ASAR** *Type*: N1 binary files *URL*: http://www.genesi-dr.eu *Description:* Advanced Synthetic Aperture Radar (ASAR) is a high-resolution imaging radar, operating at C-band, featuring enhanced capability in terms of coverage, range of incidence angles, polarisation, and modes of operation and can be operated in 5 distinct Measurement Modes: Image Mode (IM); Alternating Polarisation Mode (AP); Wide Swath Mode (WS); Global Monitoring Mode (GM) and finally Wave Mode (WV). Within each mode, several different image swaths may be used. ASAR's ability to pass relatively unaffected through clouds, illuminate the Earth's surface with its own signals, and precisely measure distances makes it especially useful for the following applications: sea ice monitoring, cartography, surface deformation detection, glacier monitoring, crop production forecasting, forest cover mapping, ocean wave spectra, urban planning, coastal surveillance (erosion) and monitoring disasters such as forest fires, floods, volcanic eruptions, and oil spills. ASAR products have a variable products size according to the mode operated and the length of the scene. The sizes range between 50-1800 Mbytes per scene.
Name: **ENVISAT AATSR** *Type*: N1 binary files *URL*: http://www.genesi-dr.eu *Description:* Advanced Along-Track Scanning Radiometer (AATSR) is designed primarily to measure Sea Surface Temperature (SST), following on from ATSR-1 and ATSR-2 on board ERS-1 and ERS-2. AATSR data have a resolution of 1 km at nadir, and are derived from measurements of reflected and emitted radiation taken at the following wavelengths: 0.55 µm, 0.66 µm, 0.87 µm, 1.6 µm, 3.7 µm, 11 µm and 12 µm. AATSR products are used in several application domains such as *ocean and coast* – sea surface temperature, *atmosphere* – atmospheric chemistry (trace gases), clouds, and *land* – vegetation.

AASTR Level 1 products have an average file size of 760 Mbytes per orbit and over 34,000 orbits at beginning of September 2008 with an increase of 14.5 orbits per day. This represents over 25 Tbytes with a daily increase of 10 Gbytes.

AASTR Level 2 products have an average file size of 130 Mbytes per orbit and over 34,000 orbits at beginning of September 2008 with an increase of 14.5 orbits per day. This represents over 4.2 Tbytes with a daily increase of 1.8 Gbytes.

Name: **ENVISAT GOMOS**

Type: N1 binary files

URL: http://www.genesi-dr.eu

Description: GOMOS is a medium resolution spectrometer covering the wavelength range from 250 nm to 950 nm. The high sensitivity requirement down to 250 nm has been a significant design driver leading to an all-reflective optical system design for the UVVIS part of the spectrum and to functional pupil separation between the UVVIS and the NIR spectral regions (thus no dichroic separation of UV). GOMOS products are used in several application domains such as *atmosphere* – air quality (ozone), atmospheric chemistry (trace gases), atmospheric phenomenon (cyclone, storm, hurricane), atmospheric winds, atmospheric temperature – and *solid earth* (volcanos).

GOMOS files have an average size ranging from 14 to 23 Mbytes per occultation. There are around 400-600 occultations per 24 hours cycle.

Name: **ENVISAT SCIAMACHY**

Type: N1 binary files

URL: http://www.genesi-dr.eu

Description: SCIAMACHY is an imaging spectrometer whose primary mission objective is to perform global measurements of trace gases in the troposphere and in the stratosphere. The solar radiation transmitted, backscattered and reflected from the atmosphere is recorded at relatively high resolution (0.2 nm to 0.5 nm) over the range 240 nm to 1700 nm, and in selected regions between 2000 nm and 2400 nm. The high resolution and the wide wavelength range make it possible to detect many different trace gases despite low concentrations. The large wavelength range is also ideally suited for the detection of clouds and aerosols. Applications using SCIAMACHY data are *agriculture* – forest fires, *atmosphere* – air quality (ozone), atmospheric chemistry (trace gases), atmospheric temperature, atmospheric radiation, clouds, *solid earth* – volcanos – and *ocean and coast* – ocean colour/biology.

SCIAMACHY Level 1 products have an average file size of 205 Mbytes per orbit and over 34,000 orbits at beginning of September 2008 with an increase of 14.5 orbits per day. This represents over 6.6 Tbytes with a daily increase of 3 Gbytes.

SCIAMACHY Level 2 products have an average file size of 6 Mbytes per orbit and over 34,000 orbits at beginning of September 2008 with an increase of 14.5 orbits per day. This represents over 200 GBytes with a daily increase of 87 Mbytes.

Name: **ENVISAT MIPAS**

Type: N1 binary files

URL: http://www.genesi-dr.eu

Description: The Michelson Interferometer for Passive Atmospheric Sounding (MIPAS) is a Fourier transform spectrometer for the detection of limb emission spectra in the middle and upper atmosphere. It observes a wide spectral interval throughout the mid infrared with high spectral resolution. Operating in a wavelength range from 4.15 microns to 14.6 microns, MIPAS detects and spectrally resolves a large number of emission features of atmospheric minor constituents playing a major role in atmospheric chemistry. Due to its spectral resolution capabilities and low-noise performance, the detected features can be spectroscopically identified and used as input to suitable algorithms for extracting atmospheric concentration profiles of a number of target species. The domains where MIPAS data is used are *atmosphere* – atmospheric chemistry (trace gases), atmospheric temperature and clouds.

MIPAS Level 1 products have an average file size of 230 Mbytes per orbit and over 34,000 orbits at beginning of September 2008 with an increase of 14.5 orbits per day. This represents over 7.5 Tbytes with a daily increase of 3.3 Gbytes per day.

MIPAS Level 2 products have an average file size of 6 Mbytes per orbit and over 34,000 orbits at beginning of September 2008 with an increase of 14.5 orbits per day. This represents over 200 GBytes with a daily increase of 87 Mbytes.

Name: **ERS-1,2 ATSR**

Type: binary files

URL: http://www.genesi-dr.eu

Description: The Along Track Scanning Radiometer (ATSR) is one of the instruments aboard the European Remote Sensing Satellites ERS-1 and ERS-2. The ATSR provides important information in scientific disciplines such as oceanography, climatology and meteorology. When combined with observations of cloud top temperatures, cloud cover, haze, aerosol and total water vapour content of the atmosphere, significant improvements may be expected in the accuracy of medium range weather forecasting. Also accurate sea surface temperatures can be of use to a number of commercial users, particularly those involved in fishing and the management of fishing areas. In the field of research, potential applications of the ATSR include distinguishing thin new ice from open water, identifying surface type, and the accumulation rate of land ice.

Name: **ERS-1,2 GOME**

Type: binary files

URL: http://www.genesi-dr.eu

Description: The Global Ozone Monitoring Experiment (GOME) is a new instrument added to the original European Remote Sensing Satellite ERS-1 payload complement for the second European Remote Sensing Satellite ERS-2. GOME is a nadir-viewing spectrometer that observes solar radiation transmitted through or scattered from the Earth's atmosphere or from its surface. The recorded spectra is used to derive a detailed picture of the atmosphere's content of ozone, nitrogen dioxide, water wapor, oxygen/oxygen dimer and bromine oxide and other trace gases. The ERS-2 orbit provides global Earth coverage every three days.

Name: **ERS-1,2 SAR**

Type: binary files

URL: http://www.genesi-dr.eu

Description: The Synthetic-Aperture Radar (SAR) is one of the instruments aboard the European Remote Sensing Satellites ERS-1 and ERS-2. The main applications of SAR operating in Image Mode are Ice mapping and monitoring, Ocean and coastal areas imaging and Land imaging.

Name: **ENVISAT MERIS Level 3 Geophysical products**

Type: HDF and netCDF

URL: http://www.genesi-dr.eu

Description: Taking advantage of the level 2 MERIS data re-processing, a number of demonstration MERIS Level 3 products have been generated for the complete mission. They have been processed using the standard L2 products processed with the last processor version. The MERIS level-3 binned data products consist of the accumulation of MERIS level-2 data (MER_RR_2P) corresponding to a specific period. The MERIS level-3 binned data are specified on a global sinusoidal grid (ISIN), with a spatial resolution of 1/12° (roughly 9.277 km), i.e. 4320 bins at the equator.

The first set of products are *marine products*:

- Chlorophyll Concentration for the open ocean (case I): Chlorophyll concentration is a very convenient measure of abundance of phytoplankton biomass, which has an important role in fixing CO2 through photosynthesis;
- Water leaving radiances at 412, 443, 490, 510 and 560 nm;
- Aerosol optical thickness at 865 nm and the angstrom coefficient. The aerosol retrieval is an important step in the atmospheric correction, but it also gives an indication on the status of the atmosphere (pollution, dust storm, etc.).

The second set of products are *land products*:

- The aerosol optical thickness at 443 nm and the angstrom coefficient. The aerosol properties give information on the air quality and aerosol types over land. It can be used as well to correct the land reflectance from the atmospheric effect;
- The MERIS Global Vegetation Index, MGVI) corresponds to the Fraction of Absorbed Photosynthetically Active Radiation (FAPAR). The FAPAR value is estimated from daily MERIS spectral measurements in the blue, red and near-infrared bands acquired at the top of the atmosphere using a physically based approach.

The third set of products are *atmosphere products*:

- The water vapour over clear sky: the water vapour is the most effective greenhouse gas in the atmosphere. It influences weather and climate and is responsible for cloud development, precipitation, and modulates the atmospheric radiative energy transfer. It effects weather and climate and is responsible for cloud development, precipitation, and modulates the atmospheric radiative energy transfer. Therefore it influences the energy balance of the earth and, in turn, also effects the temperature and circulation of the earth-atmosphere system.

In addition, two other quality products have been processed:

- The aerosol optical thickness at 550 nm both over land and sea;
- The repartition of the ABSO_DUST flag. The ABSO_DUST flag represents the Dust-like absorbing aerosol selected for atmosphere correction.

There are 17 different monthly MERIS Level 3 products with an average size of 45 Mbytes representing today some 60 Gbytes. The monthly increase is 765 Mbytes.

These products are also generated daily with a 4.6 km resolution. Each daily Level 3 product has an average size of 9.4 Mbytes. These products are available for 2007 and 2008.

Name: **ENVISAT AATSR Level 3 Geophysical products**

Type: HDF

URL: http://www.genesi-dr.eu

Description: For AASTR, global Sea Surface Temperature Demonstration Level 3 products have been generated. The sea surface temperature (SST) files and plots are monthly means of the AATSR 30 arcmin and 10 arcmin spatially averaged dual-view SST product. The average is weighted according to the number of full resolution measurements that were used to make up each AATSR SST. Envisat/AATSR and its predecessors ERS-1/ATSR-1 and ERS-2/ATSR-2 are about to complete 15 years of Sea Surface Temperature measurements with the accuracy required for climate research. In anticipation and as a demonstration, global SST Level 3 products from AATSR from September 2002 to April 2004 have been generated at two spatial resolutions, 10 arcminutes and 30 arcminutes.

The average size of the 10 arcminutes product is 44 Mbytes while for the 30 arcminutes product it's 3Mbytes. These products have been generated for 2002 to 2005 representing 1.5 Gbytes.

A.3 DRIVER

DRIVER infrastructure consists of a number of heterogeneous resources, among which the most important classes are repositories of documents, instantiated software (in the form of services) and physical ICT resources, i.e. computing, storage and networking ones. These resources span a wide European geographical area, within and outside the project's consortium.

At the time of writing this proposal DRIVER federates 130 repositories for a total of 700,554 documents gathered from 21 countries. Some of them are briefly described below. The complete list of these repositories can be found on the project's web site [22].

Main DRIVER Repositories
Name: **DSPACE@Cambridge** *Type*: Digital Content Repository.

URL: http://www.dspace.cam.ac.uk/ *Description*: DSpace@Cambridge is the institutional repository of the University of Cambridge. The repository was established in 2003 to facilitate the deposit of digital content of a scholarly or heritage nature, allowing employees and their departments at the University to share and preserve this content in a managed environment. The repository contains more than 170,000 documents.
Name: **HAL - Hyper Article on Line** *Type*: Digital Content Repository. *URL*: http://hal.archives-ouvertes.fr/ *Description*: HAL is a tool for direct scientific communication between academics. A text posted to HAL is expected to describe completed research work and to meet current scientific requirements in the field concerned, i.e., the content should be comparable to that of a paper that an investigator might submit for publication in a peer-reviewed scientific journal, conference proceedings, etc. The repository includes approximately 106,000 documents with full text.
Name: **Archimer, Archive Institutionnelle de l'Ifremer** *Type*: Digital Content Repository. *URL*: http://www.ifremer.fr/docelec/ *Description*: A database of full-text documents covering post-publications (final drafts published in peer-reviewed journals), theses, conference proceedings and Ifremer reports. These scientific publications can be read online, free of charge, by any potential user with access to the Internet. It contains more than 3,500 full documents.
Name: **EPrints Complutense** *Type*: Digital Content Repository. *URL*: http://eprints.ucm.es/ *Description*: E-Prints Complutense is an open access repository of the research literature of the Complutense University of Madrid, containing among others Articles, Book Sections, Conference or Workshop artefacts, Monographs, theses, Working Papers and Technical Reports. It contains more than 3,500 full documents.
Name: **ETH E-Collection** *Type*: Digital Content Repository. *URL*: http://e-collection.ethbib.ethz.ch/ *Description*: The ETH E-Collection is the Institutional Repository of the ETH Zurich. It serves as platform for the ETH Zurich publications, has archival functions and searching features to retrieve relevant research results. The open access platform, a service provided by the ETH-Bibliothek, comprises more than 10,000 single documents.
Name: **Fraunhofer-ePrints** *Type*: Digital Content Repository. *URL*: http://publica.fraunhofer.de/starweb/ep08/en/index.htm *Description*: Fraunhofer-ePrints is the official institutional repository of the Fraunhofer-Gesellschaft. Publications of the researchers of FhG institutes are available for public access in this repository which contains more than 3,300 documents.

A.4 AQUAMAPS

AquaMaps is a service for generating model-based, large-scale predictions of currently known natural occurrence of marine species. Models are constructed from estimates of the environmental tolerance of a given species with respect to depth, salinity, temperature, primary productivity, and its association with sea ice or coastal areas.

AquaMaps provides the maps shown in FishBase [27], the most widely used biological information system with circa 1.3 million visitors per month. Other species Information Systems using AquaMaps are

SeaLifeBase [64] and GBIF [29], the global initiative to give access to all occurrence points of all species, with respective interpretative tools[16].

Main Aquamaps data sources
Name: **C-Squares** *Type*: maps *URL*: http://www.marine.csiro.au/csquares/ *Description*: C-squares is a system for storage, querying, display, and exchange of "spatial data" locations and extents in a simple, text-based, human- and machine- readable format. It uses numbered (coded) squares on the earth's surface measured in degrees (or fractions of degrees) of latitude and longitude as fundamental units of spatial information, which can then be quoted as single squares (similar to a "global postcode") in which one or more data points are located, or be built up into strings of codes to represent a wide variety of shapes and sizes of spatial data "footprints".
Name: **Ocean Biogeographic Information System (OBIS)** *Type*: statistical *URL*: http://www.iobis.org/ *Description*: OBIS is an international information system focused on marine biodiversity. It provides expert geo-referenced data (biological, physical, and chemical) on marine species and currently contains more than 8.7 million georeferenced, accurately identified species records from more than 70 databases.
Name: **NOAA marine data** *Type*: satellite and sensor data *URL*: http://www.nesdis.noaa.gov/datainfo.html *Description*: NESDIS operates NOAA's National Data Centers for Climate, Geophysics, Oceans, and Coasts. Through these Data Centers and other centers of data, NESDIS provides and ensures timely access to global environmental data from satellites and other sources, provides information services, and develops science products.
Name: **Global Biodiversity Information Facility (GBIF)** *Type*: text and statistical, web services *URL*: http://data.gbif.org *Description*: GBIF is an international organisation that is working to make the world's biodiversity data accessible everywhere in the world. GBIF and its many partners work to mobilise the data, and to improve search mechanisms, data and metadata standards, web services, and the other components of an Internet-based information infrastructure for biodiversity.
Name: **Algaebase** *Type*: factsheets *URL*: http://www.algaebase.org/index.lasso *Description*: AlgaeBase is a database of information on algae that includes terrestrial, marine and freshwater organisms. At present, the data for the marine algae, particularly seaweeds, are the most complete.

A.5 D4Science

The D4Science is a production e-Infrastructure that by relying on EGEE address the needs of scientific communities affiliated with the broad disciplines of Environmental Monitoring (EM) and Fisheries and Aquaculture Resources Management (FARM).

[16] The technology has been implemented but the service has not been released yet.

These communities are provided with a portal through which practitioners from these fields can define and access various Virtual Research Environments giving organized and seamless access to the resources they use in their daily activity.

Environmental Monitoring (EM)
Name: **ESA Portal on Earth Observation** *Type*: Web service, data, metadata, pictures *URL*: http://www.eoportal.org/ *Description*: By giving access to a large variety of information and services, eoPortal aims to provide a single access point for Earth Observation (EO) information and services including satellite imagery, a directory to locate data and resources, direct access to earth-observing satellite data as well as map servers and cartographic resources. Relevant sub-URLs are the following: • http://directory.eoportal.org/dir_Document.html • http://images.eoportal.org/ • http://maps.eoportal.org/
Name: **ESA Earth Images** *Type*: Data, metadata, pictures *URL*: http://earth.esa.int/earthimages/ *Description*: Earth images collected and distributed by ESA. They are acquired from different instruments on-board of several Earth Observation satellites.
Name: **ESA Multimedia Gallery** *Type*: Data, metadata, pictures *URL*: http://www.esa.int/esa-mmg/mmg.pl?type=I&collection=Observing%20the%20Earth *Description*: The ESA Multimedia Gallery displays the best visual material of the ESA portal. A selection of Earth Observation related material with scientific captions, links and high-resolution formats is available for the project.
Name: **ESA MIRAVI Images** *Type*: Data, metadata, pictures *URL*: http://miravi.eo.esa.int/en *Description*: MIRAVI stands for MERIS Images RApid VIsualization and it's a data-driven system for real time image rendering and quality analysis. Those images, though fascinating, are not suitable for scientific use. Scientists usually prefer to work with MERIS data products, which fully exploit the 15 spectral bands of the instrument, and which are generated with sophisticated algorithms. MIRAVI generates the images directly from the MERIS raw data (i.e. the Level 0 data), usually available within 2 hours from data acquisition. The MIRAVI image processing does not intend to provide a very accurate geo-location or calibration.
Name: **ESA Medspiration** *Type*: Web service, data, metadata, pictures *URL*: http://www.medspiration.org/ *Description*: Medspiration is a real-time service for the production and delivery of high-resolution sea surface temperature from all available satellite sensors.
Name: **ESA G-POD Services** *Type*: Web service, data, metadata, pictures *URL*: http://eogrid.esrin.esa.int/ *Description*: The ESA Grid on-Demand Services and Infrastructure (GODIS) for Earth Observation Applications provide access to a set of services. This portal offers access to science-oriented EO Grid services and applications, including access to a number of global geophysical ENVISAT products.

Fisheries and Aquaculture Resources Management (FARM)
Name: **Fisheries Fact Sheets** *Type*: XML, HTML *URL*: http://www.fao.org/fishery/factsheets/en *Description*: FAO's Fisheries Department prepares and publishes Fishery Fact sheets for a wide variety of thematic domains such as country profiles, species, vessel types, aquatic resources, etc. that are used by decisions makers, fisheries industry, researchers and students.
Name: **FAO Fisheries and Aquaculture Time Series** *Type*: XML, HTML and CSV *URL*: http://www.fao.org/fishery/topic/16140 *Description*: FAO's Fisheries Department databases contain over forty years of species catch and production statistics, as well as statistics on commodities, fleets and supply and utilisation in tabular format.
Name: **FAO Corporate Document Repository** *Type*: MS-word documents, Acrobat documents, HTML, XML metadata *URL*: http://www.fao.org/documents *Description*: The FAO Corporate Document Repository contains more than 25,000 full text publications and related metadata in English, French, Arabic and Chinese.
Name: **GeoNetwork** *Type*: GIS maps, XML geographic metadata *URL*: http://www.fao.org/geonetwork/srv/en/main.home *Description*: FAO-GeoNetwork is a web based Geographic Data and Information Management System. It enables easy access to local and distributed geospatial information catalogues and makes available data, graphics and documents for immediate download. FAO-GeoNetwork holds approximately 5000 standardized metadata records for digital and paper maps, most of them at the global, continent and national level.
Name: **Fisheries ontologies** *Type*: XML *URL*: http://www.fao.org/aims/neon.jsp *Description*: Within the NeON project FAO is developing ontological models for its classification systems, XML data, and other major information systems. These will be exploited to organize and harmonize data within the fisheries VREs.
Name: **ASFA Abstracts** *Type*: text *URL*: http://www.fao.org/fishery/asfa *Description*: Over one million Fisheries and oceanography abstracts stretching back to the early 1970's. They can be either imported or FAO can create a web service
Name: **Compilation of Maps for Fishery Distribution** *Type*: GIS maps *URL*: http://www.fao.org/fi/website/FIRetrieveAction.do?dom=collection&xml=fish_dist_map.xml&lang=en *Description*: For the compilation of maps for fishery distribution, a new approach has undertaken in order to better represent the real distribution of each species. The main source of information to describe the species distribution was their habitat descriptions and geographic distributions as derived from the official FAO Catalogues of Species. It was possible to make use of these information using a modern GIS approach after standardization of all terminology used in the descriptive contexts.

Name: **AGROVOC** *Type*: WSDL-Compliant web services *URL*: http://www.fao.org/aims *Description*: Multilingual General Purpose Thesaurus on Agriculture. Many FAO fisheries publications are keyed to this thesaurus. It is currently being harmonized with FAOTERM, which standardizes and harmonizes the vast quantity of terms used in FAO documents and publications.
Name: **FAOLex** *Type*: XML *URL*: http://faolex.fao.org/faolex/ *Description*: Provides texts of legal documents relating to agriculture including fisheries.
Name: **FAO News and Events Management System (NEMS)** *Type*: XML *URL*: http://www.fao.org/nems/ *Description*: Provides RSS format on FAO News and Events.
Name: **FishBase** *Type*: Web service *URL*: http://www.fishbase.org *Description*: A species information system on all 30,000 finfish species of the world. More than 60 topics covered, 2 millions records.
Name: **SeaLifeBase** *Type*: Data, Metadata, Pictures *URL*: http://www.sealifebase.org/ *Description*: A FishBase-like system plus distributed databases on all marine organisms. Started in late 2005, it currently contains data on over 80,000 species.
Name: **ReefBase** *Type*: Web service *URL*: http://www.reefbase.org *Description*: An information system that gathers and index more than 25,000 documents at country-level for information on coral reefs (not on coral species) as geo-structures, threats, status, and management.
Name: **Ecopath with Ecosim** *Type*: Modelling software *URL*: http://www.ecopath.org/ *Description*: Ecopath with Ecosim (EwE) is a free ecological/ecosystem modeling software suite. EwE has three main components: Ecopath – a static, mass-balanced snapshot of the system; Ecosim – a time dynamic simulation module for policy exploration; and Ecospace – a spatial and temporal dynamic module primarily designed for exploring impact and placement of protected areas.
Name: **FIRMS (Fisheries Resources Management System)** *Type*: XML, HTML *URL*: http://firms.fao.org *Description*: Annual fish stock assessments by a network of approximately 15 regional fishery bodies.
Name: **GOOS (Global Ocean Observing System)** *Type*: Sensor data *URL*: http://www.ioc-goos.org *Description*: A worldwide network of ocean sensors collecting hydrographic data such as ocean temperature and salinity.

A.6 Other Resources

In addition to the above mentioned, a number of other data sources will be part of the knowledge ecosystem, to support the depicted scenarios. The table below shows such additional data sources.

Other resources
Name: **Fisheries Global Information System (FIGIS)** *Type*: Web service, XML, XLS *URL*: http://www.fao.org/fishery/figis/2/en *Description*: The FIGIS system, using leading edge web-based information technologies (e.g. XML. XLS and Java), has developed Internet-based tools and functionality with the view to promote more effective and focussed cooperation between information sources, within FAO as well as between FAO, regional fisheries management organisations and national centres of excellence.
Name: **Reference Tables Management System (RTMS)** *Type*: XML, CSV *URL*: http://www.fao.org/fishery/rtms *Description*: RTMS is the repository for a set of international standards used to identify and store fisheries data, such as the International Standard Statistical Classification of Aquatic Animals and Plants (ISSCAAP) and the International Standard Statistical Classification of Fishery Commodities (ISSCFC).
Name: **Lifecycle for Networked Ontologies (NeOn toolkit)** *Type*: WSDL-based web services *URL*: http://www.neon-project.org/web-content/ *Description*: NeOn is a suite of technologies partly centered on the Eclipse development environment. Interaction between NeOn and VREs is foreseen only with some of the technologies.
Name: **North Atlantic Fisheries Organization** *Type*: Vessel monitoring system catch data *URL*: http://www.nafo.int/ *Description*: Time Series data on vessel movements and catches in the NAFO management area.
Name: **WTO statistics database** *Type*: time-series *URL*: http://stat.wto.org/Home/WSDBHome.aspx?Language=E *Description*: Time Series on merchandise and commercial services trade.
Name: **UN Comtrade** *Type*: SDMX *URL*: http://comtrade.un.org/ *Description*: The United Nations Commodity Trade Statistics Database contains international trade statistics data detailed by commodities and partner countries.
Name: **UN Data** *Type*: XML, CSV *URL*: http://data.un.org/ *Description*: UNdata, developed by the Statistics Division of Department of Economic and Social Affairs, is a new powerful tool, which gives access to a unique and authoritative set of UN statistical databases.

Appendix B Letters of Support

B.1 Biodiversity Heritage Library

Biodiversity Heritage Library

08 September, 2008
Dr. Donatella Castelli
D4Science II Scientific Project Coordinator
CNR-ISTI
Via Moruzzi, 1
56124 Pisa
ITALY

Letter of Support

With this letter we would like to express interest and support on behalf of The Biodiversity Heritage Library (BHL) for the I3 project named "D4Science II" within the 7th Framework Programme, under the action line Capacities-Research Infrastructures. The BHL is a consortium of ten natural history and botanical libraries organized to digitize and structure the legacy literature of biodiversity as part of a scientific commons, freely available, and make it interoperable through open web services with major scientific projects in such fields as biodiversity, climate change, oceanography, ecology etc. The BHL has digitized 7,760,000 pages to date and will have digitized 60,000,000 pages of core biodiversity literature by 2011.

The D4Science II program will advance a world-wide information infrastructure that will support just this kind of knowledge ecology. The true value and usefulness of the vast amount of content the BHL is providing requires a large, well-planned, visionary program as outlined in the D4Science II. Integration of BHL content and many, many other holders of important scientific content with other research environments should not be done one project at time using ad local solutions. It is a situation that cries for the intelligent economies of scale in evidence in this proposal. The BHL is interested in making available its collections available to the D4Science e-infrastructure, supporting the D4Science project in accessing other biodiversity resources, and in participating in the D4Science II meetings and being informed of its future events.

Thomas Garnett

Thomas Garnett
Biodiversity Heritage Library Program Director
Smithsonian Institution Libraries
Rm. 22
MRC 154 PO Box 37012
Washington, D.C. 20013-7012
202-633-2238
202-786-2866 fax
garnettt@si.edu

B.2 DRIVER

DRIVER - Digital Repository Infrastructure Vision for European Research
www.driver-community.eu

driver

Dr. Norbert Lossau
Scientific Coordinator

Niedersächsische Staats- und
Universitätsbibliothek/
Goettingen State and University Library,
Director
Platz der Goettinger Sieben, 1
DB37073 Goettingen
Germany

Donatella Castelli
D4Science II Scientific Project Coordinator
CNR-ISTI
Via Moruzzi, 1
56124 Pisa
ITALY

27 August 2008

Letter of Support

With this letter we would like to express interest and support on behalf of DRIVER for the I3 project named "D4Science II" within the 7th Framework Programme, under the action line Capacities-Research Infrastructures.

We support the objectives of the D4Science II project in establishing interoperability between the D4Science e-Infrastructure and other large data e-Infrastructures and repositories. In particular, DRIVER expresses its interest in proposed solutions that will achieve interoperability in a manner that transparently exploits GRID technologies. We support synergies with D4Science in building enhanced publications and in developing services for the extraction of bibliometric information to improve the personalization functionality of DRIVER.

Within the framework of the DRIVER Confederation, we further support synergies with D4Science II in addressing the sustainability of the infrastructure, and aim to pursue common objectives at a high strategic level.

On behalf of the DRIVER consortium

Yours sincerely,

(Dr. Norbert Lossau, Scientific Coordinator)

DRIVER II Partners: University of Athens - Universität Bielefeld - Consiglio Nazionale Delle Ricerche - Stichting SURF - University of Nottingham - University of Bath - Uniwersytet Warszawski - Universiteit Gent - Universität Göttingen - Danmarks Tekniske Universitet - Narodna in univerzitetna knjiznica - Universidade do Minho

B.3 EGEE

ORGANISATION EUROPEENNE POUR LA RECHERCHE NUCLEAIRE
EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH

Laboratoire Européen pour la Physique des Particules
European Laboratory for Particle Physics

GENEVE, SUISSE
GENEVA, SWITZERLAND

Enabling Grids for E-sciencE

Mail address:
Dr. Bob Jones
CERN, IT Department / EGE group
CH-1211 GENEVE 23
Switzerland

E-mail: bob.jones@cern.ch

Donatella Castelli
CNR-ISTI
Via Moruzzi, 1
5124 Pisa (Italy)

Votre référence/Your reference:
Notre référence/Our reference: D4SCIENCEII_LoS

Geneva, September 2, 2008

Dear Dr Castelli,

This is a letter to show my support for the D4Science II project proposal. On behalf of the EGEE consortium, I wish to confirm our willingness to work with the D4Science II project which intends to continue the path initiated by the DILIGENT project and built on by D4Science, towards the creation of a broad European multidisciplinary scientific data e-Infrastructure that, building on the achievements of network, grid and more recent repository infrastructures provides scientists with innovative environments for sharing research products and naturally supporting knowledge ecology.

Exploiting the EGEE infrastructure, the D4Science II project proposal intends to take a further step along this path by transforming the D4Science e-Infrastructure into a driving component of a knowledge ecosystem of collaborating scientific e-Infrastructures and thus enhance the functionality offered by EGEE.

The EGEE consortium involves more than 140 leading institutions in 33 countries, federated in regional Grids, with a current combined capacity of over 70'000 CPUs, the largest international Grid infrastructure ever assembled for the benefit of a wide range of user communities totalling more than 10,000 researchers and scientists.

My colleagues and I look forward to the opportunity of a fruitful collaboration and wish you and your project the best of luck.

Yours sincerely,

Bob Jones
EGEE-III Project Director

B.4 European Grid Initiative (EGI)

LMU

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

DEPARTMENT
INSTITUT FÜR INFORMATIK
LEHR- UND FORSCHUNGSEINHEIT FÜR KOMMUNIKATIONS
SYSTEME UND SYSTEMPROGRAMMIERUNG

Prof. Dr. Dieter Kranzlmüller · Department „Institut für Informatik"
Universität München · Oettingenstr. 67 · 80538 München

Donatella Castelli
CNR-ISTI
Via Moruzzi, 1
5124 Pisa
ITALY

EMail: donatella.castelli@isti.cnr.it

Prof. Dr. Dieter Kranzlmüller
Oettingenstr. 67, 80538 München
Tel: 0 89 / 21 80- 9145
Fax: 0 89 / 21 80- 9147

kranzlmueller@ifi.lmu.de
www.nm.ifi.lmu.de

München, den 11.09.2008

Subject: Letter of Support for D4Science-II – Data Infrastructures Ecosystem for Science

Dear Donatella,

On behalf of the **EGI (European Grid Initiative) Preparation Team**, I would like to offer our strong support for the project

D4Science-II – Data Infrastructures Ecosystem for Science

and for your proposal in response to the EU Call FP7-INFRASTRUCTURES-2008-2.

The continuation of DILIGENT and D4Science in your project towards the creation of a broad European multidisciplinary scientific data e-Infrastructure is a very important activity. Building on the achievements of networks, grids and more recent repository infrastructures provides scientists with innovative environments for sharing research products and naturally supporting knowledge ecology. Your intention to take the existing solutions one step further along this path by transforming the current e-Infrastructure into a driving component of a knowledge ecosystem of collaborating scientific e-Infrastructures is seen very valuable for European science and research.

We are very interested in the aspects addressed by D4Science-II and we are looking forward in creating closer links with your project. We believe that such collaboration is necessary for the benefits of the whole environment. Once your project is accepted, we would be happy to discuss in more detail, how EGI_DS and your project could collaborate in the future.

Please feel free to contact us at contact@eu-egi.org.

Yours sincerely,

Univ.-Prof. Dr. Dieter Kranzlmüller, EGI Strategy Director
(on behalf of the EGI Preparation Team)

B.5 European Telecommunication Standards Institute (ETSI)

ETSI World Class Standards

Dr. Donatella Castelli
D4Science II Scientific Project Coordinator
CNR-ISTI
Via Moruzzi, 1
56124 Pisa
ITALY

Sophia Antipolis, 04. September 2008

Letter of Support
With this letter we would like to express interest and support on behalf of ETSI for the I3 project named "D4Science II" within the 7th Framework Programme, under the action line Capacities-Research Infrastructures.

We support the objectives of the D4Science II project in establishing interoperability between the D4Science e-Infrastructure and other large data e-Infrastructures and repositories by exploiting standards solutions. We are convinced that the feedback collected by its participants on the early adoption of standards and the experience done within the project in identifying interoperability solutions will significantly contribute to the definition and tuning of emerging standards.

ETSI expresses its interest in being able to receive and use the results of the D4Science II Project and in hosting and organising interoperability events.

Yours sincerely,

Signed on behalf of ETSI

Dr. Walter Weigel
Director General

www.etsi.org
European Telecommunications Standards Institute - Institut Européen des Normes de Télécommunication - Europäisches Institut für Telekommunikationsnormen
ETSI - 650, route des Lucioles - 06921 Sophia Antipolis Cedex - France
Tél. +33 (0)4 92 94 42 00 Fax. +33 (0)4 93 65 47 16 e-mail : secretariat@etsi.org
SIRET N° 348 623 562 00017 - APE 7112B - Association à but non lucratif enregistrée à la Sous-Préfecture de Grasse (06) N° 7803/88 - N°TVA : FR 14 348 623 562

B.6 GENESI-DR

Luigi Fusco
GENESI-DR Project Coordinator
European Space Agency
EO Programmes Directorate

Phone : +39-06-94180530
Email : Luigi.Fusco@esa.int

Frascati, 2 September 2008

Dr. Donatella Castelli
D4Science II Scientific Project Coordinator
CNR-ISTI - Via Moruzzi, 1
56124 Pisa - ITALY

Letter of Support

On behalf of the 7th Framework Programme Project GENESI-DR, short for Ground European Network for Earth Science Interoperations – Digital Repositories (Grant Agreement no. FP7-212073) lead by the European Space Agency (ESA), I would like to express interest in and support to the I3 project named "D4Science II" within the 7th Framework Programme, under the action line Capacities -Research Infrastructures.

GENESI-DR supports the objectives of the D4Science II project in establishing interoperability between the D4Science e-Infrastructure and other large data e-Infrastructures and repositories.

In particular, ESA, as coordinator of the GENESI-DR project, expresses its interest in making available GENESI-DR resources, i.e. collections, tools and computing resources to the D4Science II users according to the policies that will be specified in a MoU to be prepared in the initial phases of the D4Science II project.

Yours sincerely,

LUIGI FUSCO
GENESI-DR Project Coordinator

B.7 International Council for the Exploitation of the Sea (ICES)

ICES
CIEM

International Council for the Exploration of the Sea

Conseil International pour l'Exploration de la Mer

General Secretary
Dr Gerd Hubold

H. C. Andersens Boulevard 44–46
DK-1553 Copenhagen V
Denmark

Telephone: (45) 33 38 67 00
Telefax: (45) 33 93 42 15
E-mail: info@ices.dk
www.ices.dk

Direct Dialling:
(45) 33 38 6 + ext.

Marc Taconet
The Food and Agriculture Organization of the United Nations
Viale delle Terme di Caracalla
IT-00100 Rome
Italy

Our Ref: G.1/J.6.b 9 September 2008

ICES statement regarding the D4Science project and its follow-up phase D4Science II

Dear Marc,

The ICES Secretariat is very interested in improving the quality of fisheries statistical data both regionally and on a global level and are fully prepared to participate in the implementation of the recommendation by CWP that *FAO should consolidate the catch data of RFBs into a single database under the general guidance of CWP* (see COFI Report 2007 and CWP Report 2007).

National and regional institutes and organizations hold large amounts of relevant fisheries data (e.g. detailed catch and effort as well as biomass estimates, population dynamics, geographic distribution, etc. of marine species). Such higher resolution data are not accessible through the existing global databases e.g. the FAO NS database and these data should be readily accessible.

ICES data policy advocates that as much data as possible should be in the public domain and ICES will together with the Secretariats of other Regional Fisheries Bodies (RFB's) participate in every effort to enhance access to and transparency of such regional and national fishery data, e.g. through a global data repository that would allow individual agencies and organizations to share their data through a gridded network database environment such as the one proposed under D4Science project and its follow-up phase D4Science II. Also, ICES is interested in the development of tools that will allow the comparison of different data sets through aggregation and unit conversion. There exist conflicting data in the databases and such discrepancies should be identified, understood and if possible removed.

Interpretation of the data and refinement of their resolution might be interesting and valuable, especially in data-poor environments. Such interpretations would be done by experts and the resulting data products should be clearly identified as the work of these experts and clearly distinct from the basic data that are held in the databases. The desirability of such products for the ICES area would be judged on a case-by-case basis. There are numerous examples of highly relevant studies where data are interpreted and their resolution refined through reallocation/mapping rules, e.g. by using seasonal catch assumptions to assign annually reported catches to months, and/or by using biological characteristics of fish species (e.g. depth/habitat preferences) to assign catches reported by a larger area to a smaller geographic reference point. However, it is more difficult to envisage such data products for

Excellence in marine science since 1902

general use. ICES experience suggests that the cost involved to create and maintain such general products does not match their usefulness.

The ICES Secretariat suggest that in the D4Science project introduces a clear distinction between

(a) the inventory of existing more detailed regional and national fishery data, their submission to a gridded system and their subsequent integration including the creation of comparability and aggregation tools; and

(b) the development of products to refine the original resolution of submitted data using reallocation/mapping rules (the current proposed ICIS outputs).

ICES will be happy to contribute to the part related to point (a) but do not find that ICES can justify using resources on the part related with point (b).

Regards,

Hans

Cc. Sachiko Tsuji

B.8 OBIS

OCEAN BIOGEOGRAPHIC
INFORMATION SYSTEM
http://www.iobis.org

7 August 2008

Dr. Donatella Castelli
D4Science II Scientific Project Coordinator
CNR-ISTI
Via Moruzzi, 1
56124 Pisa
ITALY

Letter of Support

With this letter we would like to express interest and support on behalf of the Ocean Biogeographic Information System (OBIS) for the I3 project named "D4Science II" within the 7th Framework Programme, under the action line Capacities-Research Infrastructures.

We support the objectives of the D4Science II project in establishing interoperability between the D4Science e-Infrastructure and other large data e-Infrastructures and repositories. In particular, the OBIS expresses its interest in

- making available the OBIS resources, i.e. collections and related tools, to the D4Science e-infrastructure according to policies that will be defined by a specific MoU prepared in the early phases of the D4Science II project
- supporting the D4Science project in accessing the OBIS resources
- being informed of the results of D4Science II project
- participating to the D4Science II meetings and being informed on its future events

One of the explicit goals of the OBIS comunity is to make large datasets on marine biogeography readily accessible, free and open to everyone. The information technology developed by D4Science is promising as an infrastructure to make this ambition a reality.

Yours sincerely,

Dr Edward Vanden Berghe
OBIS Executive Director

OBIS Secretariat; Institute of Marine and Coastal Sciences
Rutgers, The State University of New Jersey
71 Dudley Road, New Brunswick, New Jersey 08901-8521, USA
Telephone: +1 732.932.6555 ext. 509 Facsimile: +1 732.932.8578

B.9 OGF-EUROPE

OpenGridForum
Europe

CAPACITIES

Silvana Muscella
OGF.eeig Director European
Chapter of the Open Grid Forum
OGF-Europe Technical Director
s.muscella@trust-itservices.com
Tel: + 39 050 28359

4th September 2008

Dr. Donatella Castelli
D4Science II Scientific Project Coordinator
CNR-ISTI
Via Moruzzi, 1
56124 Pisa
ITALY

Subject: Letter of Support on behalf of OGF-Europe SSA to D4SCIENCE II

With this letter we would like to express interest and support on behalf of OGF-Europe for the I3 project named ***"D4Science II"*** within the 7th Framework Programme, under the action line Capacities-Research Infrastructures.

We support the objectives of the D4Science II project in establishing interoperability between the D4Science e-Infrastructure and other large data e-Infrastructures and repositories by exploiting standards solutions. We are convinced that the feedback collected by its participants on the early adoption of standards and the experience carried out within the project in identifying interoperable solutions will significantly contribute to the definition and fine- tuning of emerging standards.

The OGF-Europe SSA expresses its interest as its mission is in promoting Open Standards and maximising interoperability in the research and industry fields in different vertical markets.

We would fully appreciate receiving and using the results of the D4Science II Project. Moreover, current discussions are underway at OGF to set up a Working Group specifically on Data for different user communities, there is a pre-meeting in December 2008 at the IDCC2008 Edinburgh to pursue this. Representatives of the D4Science project are invited to attend this meeting as would be most helpful in tapping into the user community groups of the D4Science community.

We wish you every success in the proposal and look forward to collaborating with you upon a successful outcome.

Signed on behalf of OGF-EUROPE

Yours sincerely,

SILVANA MUSCELLA
Technical Coordinator of OGF-Europe &
European Director of the European Chapter of the Open Grid Forum

References

[1] *ADS: The SAO/NASA Astrophysics Data System*. [Online] http://adswww.harvard.edu/

[2] Alpdemir, M. N.; Mukherjee, A.; Paton, N.W.; Watson, P.; Fernandes, A. A.; Gounaris, A.; Smith, J. *Service-based distributed querying on the grid*. In the Proceedings of the First International Conference on Service Oriented Computing, pages 467-482. Springer, 15-18 December 2003.

[3] Antonioletti, M.; Hastings, S.; Krause, A.; Langella, S.; Lynden, S.; Laws, S.; Malaika, S.; Paton, N. W. *Web Services Data Access and Integration – The XML Realization (WS-DAIX) Specification, Version 1.0*. Open Grid Forum Document Series, GFD-R-P.075, August 2006

[4] *APAC: Providing grid services for Australian eResearch.* [Online] http://grid.apac.edu.au/

[5] *ArXiv*. [Online] http://arxiv.org/

[6] Assante, M.; Candela, L.; Castelli, D.; Frosini, L.; Lelii, L.; Manghi, P.; Manzi, A.; Pagano, P.; Simi, M. *An Extensible Virtual Digital Libraries Generator*. In Proceedings of the 12th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2008). Aarhus, Denmark. September 2008.

[7] Banks T. *Web Services Resource Framework (WSRF) – Primer v1.2*. OASIS Committee Draft, 2006 [Online] http://docs.oasis-open.org/wsrf/wsrf-primer-1.2-primer-cd-02.pdf

[8] Beck, K.; Beedle, M.; van Bennekum, A.; Cockburn, A.; Cunningham, W.; Fowler, M.; Grenning, J.; Highsmith, J.; Hunt, A.; Jeffries, R.; Kern, J.; Marick, B.; Martin, R. C.; Mellor, S.; Schwaber, K.; Sutherland, J.; Thomas, D. *Manifesto for Agile Software Development.* 2001. [Online] *http://www.agilemanifesto.org*

[9] Bizer, C.; Grauß, T.; Cyganiak, R. *Semantic Web Client Library*. [Online] http://www4.wiwiss.fu-berlin.de/bizer/ng4j/semwebclient/index.html

[10] Boutsis, S.; Candela, L.; Di Marcello, P.; Fabriani, P.; Frosini, L.; Kakaletris, G.; Koltsida, P.; Lelii, L.; Milano, D.; Pagano, P.; Polydoras, P.; Ranaldi, P.; Simeoni, F.; Simi, M.; Tsagkalidou, V.; Turli, A. *D4Science System High-level Design.* D4Science Project Deliverable DJRA1.1a. September 2008

[11] Candela, L.; Akal, F.; Avancini, H.; Castelli, D.; Fusco, L.; Guidetti, V.; Langguth, C.; Manzi, A.; Pagano, P.; Schuldt, H.; Simi, M.; Springmann, M.; Voicu, L. *DILIGENT: Integrating Digital Library and Grid Technologies for a new Earth Observation Research Infrastructure*. In International Journal on Digital Libraries - Special Issue on Digital Libraries and eScience, vol. 7, October 2007.

[12] *CDS: CERN Document Server*. [Online] http://cdsweb.cern.ch/

[13] *CDS Invenio.* [Online] http://cdsware.cern.ch/invenio/index.html

[14] Costello, M.J.; Stocks, K.; Zhang, Y.; Grassle, J.F.; Fautin, D.G. *About the Ocean Biogeographic Information System*. [Online] www.iobis.org. August 2008

[15] *CROWN: China Research and development environment Over Wide-area Network*. [Online] http://www.crown.org.cn/

[16] *CWP Handbook of Fishery Statistical Standards.* [Online] http://www.fao.org/fishery/cwp/search/en

[17] *D4Science: DIstributed colLaboratories Infrastructure on Grid ENabled Technology for Science*. EC Contract No. 212488. [Online] http://www.d4science.eu

[18] *DarwinCore.* [Online] http://wiki.tdwg.org/twiki/bin/view/DarwinCore/WebHome

[19] *DEISA: Distributed European Infrastructure for Supercomputing Applications.* [Online] http://www.deisa.eu/

[20] *D-Grid Initiative.* [Online] http://www.d-grid.de/

[21] *DILIGENT: A DIgital Library Infrastructure on Grid ENabled Technology*. EC Contract No. 004260. [Online] www.diligentproject.org

[22] *DRIVER: Digital Repository Infrastructure Vision for European Research.* [Online] http://www.driver-repository.eu/

[23] *Dublin Core Metadata Initiative (DCMI)*. [Online] http://dublincore.org/

[24] *EGEE: Enabling Grids for E-science in Europe*. [Online] http://public.eu-egee.org

[25] *FIGIS: Fisheries Global Information System.* [Online] http://www.fao.org/fishery/figis/2/en. August 2008

[26] Foster, I.; Kishimoto, H.; Savva, A.; Berry, D.; Djaoui, A.; Grimshaw, A.; Horn, B.; Maciel, F.; Siebenlist, F.; Subramaniam, R.; Treadwell, J.; Von Reich, J. *The Open Grid Services Architecture, Version 1.5*. Open Grid Forum Document Series, GFD-I.080, July 2006

[27] Froese, R. and Pauly, D. Editors. 2008. *FishBase*. [Online] www.fishbase.org. August 2008.

[28] *GARUDA: India's National Grid Computing Initiative*. [Online] http://www.garudaindia.in/

[29] *GBIF: Global Biodiversity Information Facility*. [Online] www.gbif.org/

[30] *gCube System.* [Online] http://www.gcube-system.org/

[31] *GÉANT.* [Online] http://www.geant.net/

[32] *GENESI-DR: Ground European Network for Earth Science Interoperations – Digital Repositories.* [Online] www.genesi-dr.eu

[33] Gentil-Beccot, A.; Mele, S.; Holtkamp, A.; O'Connell, H.B.; Brooks, T. C. *Information Resources in High-Energy Physics: Surveying the Present Landscape and Charting the Future Course*. arxiv.org:0804.2701v2

[34] *GEOSS: Global Earth Observation System of Systems.* [Online] http://www.epa.gov/geoss/

[35] Ginsparg, P. *First Steps Towards Electronic Research Communication*. Computers in Physics, Vol. 8, 1994

[36] *gLite Lightweigth Middleware for Grid Computing.* [Online] http://glite.web.cern.ch/glite/

[37] *GMES: Global Monitoring for Environment and Security.* [Online] www.gmes.info

[38] *GOOS: Global Ocean Observing System.* [Online] http://www.ioc-goos.org/

[39] Graham, S.; Hull, D.; Murray, B. *Web Services Base Notification (WS-BaseNotification)*. OASIS Standard, October 2006

[40] *GriFin: Grid for Finance.* [Online] http://grb.spaci.it/grifin/

[41] *INSPIRE*. [Online] https://twiki.cern.ch/twiki/bin/view/Inspire

[42] International Organisation for Standardization. *Geographic Information – Geography Markup Language (GML)*. ISO 19136:2007. [Online] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=32554

[43] International Organisation for Standardization. *Information technology – Multimedia Framework (MPEG-21) – Part 2: Digital Item Declaration*. ISO /IEC 21000-2:2005. [Online] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=41112

[44] Jordan, C.; Kishimoto, H. *Defining the Grid: A Roadmap for OGSA® Standards.* Open Grid Forum Document Series, GFD-I.123, February 2008

[45] Kaschner, K.; Ready, J. S.; Agbayani, E.; Rius, J.; Kesner-Reyes, K.; Eastwood, P. D.; South, A. B.; Kullander, S. O.; Rees, T.; Close, C. H.; Watson, R.; Pauly, D.; Froese, R. *AquaMaps: Predicted range maps for aquatic species*. [Online] http://www.aquamaps.org. August 2008.

[46] Kaschner, K.; Watson, R.; Trites, A.W.; Pauly, D. *Mapping worldwide distributions of marine mammals using a Relative Environmental Suitability (RES) model*. Marine Ecology Progress Series 316:285-310, 2006.

[47] Lagoze, C.; Van de Sompel, H. *The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)*. [Online] http://www.openarchives.org/OAI/openarchivesprotocol.html

[48] Linking Open Data Community. The Linking Open Data dataset cloud. [Online] http://richard.cyganiak.de/2007/10/lod/

[49] Manghi, P.; Mikulicic, M.; Artini, M.; Castelli, D.; Pagano, P.; Candela, L. *Sustainable Digital Library Systems over the DRIVER Repository Infrastructure*. In Proceedings of the 12th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2008). Aarhus, Denmark. September 2008.

[50] Mele, S.; Holtkamp, A.; O'Connell, H.; Brooks, T. *High-Energy Physics Labs Join to Build a New Scientific Information System*. Interactions News Wire #37-08, May 2008. [Online] http://www.interactions.org/cms/?pid=1026243

[51] *METS: Metadata Encoding and Transmission Standard*. [Online] http://www.loc.gov/standards/mets/

[52] *MPEG: Moving Picture Experts Group.* [Online] http://en.wikipedia.org/wiki/MPEG

[53] *NAREGI: National Research Grid Initiative*. [Online] http://www.naregi.org/

[54] *NDGF: Nordic DataGrid Facility*. [Online] http://www.ndgf.org

[55] *NeOn: Lyfecycle Support for Networked Ontologies*. [Online] http://www.neon-project.org/

[56] *NGS: National Grid Service*. [Online] http://www.grid-support.ac.uk/

[57] *OAI-ORE: Open Archives Initiatives – Object Reuse and Exchange.* [Online] http://www.openarchives.org/ore/

[58] O'Dell, C.; Dallman, D.P.; Vigen, J.; Vesely, M. *50 Years of Experience in Making Grey Literature Available: Matching the Expectations of the Particle Physics Community*. In 5th International Conference on Grey Literature: Grey Matters in the World of Networked Information, GL 2003, Amsterdam, The Netherlands , 4 - 5 Dec 2003, p. 117-123.

[59] *OMII-Europe: Open Middleware Infrastructure Institute.* [Online] http://omii-europe.org/OMII-Europe/

[60] Open Geospatial Consortium. *Keyhole Markup Language (OGC KML)*. [Online] http://www.opengeospatial.org/standards/kml

[61] *Open Grid Forum*. [Online] http://www.ggf.org/

[62] *OpenSearch.* [Online] http://www.opensearch.org

[63] *OSG: Open Science Grid*. [Online] http://www.opensciencegrid.org/

[64] Palomares, M.L.D. and D. Pauly. Editors. *SeaLifeBase*. [Online] www.sealifebase.org. August 2008.

[65] *PRAGMA: Pacific Rim Applications and Grid Middleware Assembly*. [Online] http://www.pragma-grid.net/

[66] Simeoni, F.; Candela, L.; Kakaletris, G.; Sibeko, M.; Pagano, P.; Papanikos, G.; Polydoras, P.; Ioannidis, Y. E.; Aarvaag, D.; Crestani, F. *A Grid-Based Infrastructure for Distributed Retrieval*. In Research and Advanced Technology for Digital Libraries, 11th European Conference ECDL 2007, Proceedings. Budapest, Hungary, September 16-21, 2007

[67] *SPIRES*. [Online] http://www.slac.stanford.edu/spires/

[68] Sneth, A.P. *Changing focus on interoperability in information systems: from system, syntax, structure to semantics*. In Goodchild, M.; Egenhofer, M.J.; Fegeas, R.; Kottman, C. (Eds.) Interoperating Geographic Information Systems, Springer, 1999.

[69] *TeraGrid*. [Online] http://www.teragrid.org/

[70] Terradue srl. *duetopia – Discovery and Registry of Geospatial Datasets*. [Online] http://www.terradue.com/duetopia/

[71] *The North Atlantic Format (NAF). Fisheries Standard for Electronic Data Transmission.* [Online] http://www.naf-format.org

[72] *UN Comtrade.* [Online] http://comtrade.un.org/

[73] *UNdata.* [Online] http://data.un.org/

[74] Vanbenepe, W.; Graham, S.; Niblett, P. *Web Services Topics 1.3 (WS-Topics)*. OASIS Standard, October 2006

[75] W3C Semantic Web Activity. *Resource Description Framework (RDF)*. [Online] http://www.w3.org/2001/sw/

[76] W3C Semantic Web Activity. *Web Ontology Language* (OWL). [Online] http://www.w3.org/2004/OWL/