PROCEEDINGS OF SCIENCE

Realising Virtual Research Environments by Hybrid Data Infrastructures: the D4Science Experience

Leonardo Candela[a], **Donatella Castelli**[a], **Andrea Manzi**[b] **and Pasquale Pagano**[a]
[a] *Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", Consiglio Nazionale delle Ricerche, Via G. Moruzzi 1, 56124, Pisa, Italy*
E-mail: {candela | castelli | pagano}@isti.cnr.it
[b] *European Organisation for Nuclear Research CERN, CH – 1211 Geneva 23, Switzerland*
E-mail: andrea.manzi@cern.ch

Modern science tend to be more than ever multidisciplinary, collaborative and demanding. Scientific investigations span the boundaries of single institutions, disciplines, and countries. In order to serve these scenarios innovative working environments are needed. In this paper it is presented an innovative environment consisting of an Hybrid Data Infrastructure supporting the dynamic deployment and operation of an array of Virtual Research Environments, each tailored to serve the needs of a scientific community towards a research endeavour. The paper reports on the enabling technology and how it has been deployed to realise the D4Science infrastructure and serve the needs of different communities.

International Symposium on Grids and Clouds (ISGC) 2014,
23-28 March 2014
Academia Sinica, Taipei, Taiwan

PoS(ISGC2014)022

1. Introduction

Science and scientists are calling for innovative approaches and practices aiming at facilitating research collaborations that span institutions, disciplines, and countries. These approaches should be offered under the "as-a-Service" paradigm, i.e., scientists expect to be provided with innovative working environments that give them the facilities they need while allowing them to save time and money without compromising research quality. The expected facilities include (*i*) data, usually falling in the big data domain and spreading across multiple information systems and repositories, (*ii*) services, i.e., an open ended set of processes and workflows supporting data analysis and mining, and (*iii*) computing capabilities, i.e., the power to elastically acquire the amount of computing resources needed to effectively and efficiently execute onerous tasks.

To serve these scenarios, D4Science.org is operating an *Hybrid Data Infrastructure* [1]. This is an IT infrastructure built as a "system of systems", i.e., implemented by nicely integrating other infrastructures, services and information systems. This HDI is conceived to enable the delivery of *Virtual Research Environments* "as-a-Service" [2]. Virtual Research Environments are web-based working environments where groups of scientists, possibly geographically distant from each other, have user friendly, transparent and seamless access to the flexible and shared set of remote resources (data, services and computing capabilities) needed to perform their work.

In this paper, we present the D4Science Hybrid Data Infrastructure (HDI) and the rich array of Virtual Research Environments deployed and operated to serve communities of practice in domains including biodiversity, environment and fisheries.

The remainder of this paper is organised as follows. Section 2 extensively describes the D4Science enabling technology, i.e., gCube. Section 3 documents the D4Science Infrastructure. Section 4 describes the approach leading to the creation of VREs and presents the currently existing ones. Finally, Section 5 concludes the paper.

2. gCube: the enabling technology

gCube [3, 5] is a software system specifically conceived to enable the creation and operation of an innovative typology of infrastructure, i.e., an *Hybrid Data Infrastructure* [1], that by leveraging Grid [6], Cloud [7], Digital Library [8] and Service-orientation [9] principles and approaches is delivering a number of data management facilities *as-a-Service*. One of its distinguishing feature is the orientation to serve the needs of diverse Communities of Practice [10] by providing each of them with a dedicated, flexible, ready-to-use, web-based working environment, i.e., a *Virtual Research Environment* [2, 11].

gCube hosts a compelling portfolio of applications having vast and heterogeneous target audience ranging from scientists willing to perform their investigations in a more "simple" way to service providers willing to develop innovative facilities for scientists. The current catalogue of applications captures six main domain bundles that can be customized to meet specific needs (Fig. 1).

2.1 AppsCube

AppsCube is a framework conceived to support practitioners willing to develop applications

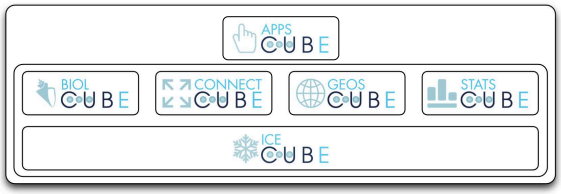

Figure 1: gCube Application Bundles

interfacing with and benefitting from gCube facilities. It includes applications ranging from the *Application Service Layer*, i.e., a framework acting as a middleware between the gCube lower level services and the presentation layer, to the *Featherweight Stack*, i.e., a set of *microlibs* enabling the interaction with gCube services, and the *SmartGears*, i.e., a set of Java libraries that transparently turns Servlet-based containers and applications into gCube resources. More details are available in the Developer's Guide [4].

2.2 BiolCube

BiolCube is a gCube application bundle offering facilities to practitioners working with species occurrence data and taxonomic profiles.

In particular, it offers facilities for discovering and accessing species occurrence and taxonomic data within major repositories and information systems, e.g., OBIS, GBIF, Catalogue of Life, WoRMS [12]. The discovery mechanism is simple (based on species common names or scientific names) yet powerful since it supports query expansion and additional filters. The identified datasets are enriched with links to other species, can be displayed on a map as well as saved in standard formats (e.g., DarwinCore, DarwinCore-Archive, CSV) for future uses.

Moreover, it offers facilities for *species occurrence datasets processing* [12], facilities for *species distribution modeling* [13], and facilities for taxonomic and nomenclature data comparison [14, 15]. The facilities species occurrence datasets processing include algebraic operations (union, intersection, subtraction, and duplicates deletion) based on spatial and syntactic similarity measurements, clustering (e.g., density based algorithms such as DBScan, distance based algorithms such as K-means), outliers detection (e.g., Local Outlier Factor approach), occurrence points representativeness (e.g., Habitat Representativeness Score technique), and occurrence points enrichment with chemical and physical environmental parameters. The facilities for species distribution modeling offer a rich array of dedicated algorithms and approaches including AquaMaps [13]. AquaMaps is actually a family of approaches (e.g., suitable, native) producing species distribution probabilities on half-degree cells by relying on (*i*) HSPEN, a table containing species envelops, (*ii*) HCAF, a table containing environmental parameters and (*iii*) a table containing species occurrences points (half-degree cells). Moreover, if offers methods for producing new versions of HSPEN and HCAF. The facilities for comparing taxonomic and nomenclature data include (*a*) a flexible environment for comparing any two taxonomic checklists in DarwinCore-Archive format to detect, analyse and report relationships among taxa of the compared checklists (e.g., corresponds, includes, overlaps, not found in) [14], and (*b*) Bionym [15], i.e., a taxonomic data matching workflow based approach that enables users to combine a number of matchers (e.g., GSay, FuzzyMatcher, Levensthein, Tri-

gram) and tunes their contribution while identifying the proper scientific names recognised by a number of authoritative sources.

2.3 ConnectCube

ConnectCube is a gCube application bundle offering facilities to practitioners wanting to produce information-rich objects, resulting from the aggregation and synthesis of data from multiple sources.

In particular, if offers an innovative collaboration oriented environment integrating social networking practices in research environments [16]. It is conceptually close to the common facilities promoted by social networks – e.g., posting news, commenting on posted news, re-share news – yet adapted to promote large scale collaboration and cooperation on comprehensive scientific products, datasets, theories, and tools. Apart from post-oriented facilities, the environment offers a *shared workspace* and a *messaging* application. The workspace resembles a folder-based file system for managing information objects. The added value is represented by the type of information objects it can manage in a seamless way. It supports items ranging from binary files to information objects representing datasets, workflows, species distribution maps, time series, and comprehensive research products. Through it, data sharing is fostered, making results, workflows, annotations and documents immediately available to co-workers, to anyone provided with URIs, and to any other person authorized via WebDAV. The messaging application resembles an email environment with the distinguishing feature of being integrated with the rest, e.g., it is possible to send as attachment any dataset residing in the workspace without consuming bandwidth.

Moreover, it offers an application for creating and managing enhanced documents [17], i.e., rich information objects resembling documents yet aggregating multiple parts. Parts include images, datasets, maps, and graphs. It offers functionality for defining templates the documents should adhere to, as well as facilities for defining and monitoring workflows driving the collaborative production of these "documents".

A set of facilities providing its users with an integrated discovery and access to heterogeneous data completes the offering of this bundle. In particular, the Information Object Discovery application offers facilities for retrieving information objects from multiple collections and information systems in a seamless way [18]. It offers both a Google-like approach and an advanced search allowing users to characterise in detail the information objects they are looking for. It includes facilities for presenting the results according to semantic-based clustering [19]. Moreover, it offers a unifying domain specific top level ontology providing users with an integrated view over a number of information [20].

2.4 GeosCube

GeoCube is a gCube application bundle offering facilities to practitioners dealing with geospatial information.

In particular, it offers facilities for geospatial data discovery and processing [21].

The Geospatial Data Discovery application offers facilities for browsing and visualising geospatial data. In particular, these include facilities to navigate, search and discovery layers within a GeoNetwork instance via the OGC CSW protocol [22]. Moreover, these include facilities to interactively explore, manipulate, visualize, compare, and analyse geospatial data.

The Geospatial Data Processing application offers facilities for executing a rich array of data processing tasks on geospatial data. The current set of supported tasks includes maps comparison algorithms (supported formats include WFS, OPeNDAP and ASC), an intersection algorithm, i.e., an approach that computes the percentages of overlap between different areas on the compared maps, and many more. Moreover, the application enables the invocation of services exposing their capabilities via the OGC WPS protocol [23].

2.5 IceCube

IceCube is a gCube application bundle offering core facilities for the deployment, operation and management of a gCube based infrastructure. This is the basic building block of this technology and it is particularly relevant for the mechanism enabling the development of the Virtual Research Environments. It includes the following applications:

Policy-oriented Security Framework This application offers facilities for authorisation, authentication and accounting as-a-Service. It is based on standard protocols and technologies (e.g., SAML) providing: (*a*) an open and extensible architecture; (*b*) interoperability with external infrastructures and domains while obtaining the so-called Identity Federation.

Resources Management This application offers facilities for the management (discovery, deployment, monitoring) of *resources* including hosting nodes, services, software and datasets. It includes (*a*) an *Information System* acting as a registry of the infrastructure by offering global and partial views of its resources and their current status and notification instruments; (*b*) a *Resource Management Service* that builds on the Information Service to realise resource allocation and deployment strategies. For resource allocation, it enables the dynamic assignment of a number of selected resources to a given community (e.g., the creation of a VRE requires that a set of hosting nodes, service instances and data collections are allocated to a given application). For deployment, it enables the allocation and activation of both gCube software and external software on hosting nodes, i.e., servers able to host running instances of services; (*c*) an *hosting node*, a software component that once installed on a server transforms it into a gCube hosting node. A gCube hosting node can be dynamically endowed with a number of services including a local worker to execute computing tasks on that server.

File-oriented Storage Facilities This is a scalable high-performance storage service. In particular, it relies on a network of distributed storage nodes managed via specialized open-source software for document-oriented databases. This facility is offered by the gCube Storage Manager, a Java based software that presents a unique set of methods for services and applications running on the e-Infrastructure. In its current implementation, three possible document store systems are used [24]: MongoDB, Terrastore and USTO.RE [25]. The Storage Manager was designed to reduce the time required to add a new storage system to the e-Infrastructure. This promotes openness versus other document stores, e.g., CouchDB [26], while hiding the heterogeneous protocols of those systems from the services and applications exploiting the infrastructure storage facility.

Virtual Research Environments Management This application offers facilities for dynamically creating and managing virtual research environments [2, 11], i.e., internet-based working

environments tailored to serve the needs of diverse and evolving user communities. The application supports the specification and deployment of complete VREs in terms of the data and services they should offer by automatically acquiring and aggregating the needed resources including the user interface constituents from the infrastructure [27, 28].

Workflow Management Facilities This application offers facilities for executing complex processes, i.e., a workflow of tasks. It includes a process execution engine (PE2ng) that manages the execution of software elements in a distributed infrastructure under the coordination of a composite plan that defines the data dependencies among its actors. It provides a powerful, flow-oriented processing model that supports several computational middleware without performance compromises [29]. Thus, a process can be designed as a workflow of invocations of components (including, services, binary executables, scripts, map-reduce jobs) by ensuring that prerequisite data are prepared and delivered to their consumers through the control of the flow of data. PE2ng aims to bring together and integrate computing paradigms, execution patterns and Infrastructures. Overall, an unrestrictive meta-Infrastructure is offered with a single submission, monitoring and access execution point offering a single language for "Programming in the Large" [30].

2.6 StatsCube

StatsCube is a gCube application bundle offering facilities to practitioners working with a rich array of information, ranging from observational data to statistical data. In particular, it includes applications for data analytics at scale and applications for tabular data and code lists management.

The Statistical Service offers facilities for efficiently and effectively executing a rich array of statistical data processing algorithms [31]. The application relies on the distributed and elastic computing capacities offered by the underlying infrastructure. It offers a set of off-the-shelf algorithms including clustering algorithms such as DBScan. Moreover, it enables a simple integration and execution of user-defined algorithms expressed in a number of programming and scripting languages including R [32]. It currently embeds more than 100 different algorithms ranging from Anomalies Detection, Classification, Clustering, Simulation, Training, Bayesian Methods, Trends, and many more [33]. These algorithms are then executed on a distributed infrastructure by completely hiding the complexity of such an execution while ensuring robustness, throughput, fault-tolerance, and privacy.

The Tabular Data Manager offers facilities for discovery, management and processing of tabular data. In particular, it offers facilities for supporting the entire workflow of tasks on tabular data including tabular data creation, collaborative curation and publishing. It offers also a number of facilities for tabular data manipulation including filtering, grouping, unions and intersections. In addition to that, it is equipped with a powerful mechanism for versioning and a rich set of metadata for describing the tabula data resource including provenance. For discovery, it offers both a Google-like approach and an advanced search allowing users to characterise in detail the information they are looking for. For the processing, it relies on the Statistical Manager to offer effective tabular data manipulation facilities including geocoding, maps projection, clustering, outlier identification, hidden trends, trends comparison, and many more.

COTRIX offers facilities for code lists, i.e., recognised controlled vocabularies, management. It includes facilities for code lists creation (also via ingestion), collaborative curation, and publishing. It offers both a Google-like approach and an advanced search allowing users to characterise in detail the information they are looking for.

3. The D4Science infrastructure

The D4Science infrastructure was designed, developed and put in production back in 2007 with the support of a series of EU projects [34]. It started to serve several use cases and communities ranging from Fisheries to Digital Libraries, Earth Observation and Biodiversity, and it has acquired its maturity specializing its scope serving Biodiversity scientific communities (both marine and land). The two main user communities at the moment are those participating to the EU project iMarine[1] and EUBrazilOpenBio[2] [12, 14].

The D4Science infrastructure is geographically distributed by design and its main feature is to enable interconnections among different technology providers and data providers. In addition, it offers interoperability at the level of the computing infrastructure, such as cloud and grid computing.

The D4Science infrastructure hosting resources dedicated to iMarine and EUBrazilOpenBio are provided by project members plus an external partner (ASGC Taiwan). Therefore in total 6 projects members sites contribute to the infrastructure: CNR - Pisa, Italy; FAO - Rome, Italy; NKUA - Athens, Greece; VLIZ - Ostende, Belgium; UFF - Niteroi, Brazil; UPV - Valencia, Spain. In addition an external projects partner is also providing resources, namely ASGC - Taiwan.

Table 1 provides detailed information about the contribution from each site. The column "type" either reports Hardware (HW) together with type of CPU or Virtual Machine (VM) and the type of virtualization system.

Site	Type	Resource	RAM (GB)	Disk (TB)	CPUs
ASGC	HW	Two Quad-Core Intel Xeon®CPU 5130 @ 2.00GHz	8	0.2	8
	HW	Quad-Core Intel Xeon®CPU 5150 @ 2.66GHz	4	0.1	4
CNR	VM	Xen hypervisor	844	35.4	754
FAO	HW	Two Quad-Core Intel Xeon®X5450 @ 3.0 GHz	8	0.3	8
NKUA	VM	Xen hypervisor	106.25	1.8	56
VLIZ	HW	Intel Xeon®CPU E5649 @ 2.53GHz	3	0.4	4
UFF	HW	Intel Xeon®CPU 3.06 GHz	2.5	0.5	2
UPV	HW	Intel Xeon®CPU 2.00 GHz	38	0.75	16
Total			1011.25	39.5	852

Table 1: D4Science Infrastructures: resources by partner

Part of the the resources allocated to the infrastructure are dedicated to the pre-production infrastructure, also called Quality Assurance environment where the software is validated before reaching the production environment. For what concerns the production infrastructure, the hosted resources can be categorized in 3 main areas:

[1] http://www.i-marine.eu

[2] http://www.eubrazilopenbio.eu

gCube Resources are dedicate to host gCube Service containers (both gHN and tomcat);

UMD Resources are dedicated to host the UMD middleware[3], 2 sites (CNR and NKUA) host production services for the EGI infrastructure[4]

Third Party Resources are hosting third party services classified under their functional category:

- Clusters: MongoDB, Cassandra, Hadoop, OGC Services (Geoserver, Thredds, North52 WPS), CouchBase, ElasticSearch;
- Services: ActiveMQ MessageBroker, JackRabbit, RStudio;
- Databases: several PostgreSQL and MySQL Database instances;

In addition to hosted resources the infrastructure exploits external resources via federated access. This includes access to resources agreed upon signed Memorandum of Understanding or collaborations with project members. This includes the resources offered by the EGI infrastructure which extend the storage and computing capacity available under the D4Science Infrastructure. In detail, the EGI sites supporting the D4science infrastructure Virtual Organisation (d4science.research-infrastructures.eu) are in Table 2.

Site Short Name	Site Official Name	CREAM	WN	SE	CPUs	Storage (TB)
INFN-TRIESTE	INFN-TRIESTE	✓	✓	✓	2380	3.7
Taiwan-LCG2	Academia Sinica Grid Computing Center	✓	✓	✓	240	0.05
Total					2620	3.75

Table 2: EGI services supporting the D4Science Virtual Organisation

In addition to grid resources, by collaborating with the VENUS-C project[5], D4Science has been granted access to Microsoft Azure[6] computation and storage cloud resources, namely 1.5 M CPU Hours and 1.5 TB Storage.

For the data, the D4Science infrastructure offers services for seamless access to a wide spectrum of data including *species data* (cf. Tab. 3), *geospatial data* (cf. Tab. 4), *statistical data* (cf. Tab. 5), and *semi-structured data* (cf. Tab. 6) from multiple data providers and information systems.

4. Virtual Research Environments

Virtual Research Environments (VREs) are " dedicated systems" that provide their users with a web-based set of facilities (including services, data and computational facilities) to accomplish a set of tasks by dynamically relying on the underlying infrastructure.

The development of VREs is actually based on three main activities: two of them are "preparatory" and consist in (*i*) the development of software artefacts that realise a set of functions (cf. Sec

[3] http://repository.egi.eu/category/umd_releases/
[4] http://www.egi.eu
[5] http://www.venus-c.eu/
[6] https://www.windowsazure.com/

Data source	Description
Catalogue of Life	The data source offers an integrated checklist and a taxonomic hierarchy of more than 1.3 million species of animals, plants, fungi and micro-organisms.
FAO ASFIS	The List of Species for Fishery Statistics Purpose includes 12,000+ species of interest or relations to fisheries and aquaculture; `www.fao.org/fishery/collection/asfis/en`.
GBIF	The data source offers more than 430 million of records on species and more than 14,000 datasets aggregated from 580+ publishers; `www.gbif.org`.
Fishbase	The data source offers access to 32,700 Species, 302,900 Common names, 53,600 Pictures, 49,700 References aggregated thanks to the effort of thousand collaborators.
IRMNG	The Interim Register of Marine and Nonmarine Genera data source offers access to over 465,000 genus names and 1.6 million species names; `www.obis.org.au/irmng`.
ITIS	The Integrated Taxonomic Information System data source offers authoritative taxonomic information on plants, animals, fungi, and microbes of North America and the world; `www.itis.gov`.
NCBI Taxonomy	The National Center of Biotechnology Information data source offers a curated classification and nomenclature for all of the organisms in the public sequence databases. This currently represents about 10% of the described species of life on the planet; `www.ncbi.nlm.nih.gov/taxonomy`.
OBIS	The Ocean Biogeographic Information System data source offers more that 37 million records on species and 1,300+ datasets; `www.iobis.org`.
SeaLifeBase	The data source offers access to 126,000 Species, 27,300 Common names, 11,900 Pictures, 18,200 References aggregated thanks to the effort of hundred collaborators.
WoRMS	The World Register of Marine Species data source offers species "names" for more than 200,000 species including 300,000+ species names and synonyms and 400,000+ taxa; `www.marinespecies.org`.
WoRDSS	The World Register of Deep-Sea Species data source offers species "names" for deep-sea species based on WoRMS. `www.marinespecies.org/deepsea`.

Table 3: Species Data Databases and Information Systems Integrated in D4Science

Data source	Description
FAO GeoNetwork	The data source exposes spatial data maintained by FAO and its partners; `geonetwork.fao.org`.
World Ocean Atlas	The data source gives access to a number of environmental variables. In particular, iMarine focuses on some indicators including Apparent Oxygen Utilisation, Dissolved Oxygen, Nitrate, Oxygen Saturation, Phosphate, Sea Water Salinity, Sea Water Temperature, and Silicate; `www.nodc.noaa.gov/OC5/WOA09/pr_woa09.html`.
Marine Regions	The data source gives access to a standard list of marine georeferenced place names and areas including EEZ; `www.marineregions.org`.
myOceans	The data source gives access to a number of environmental variables. In particular, D4Science focuses on some indicators including ice concentration, ice thickness, ice velocity, mass concentration of chlorophyll in sea water, meridional velocity, mole concentration of dissolved oxygen in sea water, mole concentration of nitrate in sea water, mole concentration of phosphate in sea water, mole concentration of phytoplankton expressed as carbon in sea water, net primary production of carbon, salinity, sea surface height, temperature, zonal velocity, wind speed, and wind stress. `www.myocean.eu`.

Table 4: Spatial Data Databases and Information Systems Integrated in D4Science

Data source	Description
IRD Datasets	The UMR EME/Observatoire Thonier SDMX Registry and Repository exposes the Sardara database that contains tuna captures data from several countries, aggregated according to CWP statistical squares (1'x1' or 5'x5') and the ObServe database that contains tuna and bycatches captures observed by scientific observers on-board of French industrial purse seiners.
Codelists	A set of SDMX Codelists either directly accessed from the FAO Registry, or manually uploaded through the facility developed in the context of ICIS.
StatBase	This data source collects and organises data about several sectors including Agriculture, Education, Energy, Environment, Industry, Population. Data are collected from several data providers including African Development Bank, Central Bank of Central African States, Freedom House, International Energy Agency, OECD, United Nations Industrial Development Organization.

Table 5: Statistical Data Databases and Information Systems Integrated in D4Science

2); and (*ii*) the deployment of these artefacts in an operational infrastructure and the population of the infrastructure itself (cf. Sec 3). The third activity is the actual deployment and operation of a VRE which thanks to gCube and the D4Science infrastructure is a very straightforward activity consisting of: (*i*) a *design* phase where authorised users are provided with a wizard-based approach to specify the data and the services characterising the envisaged environment by selecting among the available ones; (*ii*) a *deployment* phase where authorised users are provided with a wizard-based approach to approve a VRE specification and monitor the automatic deployment of the real components needed to satisfy the specification; and (*iii*) an *operation* phase where authorised users are provided with facilities for managing the users of the VRE and altering the VRE specification if needed. Details on this approach have been presented in previous works [27, 28] while a screenshot of the wizard supporting the VRE specification is in Figure 2.

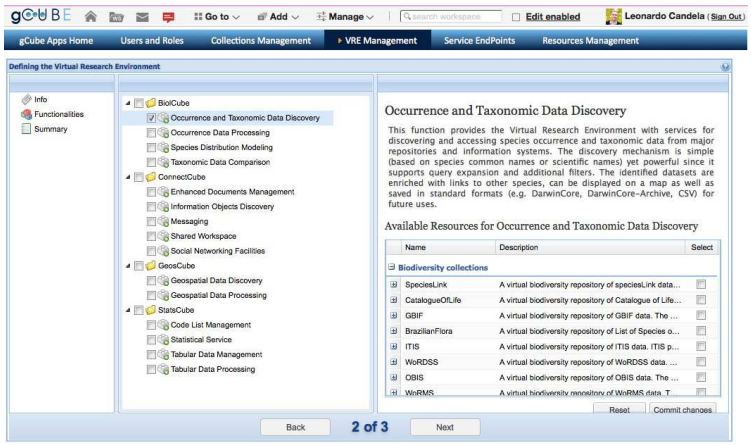

Figure 2: Virtual Research Environment definition phase: selecting the expected facilities

At the end of March 2014, 21 VREs are concurrently hosted and operated by the D4Science infrastructure. Some of them have been in constant use for over four years with over 750 users. A detailed list is in Table 7 where for each VRE it is reported: the name of the VRE, the domain the VRE is serving, whether the VRE membership is "open", i.e., any user can join it, or not, and the number of users.

Data source	Description
Aquatic Commons	The data source offers access to thematic material covering natural marine, estuarine/brackish and fresh water environments; aquaticcommons.org.
BHL	The Biodiversity Heritage Library data source offers access to legacy literature of biodiversity held by a consortium of natural history and botanical libraries; www.biodiversitylibrary.org.
Bioline	The Bioline International data source offers access to open access quality research journals published in developing countries; www.bioline.org.br.
CEEMar	The Central and Eastern European Marine Repository data source offers material covering marine, brackish and fresh water environments; www.ceemar.org/dspace.
DataCite	The data source offers access to the same service whose mission is to give access to research data; www.datacite.org.
DBPedia	The knowledge base results from Wikipedia. It contains over 4 millions things including persons, places, creative works, organisations, species and diseases; dbpedia.org/About.
DRS	The data source at National Institute of Oceanography offers institutional publications including journal articles and technical reports; drs.nio.org/drs.
Dryad	The data source offers access to the same service whose mission is to give access to research data underlying research publications; datadryad.org.
FactForge	The knowledge base results from the integration of a number of datasets including DBPedia, WordNet, Geonames, and Freebase; factforge.net.
FAO Factsheets	The data source gives access to the Aquatic Species Fact Sheets developed by the same FAO programme; www.fao.org/fishery/fishfinder.
FAO FLOD	A semantic knowledge base hosted by FAO containing a dense network of relationships among the major entities of the fishery domain, including marine species, water areas, land areas, and exclusive economic zones; www.fao.org/figis/flod.
iMarine TLO	The warehouse integrates information from FishBase, WoRMS, ECOSCOPE, FLOD and DBPedia by using the same top-level ontology developed for the marine domain. It currently contains approximately 3 millions of triples about more than 40,000 entities including marine species, ecosystems, water areas, and vessels;
Nature	The data source offers access to the articles published by nature.com.
OceanDocs	The data source offers research and publication materials in Marine Science by aggregating content form 256 repositories; www.oceandocs.net.
OpenAIRE	The data source give access to the publications aggregated by the same European funded project; www.openaire.eu.
PANGAEA	The data source offers georeferenced data from earth system research via OAI-PMH. The system guarantees long-term availability of its content through a commitment of the operating institutions. The aggregated repositories are 475; www.pangaea.de.
PenSoft Journals	The data source gives access to a number of open-access journals. In particular, iMarine focuses on BioRisk, Comparative Cytogenetics, International Journal of Myriapodology, Journal of Hymenoptera Research, MycoKeys, Nature Conservation, NeoBiota, PhytoKeys, Subterranean Biology, and ZooKeys.
SmartFish	The SmartFish Chimaera knowledge base offers a unified and integrated view on three marine fisheries information sources, i.e., FIRMS – an international knowledge base including fisheries and resource from West Indian Ocean; StatBase – a statistical database containing statistics provided by West Indian Ocean countries; and WIOFish – a regional knowledge base on West Indian Ocean Fisheries.
WHOAS	The data source offers the production of Woods Hole scientific community including articles and data sets; www.mblwhoilibrary.org/services/whoas-repository-services.
YAGO2	The knowledge base extends the YAGO knowledge base by anchoring entities, facts and events in time and space. The knowledge base is built from Wikipedia, GeoNames and WordNet and contains more than 440 million facts about 9.8 million entities.

Table 6: Various Databases and Information Systems Integrated in D4Science

VRE Name	Domain	Open	Users
AquaMaps	Biodiversity		57
BiodiversityLab	Biodiversity		64
BiodiversityResearchEnvironment	Biodiversity	✓	61
DocumentsWorkflow	Any	✓	40
EcologicalModeling	Biodiversity	✓	59
ENVRI	Environment		22
FCPPS	Fisheries		32
FishFinderVRE	Fisheries		12
gCube	Software		43
ICIS	Fisheries		43
iMarineBoard	Policy		21
iSearch	Any	✓	26
MarineSearch	Marine		4
ScalableDataMining	Analytics	✓	41
SpeciesLab	Biodieversity		50
TBTI	Fisheries		10
TCom	Software		35
TimeSeries	Any	✓	52
VesselActivitiesAnalyzer	Analytics	✓	43
VTI	Analytics		28
VME-DB	Fisheries		10

Table 7: D4Science Virtual Research Environments

In some cases the users served by the VRE represent a small fraction of the community benefitting from the VRE services, e.g., the AquaMaps VRE [35, 13] is actually exploited by data managers to produce the maps disseminated via the AquaMaps service.

5. Conclusions

Modern science calls for innovative working environments crossing the boundaries and capacities of single scientists, laboratories and institutions.

In this paper we have presented one of such innovative working environments, i.e., that offered by the D4Science organisation. This organisation is making available a Hybrid Data Infrastructure enabling the dynamic deployment and operation of an array of Virtual Research Environments, each tailored to serve the needs of a scientific community towards a research endeavour. We have presented gCube, i.e., the enabling technology, as well as the resulting infrastructure and the existing VREs. The served communities and scenarios have demonstrated that the proposed approach is suitable and can be applied to a range of scientific applications.

The gCube technology is in constant evolution and its community is particularly active[7]. Future activities and work will mainly focus on three typologies of actions: (*i*) the creation of new virtual research environments to serve the needs of new scenarios by benefitting from the rich array of facilities so far developed; (*ii*) the development of plug-ins and mediator services enlarging

[7]In the period August '13 – August '14, 40 contributors have been involved in its development with more than 11,000 software commits `https://www.openhub.net/p/gCube`

the set of data sources integrated in the D4Science infrastructure; and (*iii*) the development of new algorithms and approaches aiming at enlarging the offering of the Statistical Manager service. Thanks to the openness of the gCube system, some of these developments can be performed by the community in the large, e.g., every scientists owning an algorithm worth to share can decide to integrate it into the Statistical Manager and benefit from a boost in performances [32].

Acknowledgements The work reported has been partially supported by the *iMarine* project (FP7 of the European Commission, FP7-INFRASTRUCTURES-2011-2, Contract No. 283644).

References

[1] L. Candela, D. Castelli, and P. Pagano. Managing big data through hybrid data infrastructures. *ERCIM News*, (89):37–38, 2012.

[2] L. Candela, D. Castelli, and P. Pagano. Virtual research environments: an overview and a research agenda. *Data Science Journal*, 12:GRDI75–GRDI81, 2013.

[3] gCube Development Team. gCube Website. `https://www.gcube-system.org`, 2008.

[4] gCube Development Team. gCube Developer's Guide. `http://gcube.wiki.gcube-system.org/gcube/index.php/Developer\%27s_Guide`, 2008.

[5] L. Candela, D. Castelli, and P. Pagano. gCube v1.0: A Software System for Hybrid Data Infrastructures. Technical Report 2008-TR-035, Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", CNR, 2008.

[6] I. Foster and C. Kesselman. *The Grid: Blueprint for a Future Computing Infrastructure*. Morgan-Kaufmann, 2004.

[7] M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia. A view of cloud computing. *Communications of the ACM*, 53(4):50–58, April 2010.

[8] L. Candela, D. Castelli, N. Ferro, Y. Ioannidis, G. Koutrika, C. Meghini, P. Pagano, S. Ross, D. Soergel, M. Agosti, M. Dobreva, V. Katifori, and H. Schuldt. *The DELOS Digital Library Reference Model - Foundations for Digital Libraries*. DELOS: a Network of Excellence on Digital Libraries, February 2008. ISSN 1818-8044 ISBN 2-912335-37-X.

[9] M.N. Huhns and M.P. Singh. Service-oriented computing: key concepts and principles. *IEEE Internet Computing*, 9:75–81, 2005.

[10] E. Wenger. *Communities of Practice: Learning, Meaning and Identity*. Cambridge University Press, 1998.

[11] L. Candela, D. Castelli, and P. Pagano. Making Virtual Research Environments in the Cloud a Reality: the gCube Approach. *ERCIM News*, (83):32–33, October 2010.

[12] L. Candela, D. Castelli, G. Coro, L. Lelii, F. Mangiacrapa, V. Marioli, and P. Pagano. An infrastructure-oriented approach for supporting biodiversity data. *Ecological Informatics*, n/a:n/a, 2014. Article first published online: 1 August 2014 DOI: 10.1016/j.ecoinf.2014.07.006

[13] L. Candela, D. Castelli, G. Coro, P. Pagano, and F. Sinibaldi. Species distribution modeling in the cloud. *Concurrency and Computation: Practice and Experience*, n/a:n/a, 2013. Article first published online: 11 July 2013 `http://onlinelibrary.wiley.com/doi/10.1002/cpe.3030/abstract`.

[14] R. Amaral, R. M. Badia, I. Blanquer, Ricardo Braga-Neto, L. Candela, D. Castelli, C. Flann, R. De Giovanni, W. A. Gray, A. Jones, D. Lezzi, P. Pagano, V. Perez-Canhos, F. Quevedo, R. Rafanell, V. Rebello, M. Sousa-Baena, and E. Torres. Supporting biodiversity studies with the EUBrazilOpenBio Hybrid Data Infrastructure. *Concurrency and Computation: Practice and Experience*, n/a:n/a, 2014.

[15] E. Vanden Berghe, N. Bailly, G. Coro, F. Fiorellato, C. Aldemita, A. Ellenbroek, and P. Pagano. BiOnym - a flexible workflow approach to taxon name matching. *Technical Report 2014-TR-022*, Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", CNR 2014.

[16] M. Assante, L. Candela, D. Castelli, and P. Pagano. The D4Science Research-Oriented Social Networking Facilities. *ERCIM News*, 96:n/a, 2014.

[17] M. Assante, L. Candela, and P. Pagano. An environment supporting the production of live research objects. *The Grey Journal*, 9(1), 2013.

[18] F. Simeoni, L. Candela, G. Kakaletris, M. Sibeko, P. Pagano, G. Papanikos, P. Polydoras, Y.E. Ioannidis, D. Aarvaag, and F. Crestani. A Grid-Based Infrastructure for Distributed Retrieval. In L. Kovács, N. Fuhr, and C. Meghini, editors, *Research and Advanced Technology for Digital Libraries, 11th European Conference, ECDL 2007, Budapest, Hungary, September 16-21, 2007, Proceedings*, volume 4675 of *Lecture Notes in Computer Science*, pages 161–173. Springer-Verlag, 2007.

[19] P. Fafalios and Y. Tzitzikas Post-Analysis of Keyword-based Search Results using Entity Mining, Linked Data and Link Analysis at Query Time. IEEE 8th International Conference on Semantic Computing (ICSC'14), Newport Beach, California, USA, June 2014

[20] Y. Tzitzikas, C. Allocca, C. Bekiari, Y. Marketakis, P. Fafalios, M. Doerr, N. Minadakis, T. Patkos and L. Candela Integrating Heterogeneous and Distributed Information about Marine Species through a Top Level Ontology. Proceedings of the 7th Metadata and Semantic Research Conference, MTSR'13, Thessaloniki, Greece, November 2013.

[21] Candela L., Coro G., Cossu R., Pagano P. Realizing spatial data infrastructure solutions in ENVRI. In: GEPW-8 - GEO European Projects' Workshop (Athens, Greece, 12-13 June 2014).

[22] OpenGIS Catalogue Service `http://www.opengeospatial.org/standards/cat`

[23] OpenGIS Web Processing Service `http://www.opengeospatial.org/standards/wps`

[24] R. Cattell. Scalable SQL and NoSQL data stores. *SIGMOD Rec.*, 39(4):12–27, May 2011.

[25] F.A. Durão, R.E. Assad, A.F. Silva, J.F. Carvalho, V.C. Garcia, and F.A.M. Trinta. USTO.RE: A Private Cloud Storage System. In *13th International Conference on Web Engineering (ICWE 2013) - Industry track*, Aalborg, 2013.

[26] J.C. Anderson, J. Lehnardt, and N. Slater. *CouchDB: The Definitive Guide*. O'Really, 2009.

[27] M. Assante, L. Candela, D. Castelli, L. Frosini, L. Lelii, P. Manghi, A. Manzi, P. Pagano, and M. Simi. An Extensible Virtual Digital Libraries Generator. In B. Christensen-Dalsgaard, D. Castelli, B.A. Jurik, and J. Lippincott, editors, *12th European Conference on Research and Advanced Technology for Digital Libraries, ECDL 2008, Aarhus, Denmark, September 14-19*, volume 5173 of *Lecture Notes in Computer Science*, pages 122–134. Springer, 2008.

[28] M. Assante, P. Pagano, L. Candela, F. De Faveri, and L. Lelii. An approach to virtual research environment user interfaces dynamic construction. In *Proceedings of the International Conference on Theory and Practice of Digital Libraries (TPDL 2011)*, volume 6966 of *Lecture Notes in Computer Science*, pages 101–109. Springer, 2011.

[29] M.M. Tsangaris, G. Kakaletris, H. Kllapi, G. Papanikos, F. Pentaris, P. Polydoras, E. Sitaridi, V. Stoumpos, and Y.E. Ioannidis. Dataflow processing and optimization on grid and cloud infrastructures. *IEEE Data Eng. Bull.*, 32(1):67–74, 2009.

[30] G. Wiederhold, P. Wegner, and S. Ceri. Toward Megaprogramming. *Communication of the ACM*, 38(11):89–99, November 1992.

[31] G. Coro, P. Pagano, and L. Candela. Providing statistical algorithms as-a-service. TDWG 2013 - Taxonomic Database Working Group 2013 (Firenze, 28-31 October 2013), 2013. Abstract.

[32] G. Coro, L. Candela, P. Pagano, A. Italiano and L. Liccardo. Parallelising the Execution of Native Data Mining Algorithms for Computational Biology. *Concurrency and Computation: Practice and Experience*, n/a:n/a, 2014. Under review.

[33] G. Coro and L. Candela. gCube statistical manager: the algorithms. *Technical Report 2014-TR-027*, Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", CNR 2014.

[34] P. Andrade, L. Candela, D. Castelli, A. Manzi, and P. Pagano. The D4Science Production Infrastructure. Technical Report 2009-TR-054, Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", CNR, 2009.

[35] L. Candela and P. Pagano. The D4Science Approach toward Grid Resource Sharing: The Species Occurrence Maps Generation Case. In Simon C. Lin and Eric Yen, editors, *Data Driven e-Science - Use Cases and Successful Applications of Distributed Computing Infrastructures (ISGC 2010)*, pages 225–238. Springer, 2011.

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