Accepted Manuscript

The gCube system: Delivering Virtual Research Environments as-a-Service

M. Assante, L. Candela, D. Castelli, R. Cirillo, G. Coro, L. Frosini, L. Lelii, F. Mangiacrapa, V. Marioli, P. Pagano, G. Panichi, C. Perciante, F. Sinibaldi

PII: S0167-739X(17)32836-4
DOI: https://doi.org/10.1016/j.future.2018.10.035
Reference: FUTURE 4540

To appear in: *Future Generation Computer Systems*

Received date : 15 October 2018
Accepted date : 20 October 2018

Please cite this article as: M. Assante, et al., The gCube system: Delivering Virtual Research Environments as-a-Service, *Future Generation Computer Systems* (2018), https://doi.org/10.1016/j.future.2018.10.035

The gCube System: Delivering Virtual Research Environments as-a-Service

M. Assante[a], L. Candela[a,*], D. Castelli[a], R. Cirillo[a], G. Coro[a], L. Frosini[a,b], L. Lelii[a], F. Mangiacrapa[a], V. Marioli[a], P. Pagano[a], G. Panichi[a], C. Perciante[a,b], F. Sinibaldi[a]

[a]*Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" – Consiglio Nazionale delle Ricerche – via G. Moruzzi, 1 – 56124, Pisa, Italy*
[b]*Dipartimento di Ingegneria dell'Informazione – Università di Pisa – Via G. Caruso 16 – 56122 – Pisa, Italy*

Abstract

Important changes have characterised research and knowledge production in recent decades. These changes are associated with developments in information technologies and infrastructures. The processes characterising research and knowledge production are changing through the digitalization of science, the virtualisation of research communities and networks, the offering of underlying systems and services by infrastructures. This paper gives an overview of gCube, a software system promoting elastic and seamless access to research assets (data, services, computing) across the boundaries of institutions, disciplines and providers to favour collaboration-oriented research tasks. gCube's technology is primarily conceived to enable *Hybrid Data Infrastructures* facilitating the dynamic definition and operation of *Virtual Research Environments*. To this end, it offers a comprehensive set of data

*Corresponding author

Email addresses: massimiliano.assante@isti.cnr.it (M. Assante), leonardo.candela@isti.cnr.it (L. Candela), donatella.castelli@isti.cnr.it (D. Castelli), roberto.cirillo@isti.cnr.it (R. Cirillo), gianpaolo.coro@isti.cnr.it (G. Coro), luca.frosini@isti.cnr.it (L. Frosini), lucio.lelii@isti.cnr.it (L. Lelii), francesco.mangiacrapa@isti.cnr.it (F. Mangiacrapa), valentina.marioli@isti.cnr.it (V. Marioli), pasquale.pagano@isti.cnr.it (P. Pagano), giancarlo.panichi@isti.cnr.it (G. Panichi), costantino.perciante@isti.cnr.it (C. Perciante), fabio.sinibaldi@isti.cnr.it (F. Sinibaldi)

management commodities on various types of data and a rich array of "mediators" to interface well-established Infrastructures and Information Systems from various domains. Its effectiveness has been proved by operating the D4Science.org infrastructure and serving concrete, multidisciplinary, challenging, and large scale scenarios.

Keywords: Virtual Research Environments, Social Networking, Science Gateway

1. Introduction

Research and knowledge production practices have been changed in recent decades [1, 2]. These changes are induced by developments in information technologies and infrastructures. The changes are impacting the whole research lifecycle – from data collection and curation to analysis, visualisation and publishing. Science is digital, collaborative and multidisciplinary, research communities are dynamically aggregated, working environments conceived to support research tasks are virtual, heterogeneous and networked across the boundaries of research performing organisations. Scientists are thus asking for integrated environments providing themselves with seamless access to data, software, services and computing resources they need in performing their research activities independently of organisational and technical barriers. In these settings, approaches based on ad-hoc and "from scratch" development of the envisaged supporting environments are neither viable (e.g. high "time to market") nor sustainable (e.g. technological obsolescence risk).

The *as-a-Service* delivery model promoted by the Cloud computing [3, 4] is a suitable model for providing scientists with the services, infrastructures and environments they are expecting [5]. This model consists in services (a) delivered over the Internet rather than provided locally or on-site and (b) managed by professional and dedicated providers rather than the primary consumers. This model makes it possible (i) for the service provider, to leverage economies of scale to keep developments and operational costs low; (ii) for the service consumer, to acquire the services and the capacity needed in an elastic way. The presence of services delivered with the as-a-Service delivery model is potentially reducing the efforts and costs needed to implement research supporting environments yet it is not nullifying neither the efforts nor the costs. Depending on the typologies of service(s) that are made available by service providers, scientists might be requested to implement and operate on their own what is missing to get the research supporting environment they need. In practice, there might be a functional mismatch between scientists' expectations and service providers' offerings. For instance, in the case of Infrastructure-as-a-Service, scientists (actually, technical staff supporting them) are supported in the creation of virtual machines (compute, storage, networking) yet they have to install and configure on their own the software they need.

The consumption of the available services and their exploitation to realise the scientific workflows should be an easy task that does not require extra skills nor distract effort from the pure scientific investigation (long learning curves, "entry barriers"). Science Gateways (SGs) [6] and Virtual Research Environments (VREs) [7] have been proposed to close the gap between service providers' offerings and scientific communities' expectations. Very often SGs / VREs are ad-hoc portals built to serve the needs of a specific community only. The development of such environments aiming at facilitating scientists tasks is challenging from the system engineering perspective. Several technologies and skills are needed [8, 9]. Moreover, (*a*) people having the requested expertise (mainly expertise related with IT) are not always available in the scientific contexts calling for VREs, and (*b*) technology is in continuous evolution thus offering new opportunities for implementing existing facilities in innovative ways or integrating innovative facilities in existing VREs. This should discourage scientific communities from building their own solutions. Rather it should suggest them to outsource the task of developing and operating Virtual Research Environments to providers delivering them with the as-a-Service model.

This paper gives an overview of gCube[1]. gCube is a software system specifically conceived to enable the creation and operation of an innovative typology of e-Infrastructures, i.e. *Hybrid Data Infrastructures*, that by aggregating a wealth of resources from other infrastructures offers cutting-edge *Virtual Research Environments as-a-Service*. gCube complements the offerings of the aggregated infrastructures by implementing a comprehensive set of value-added services supporting the entire data management lifecycle in accordance with collaborative, user friendly, and Open Science compliant practices. gCube supports the D4Science.org infrastructure that hosts hun-

[1] gCube Software System website `www.gcube-system.org`

dreds Virtual Research Environments to serve the biological, ecological, environmental, social mining and humanities communities world-wide.[2] Overall, the VREs are connecting more than 6000 scientists spread all over the world.

2. gCube System Overview

In order to offer VREs as-a-Service, the gCube system has been designed according to a number of guiding principles described below.

Component orientation. gCube is primarily organised in a number of physically distributed and networked *services*. These services offer functionality that can be combined together. In addition, it consists of (a) auxiliary components (*software libraries*) supporting service development, service-to-service integration, and service capabilities extension, and (b) components dedicated to realise user interfaces (*portlets*) facilitating the exploitation of one or more services.

Autonomic behaviour. Some components are dedicated to manage the operation of a gCube-based infrastructure and its constituents, e.g. automatic (un-)deployment, relocation, replication. These components realise a middleware providing the resulting infrastructure with an autonomic behaviour that reduces its deployment and operation costs.

Openness. gCube supplies a set of generic frameworks supporting data collection, storage, linking, transformation, curation, annotation, indexing and discovery, publishing and sharing. These frameworks are oriented to capture the needs of diverse application domains through their rich adaptation and customisation capabilities [10].

System of Systems. gCube includes components realising a rich array of mediator services for interfacing with existing "systems" and their enabling technologies including middlewares for distributed computing (e.g. EMI [11]), cloud (e.g. Globus [12], OCCI [13]) and data repositories (e.g. OAI-PMH [14], SDMX [15], OGC WP*[3].). Via these mediator services, the storage facili-

[2] The list of supported VREs is evolving and it is always available at https://services.d4science.org/explore.

[3] Open Geospatial Consortium Standards and Supporting Documents http://www.opengeospatial.org/standards

ties, processing facilities and data resources of external infrastructures are conceptually unified to become gCube resources.

Policy-driven resources sharing. gCube manages a resource space where (*a*) resources include gCube-based services as well as third party ones, software libraries, portlets and data repositories, (*b*) resources exploitation and visibility is controlled by policies realising a number of overlay sets on the same resource space. This approach is key to have a flexible and dynamic mechanism for VRE creation, since VREs are actually realised through dynamic aggregations of resources.

As a Service. The gCube offering is exposed according to the "as a Service" delivery model [3]. The advantage is that the actual management is in the hand of expert operators who manage the infrastructure (*i*) by providing reliable services, (*ii*) by leveraging economies of scale, and (*iii*) by using elastic approaches to scale. Via gCube nodes (servers enriched with microservices) the system offers storage and computing capacities as well as management of service instances (dynamic (un)deployment, accounting, monitoring, alerting). Via gCube APIs the system gives a flexible and powerful platform to which developers can outsource data management tasks. Via gCube services the system offers a number of ready to use applications.

These guiding principles allow providing VREs as-a-Service, i.e. authorised users can aggregate – by using a wizard – existing resources (including data) to form innovative working environments and make them available via a plain web browser or even via a thin client. Of course, the set of resources that can be aggregated cannot be considered sufficient for any exploitation scenario. Whenever a gap between gCube capabilities and user expectations emerges it must be filled by activities ranging from the development of service *plug-ins* adding capabilities to existing components (e.g. for accessing specific data, exploiting a specific protocol, making available a specific analytics method) to the development of new portlets, web apps, or even new services devised to support domain specific workflows.

gCube key components resulting from the above principles are organised according to Fig. 1.

The *gateway* is a portal based on Liferay technology[4]. In the reality the gateway is conceived to be deployed by a number of servers hosting

[4] Liferay website `https://www.liferay.com/`.

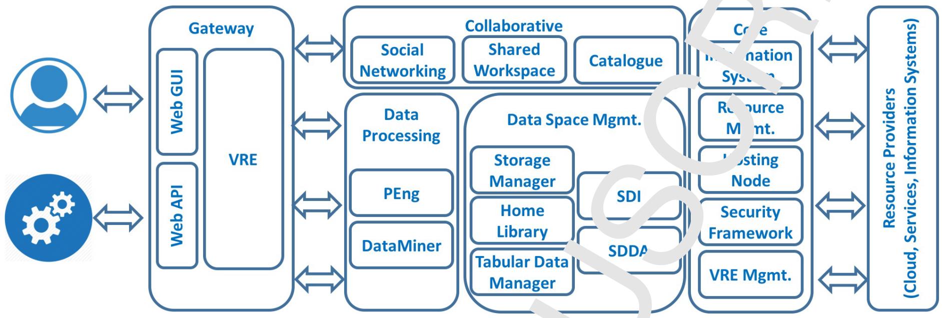


Figure 1: gCube Functional Areas

Liferay instances all made available by a proxy guaranteeing high availability and load balancing. The gateway hosts all the portlets implementing GUI components facilitating access to the services. Moreover, for every VRE a proper application context (i.e. a site) is created with the portlet needed by the specific VRE organised according to the specific needs.

The rest of the areas are described in the next sections.

2.1. Core Services

gCube core services offer the basic facilities for resources management, security, and VRE management described below.

Resource Management. This set of services (Information System, Resource Management Service and Hosting Node) offers facilities for the secure and seamless management (discovery, deployment, monitoring, accounting) of *resources* (proprietary or third party ones) encompassing hosting nodes, services, software and datasets. The *Information System* acts as the registry of the entire infrastructure. It gives global and partial views of the resources and their operational state through query answering or notifications. The *Resource Management Service* is responsible for resource allocation and deployment strategies (e.g. dynamically assigning selected resources to a given context such as a VRE, assigning and activating both gCube software and external software on hosting nodes). The *Hosting Node* is a software component that once installed on a (virtual) machine transforms it into a server managed by the infrastructure and makes it capable to host running instances of services and manage their lifecycle. This type of node can be configured to host a worker service thus making the machine capable to execute computing tasks (cf. Sec. 2.3).

Security Framework. Facilities for authentication and authorisation are supported. They are based on standard protocols and technologies (e.g. SAML 2.0) providing: (*a*) an open and extensible AA architecture; (*b*) interoperability with external infrastructures and domains obtaining Identity Federation (e.g. OpenID). For authorisation, gCube implements a token based authorization system with an attribute-based access control paradigm. For authentication, users are requested to sign-in with their account (including third-party accounts like Google or LinkedIn). Once logged in, the user is provided with a user token that is transparently used to perform calls on behalf of the user. Whenever a user action implies a call to a service requiring authorization, the gCube security framework (*a*) automatically collects the credentials to be used by the credential wallet, i.e. a service where the credentials to access each service are stored in encrypted form by using VRE-specific symmetric keys, (*b*) decrypts the user credentials with the specific VRE key, and (*c*) performs the authorized call by passing the credentials and the user token. In this way the connection to the service is established in a secure way while the token is used to verify that the specific user is authorized to call the service.

VRE Management. Facilities for the specification (wizard-based) and automatic deployment of complete VREs in terms of the data and services they should offer are supported [16]. These facilities are implemented by dynamically acquiring and aggregating the needed resources, including user interface constituents, from the resource space. This 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. This is enacted by allowing users to select the items of interest among the available ones; (*ii*) a *deployment* phase where authorised users are provided with a wizard-based approach to approve a VRE specification. Once approved, the deployment starts and the manager is enacted to 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 [16, 17] while a screenshot of the wizard supporting the VRE specification is in Figure 2.

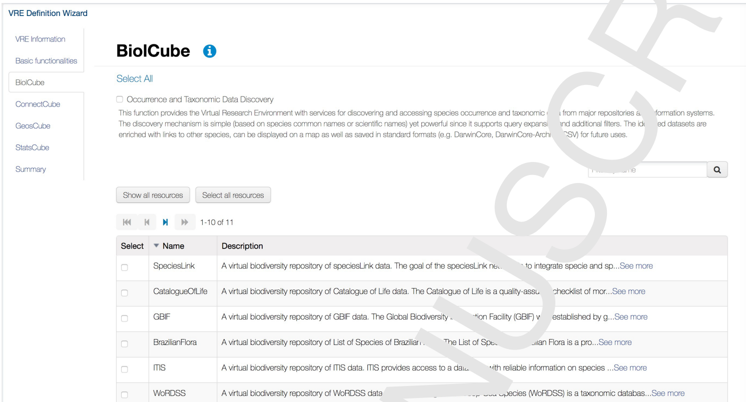


Figure 2: Virtual Research Environment definition phase: selecting the expected facilities

2.2. Data Space Management Services

Data occupies a key role in science and scientific workflows, thus VREs are called to support their effective management. However, data to be managed are very heterogeneous (formats, typologies, semantics), disaggregated and dispersed in multiple sites (including researchers' drawers), possibly falling under the big data umbrella. In the context of a VRE, it is likely that compound information units are produced by using constituents across the various solutions. To cope with this variety, gCube offers an array of solutions ranging from those aiming at abstracting from the heterogeneity of data (file-oriented and information objects) to those focusing on specific data typologies having different levels of semantic embodiment (tabular, spatial, biodiversity data). Independently of data typologies, all these solutions are characterised by (a) support for aggregation of data residing in existing repositories; (b) scalability strategies enabling users to dynamically adding more capacity; (c) comprehensive metadata to capture key aspects like context, attribution, usage licenses, lineage [18]; and (d) policy-driven configurability to adapt the data space to specific needs, e.g. by selecting repositories and datasets. In particular, the following facilities are supported.

File-oriented data. The *Storage Manager* is a Java based software library supporting a unique set of methods for services and applications to manage files efficiently. It relies on a network of distributed storage nodes managed by specialized open-source software for document-oriented databases. In its current implementation, three possible document store systems [19] can be seamlessly used, namely MongoDB, Terrastore and U.STORE [20], while new ones can be added by implementing a specific mediator.

Information Objects. The *Home Library* is a Java based software library enabling objects consisting of a tree of nodes with associated properties. It is compliant with the Java Content Repository API and implemented by relying on Apache Jackrabbit for the object structure and node properties. Node content is outsourced to other services. Every data to be managed in a VRE has a manifestation in terms of Information Objects. The unification of the entire data space makes it possible to realise services across the boundaries of specific data typologies. Among these unifying services there is an innovative search engine [21] that makes it possible to seamlessly discover objects in the data space.

Tabular data. The *Tabular Data Manager* offers a comprehensive and flexible working environment for accessing, curating, analysing and publishing tabular data. It enables a user to ingest data – from a file or a web location – that are represented in formats including CSV, JSON and SDMX. In the case of "free" formats, namely CSV, the system offers facilities to transform data in a well-defined table format where the types of the columns are basic data types including temporal and spatial dimensions as well as references to controlled vocabularies, e.g. Code List. Table formats can be defined in an interactive way as well as by relying on templates. Besides constraints on table column types, templates can contain additional validation rules as well as specification of data operations to be performed when an error occurs. Once a tabular data resource is created, it can be manipulated and analysed by benefitting of well known tabular data operations, e.g. adding columns, filtering, grouping, as well as advanced data analytics tasks (cf. Sec. 2.3). To guarantee a proper and real time management of the data lineage, the service relies on an underlying cluster of RDBMs where any operation on a tabular dataset leads to a new referable version.

Spatial data. gCube owns services realising the facilities of a *Spatial Data Infrastructure* (SDI) by relying on state-of-the-art technologies and standards [22]. It offers standard-based services for data discovery (a catalogue), storage and access (a federation of repositories), and visualisation (a map container). The catalogue service enables the discovery of geospatial data residing in dedicated repositories by relying on GeoNetwork and its indexing facilities. For data storage and access, gCube offers a federation of repositories based on GeoServer and THREDDS technologies. In essence, the infrastructure hosts a number of repositories and a *GIS Publisher Service* that enables a seamless publication of geospatial data while guaranteeing load balancing, failure management and automatic metadata generation. It relies on an open set of back-end technologies for the actual storage and retrieval of the data. Because of this, the GIS Publisher Service is designed with a plug-in-oriented approach where each plug-in interfaces with a given back-end technology. To enlarge the array of supported technologies it is sufficient to develop a dedicated plug-in. Metadata on available data are published by the catalogue. For data visualisation, the infrastructure offers *Geo Explorer* and *GIS Viewer*, two components dedicated to support browsing and visualisation of geospatial data. In particular: Geo Explorer is a web application that allows users to navigate, organize, search and discover layers from the catalogue via the CSW protocol; GIS Viewer is a web application that allows users to interactively explore, manipulate and analyse geospatial data.

Biodiversity data. The *Species Data Discovery and Access Service* (SDDA) [23] provides users with facilities for the management of nomenclature data and species occurrences. As data are stored in authoritative yet heterogeneous information systems, SDDA is mainly conceived to dynamically "aggregate" data from these systems and unify their management. It is designed with a plug-in based architecture. Each plug-in interacts with an information system or database by relying on a standard protocol, e.g. TAPIR, or by interfacing with its proprietary protocol. Plug-ins conform queries and results from the query language as well as data model envisaged by SDDA to the features of a particular database. SDDA promotes a unifying discovery and access mechanism based on the names of the target species, whether the scientific or the common ones. To overcome the potential issues related to taxonomy heterogeneities across diverse data sources, the service supports an automatic query expansion mechanism, i.e. upon request the query is augmented with "similar" species names. Also, queries can be augmented with criteria aiming at explicitly selecting the databases to search in and the spatial and temporal coverage of the data. Discovered data are presented in a homogenised form, e.g. in a typical Darwin Core format.

2.3. Data Processing Services

In addition to the facilities for managing datasets, VREs offer services for processing them. It is almost impossible to figure out "all" the processing tasks needed by scientists, thus the solution is to have environments where scientists can easily plug and execute their tasks. gCube offers two typologies of engines: one oriented to enact tasks executions at system level, another oriented to enact task execution at user level. Both of them are conceived to rely on a distributed computing infrastructure to execute tasks. A description of these two engines is given below.

System oriented workflow engine. The *Process Execution Engine* (PEng) is a system orchestrating flows of invocations (processes). It builds on principles of data flow processing appropriately expanded in the direction of interoperability [24]. According to this, PEng includes the plan (flow of execution), the operators (executable logic), the transport and control abstraction, the containers (areas of execution), the *state* holders (e.g. storage), the *resource profiles* (definitions of resources characteristics for exploitation in a plan). It allows to distribute jobs on several machines. Each job defines an atomic execution of a more complex process. Relations and hierarchies among the jobs are defined by means of Direct Acyclic Graphs (DAG). DAGs are statically defined according to the Job Description Language (JDL) specifications [25].

Data analytics engine. The *Data Miner* (DM)[5] is conceived to provide end users with an environment to execute computational analysis of datasets through both service provided algorithms and user defined algorithms [26]. At VRE creation time, the DM is configured with respect to the algorithms to be offered in that context. The DM currently supplies more than 200 ready to use algorithm implementations which include real valued features clustering, functions and climate scenarios simulations, niche modelling, model performance evaluation, time series analysis, and analysis of marine species and geo-referenced data. New algorithms can easily be integrated, in fact the DM comes with a development framework dedicated to this. A scientist willing to integrate a new algorithm should develop it by implementing some basic Java interface defining algorithm's inputs and outputs. In the case of non-Java algorithms, e.g. R scripts, the framework provides facilities

[5]Formerly known as Statistical Manager. In practice, DataMiner results from the re-engineering of the Statistical Manager service.

to integrate them [27]. Integrated algorithms can be shared with coworkers by simply making them publicly available. The DM is designed to operate as a federation of DM instances, all sharing the same capabilities in terms of algorithms. Depending on the characteristics of the algorithm and the data, each DM instance executes the algorithm locally or outsources part of it to the underlying infrastructure including gCube workers (cf. Sec. 2.1). A queue-based messaging system dispatches information about the computation, which includes (*i*) the location of the software containing the algorithm, (*ii*) the subdivision of the input data space, which establishes the portion of the input to assign to each node, (*iii*) the location of the data to be processed, (*iv*) the algorithm parameters. Workers are data and software agnostic, which means that when ready to perform a task, they consume information from the queue and execute the software in a sandbox passing the experimental parameters as input. DM instances and workers share a data space for input and output consisting of a RDBMS, the Storage Manager, and the Home Library (cf. Sec. 2.2).

2.4. Collaborative Services

VREs are easy to use (e.g. requested skills do not exceed the average scientist's ones), have limited adoption costs (e.g. no software to be installed), look like an integrated whole (e.g. the boundaries of the constituents are not perceived), and have an added value with respect to the single constituent's capabilities (e.g. simplify data exchange).

gCube offers its VREs via thin clients, e.g. a plain web browser. All the facilities so far described are made consumable via specific components, i.e. *portlets*, that are web-based user interface constituents conceived to be aggregated, configured and made available by a portal at VREs creation.

To complete its offering and provide its users with added value services, gCube equips its VREs with (*i*) a social networking area, (*ii*) a catalogue, and (*iii*) a user management dashboard.

Social Networking. gCube offers facilities promoting innovative practices that are compliant with Open Science [28]. Among the services there is a *Home Social* resembling a social network timeline where VRE users as well as applications can post messages, information objects, processing results and files. Such posts can be discussed and favourited (or questioned) by VRE members in a very open way. Every member is also provided with a *Workspace*, i.e. a folder-based virtual "file system" allowing complex information objects, including files, datasets, workflows, and maps. Objects residing in the workspace can pre-exist the VRE or be created during the VRE lifetime, all of them are managed in a simple way (e.g. drag & drop), can be downloaded as well as shared in few clicks.

Catalogue. gCube offers a catalogue-like facility VRE users can rely on to have a flexible and powerful publishing environment to announce, or being informed of, the availability of research artefacts. The publishing platform is based on the CKAN open source technology[6]. This technology has been largely customised and extended to meet the needs arising in gCube application scenarios. One of the major extensions is related to the "data model" to be supported by the platform. CKAN natively supports the notion of *dataset*, i.e. the managed item. Each of such items is characterised by (*i*) a basic set of attributes (e.g. author, title, description), (*ii*) an open-ended set of additional attributes captured by any (key, value) pair, (*iii*) a set of resources forming the payload of the item, and (*iv*) the organisation the dataset belongs to. This model has been extended by adding the notion of publishable "item type" that by using a templating approach enables to characterise the publishable item typologies by carefully defining the additional attributes (with controlled vocabularies and allowed values) characterising them. In addition to that, it is possible to define systematic tagging strategies per item type as well as systematic assignment of items to *groups*, i.e. collections of items defined for discovery and management purposes. This make it possible for every community served by a VRE to carefully define the publishing practices as well as the objects to be published (e.g. datasets, software, services) and how they are expected to be described and managed. The catalogue is equipped with a dedicated portlet allowing to navigate and access the available content (taking into account the access policies of the items and organizations) and supporting search (keyword based and faceted) and browse (by tag, organization, group, type) facilities.

User Management. gCube offers a rich array of portlets organised in a dashboard to enact VRE managers to easily manage the users of a VRE. Management facilities offered includes (*i*) to be provided with an ever updated list

[6]CKAN is an open-source system conceived to enact the construction of data hubs and data portals `http://ckan.org/`

of current VRE members with their roles; (*ii*) to easily manage membership, i.e. accept / reject requests for new membership, withdraw membership; (*iii*) to create and manage groups; (*iv*) to assign / revoke roles to members.

3. Related work

Virtual Research Environments, Science Gateways, Virtual Laboratories and other similar terms [7] are used to indicate web-based systems emerged to provide researchers with integrated and user friendly access to data, computing and services of interest for a given investigation that are usually spread across many and diverse data and computing infrastructures. Moreover, they are conceived to enact and promote collaboration among their members for the sake of the investigation.

There are many frameworks that can be used to build such systems. Shahand et al. [9] have identified eleven frameworks explicitly exploited to develop Science Gateway including Apache Airavata, Catania SG Gateway, Globus, HUBzero(+Pegasus), ICAT Job Portal, and WS-PGRADE/gUSE. Such frameworks are quite diverse, e.g. Apache Airavata offers its facilities via an API while the Catania SG Gateway offers its facilities via a GUI and a RESTful API. However, they share certain characteristics that make them operate at a lower level of abstraction with respect to the one of gCube. For data management, these frameworks mainly focus on files while gCube tries to capture an extensive domain offering specific services (cf. Sec. 2.2). Moreover, such specific services are conceived to make it easy to collect data from / interface with existing data providers thus to make their content available to VRE members. For data processing, the frameworks analysed by Shahand et al. focus on executing jobs while the gCube Data analytics engine (cf. Sec. 2.3) complements this key yet basic facility with mechanisms enabling scientists to easily plug their methods into an environment transparently relying on distributed computing solutions. Moreover, every single algorithm once successfully integrated is automatically exposed with a RESTful API (OGC Web Processing Service) thus making it possible to invoke it by workflows. Finally, the mechanism gCube offers for the creation of a VRE is unique (cf. Sec. 2.5). In essence, authorised users can simply create a new VRE via a wizard driving them to produce a characterisation of the needed environment in terms of existing resources. The software (including GUI constituents) and the data needed to satisfy the VRE specification are automatically deployed, no sysadmin intervention is needed.

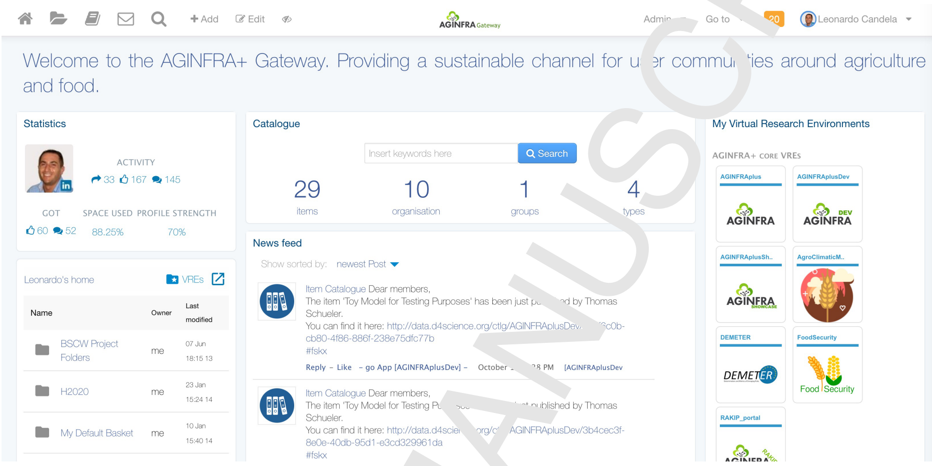


Figure 3: AGINFRA+ Gateway Home

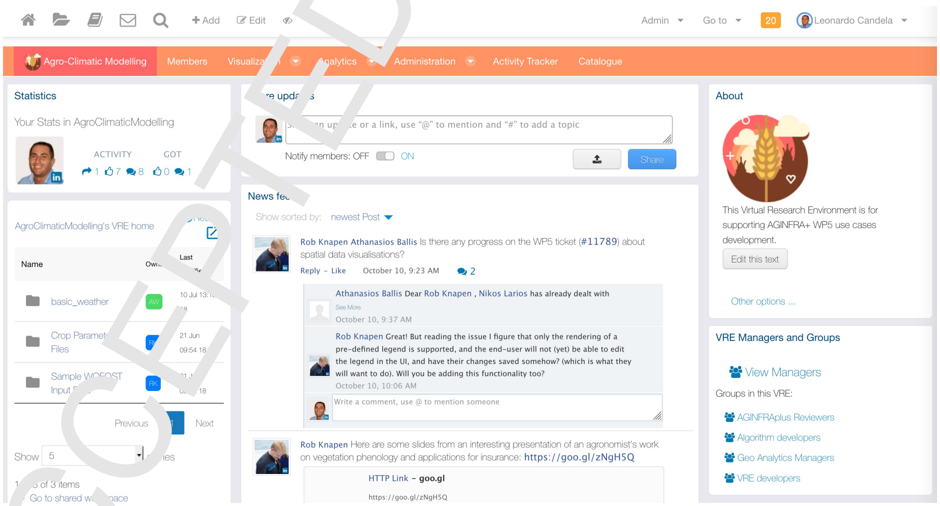


Figure 4: A typical Virtual Research Environment application

4. Conclusion

This paper provided a comprehensive description of the design principles characterising the gCube software system and the facilities this system offers to enact the operation of an IT infrastructure enabling the development of *Virtual Research Environments*, i.e. ready to use web-based working environments specifically conceived to provide their designated community with the facilities (services, data, capacity) they need. Virtual research environments are expected to be used primarily via a plain web browser. A screenshot of a typical home of the GUI is in Fig. 3 where the user is provided with the list of VREs he/she is member of (right column), some statistics on his/her activity and a direct access to the workspace (left column), access to the catalogue and the messages and discussions occurring (central column). Fig. 3 showcases a typical home page of a VRE with its own specific menu for accessing the facilities it offers. In addition to web GUIs, there are RESTful APIs for accessing and using VRE services in a programmatic way.[7].

The development of a new VRE by gCube is a task usually requiring few ours including the use of the wizard to define the VRE, the approval of the specification and the deploy of the components including the portlets. Once the VRE is in place, the VRE managers can customize the GUI (if needed) by reshuffling the portlets across menu and pages, add some content (e.g. in the workspace) and finally start inviting members. This is the ideal scenario based on the assumption that everything needed to create the VRE has been already developed. In case there are pieces of technology to be developed the time and effort needed to develop such technology depends on several factors including their complexity. In some cases, it is a matter of simply exploiting some gCube services. For instance, if a VRE Manager / member is willing to make available to the rest of VRE members his/her own analytics method (already implemented), he/she should just import such a method by relying on the data analytics importing mechanism (cf. Sec. 2.3). Another example of extension is about the enlargement of the available data sources. A number of mediators have been developed for accessing biodiversity information systems, the one of interest are simply selected at VRE specification time yet new ones can be developed to deal with new sources. Entire web sites can be integrated in a VRE by relying on a specific portlet taking care of embedding the content of the third party web site and

[7]More information are available at `https://dev.d4science.org/`

invoking the third party website with a security token enabling the web site to get information on the user invoking it. All in all, the patterns governing VRE development are many and diverse. Developing a new component suitable for being integrated in a VRE also depends on the degree of integration expected, they range from a legacy component simply integrated from the GUI perspective with aspects of authentication and authorization up to fully fledged new gCube components designed to interface with the rest of gCube components.

The experiences made while exploiting gCube to operate the D4Science.org infrastructure somehow demonstrate that the principles governing the VREs delivery and the system openness are key in the modern science settings [29]. The currently supported VREs are available via dedicated portals, some of these VREs are openly available for exploitation and test.[8] Several use cases have been developed by relying on gCube-based VREs, e.g. to estimated the spread of the puffer fish Lagocephalus sceleratus in the Mediterranean Sea due to suitable environmental conditions in this area and favoured by climate change [30]; to develop a bayesian hierarchical approach for the estimation of length?weight relationships in fishes [31]; to develop a global record of fish stocks and fisheries [32]; to develop a workflow where by sharing photos of an object or an environment it is possible to produce a virtual reality scene as a navigable 3D reconstruction that can be shared with other people [33]. Overall, D4Science is currently serving more that thousands of users (more than 7000 in September '18). In the period January-September 2018 the users served by this infrastructure and its VREs performed: a total of 50,127 sessions, with an average of circa 5569 sessions per month; a total of 4,288 social interactions, with an average of circa 476 interactions per month; a total of 150 millions of analytics tasks, with an average of circa 16 millions tasks per month.

Acknowledgments. This work has received funding from the European Union's Horizon 2020 research and innovation programme under the AGINFRA PLUS project (Grant agreement No. 731001), the BlueBRIDGE project (Grant agreement No. 675680), the ENVRI PLUS project (Grant agreement No. 654182) and the EOSCpilot project (Grant No. 739563). The gCube tech-

[8] The D4Science Gateway `http://services.d4science.org/` offers an up to date list of gCube-based Virtual Research Environments.

ACCEPTED MANUSCRIPT

nology results from several efforts and contributions[9].

References

[1] T. Hey, S. Tansley, K. Tolle, The Fourth Paradigm: Data-Intensive Scientific Discovery, Microsoft Research, 2009.

[2] S. Bartling, S. Friesike, Towards another scientific revolution, in: Opening Science, Springer International Publishing, 2014, pp. 3–15. doi:10.1007/978-3-319-00026-8_1.

[3] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, M. Zaharia, A view of cloud computing, Communications of the ACM 53 (4) (2010) 50–58. doi:10.1145/1721654.1721672.
URL http://doi.acm.org/10.1145/1721654.1721672

[4] B. Allen, J. Bresnahan, L. Childers, I. Foster, G. Kandaswamy, R. Kettimuthu, J. Kordas, M. Link, S. Martin, K. Pickett, S. Tuecke, Software as a service for data scientists, Communications of the ACM 55 (2) (2012) 81–88. doi:10.1145/2076450.2076468.

[5] I. Foster, Service-oriented science, Science 308 (5723) (2005) 814–817. doi:10.1126/science.1110411.

[6] S. Gesing, N. Wilkins-Diehr, Science gateway workshops 2014 special issue conference publications, Concurrency and Computation: Practice and Experience 27 (16) (2015) 4247–4251. doi:10.1002/cpe.3615.

[7] L. Candela, D. Castelli, P. Pagano, Virtual research environments: an overview and a research agenda, Data Science Journal 12 (2013) GRDI75–GRDI81. doi:10.2481/dsj.GRDI-013.

[8] K. A. Lawrence, N. Wilkins-Diehr, J. A. Wernert, M. Pierce, M. Zentner, S. Marru, Who cares about science gateways? a large-scale survey of community use and needs, in: 9th Gateway Computing Environments Workshop, 2014, pp. 1–4. doi:10.1109/GCE.2014.11.

[9] https://www.gcube-system.org/credits

[9] S. Shahand, A. H. C. van Kampen, S. D. Olabarriaga, Science gateway canvas: A business reference model for science gateways, in: SCREAM '15 Proceedings of the 1st Workshop on The Science of Cyberinfrastructure: Research, Experience, Applications and Models, 2015. doi:10.1145/2753524.2753527.

[10] F. Simeoni, L. Candela, D. Lievens, P. Pagano, M. Simi, Functional adaptivity for Digital Library Services in e-Infrastructures: the gCube Approach, in: 13th European Conference on Research and Advanced Technology for Digital Libraries, ECDL 2009, 2009.

[11] C. Aiftimiei, A. Aimar, A. Ceccanti, M. Cecchi, A. Di Meglio, F. Estrella, P. Fuhrmam, E. Giorgio, B. Konya, L. Field, J. Nilsen, M. Riedel, J. White, Towards next generations of software for distributed infrastructures: The european middleware initiative, in: E-Science (e-Science), 2012 IEEE 8th International Conference on, 2012. doi:10.1109/eScience.2012.6404415.

[12] R. Ananthakrishnan, K. Chard, I. Foster, S. Tuecke, Globus platform-as-a-service for collaborative science applications, Concurrency and Computation: Practice and Experience n/a (2014) n/a. doi:10.1002/cpe.3262.

[13] A. Edmonds, T. Metsch, A. Papaspyrou, A. Richardson, Toward an open cloud standard, IEEE Internet Computing 16 (4) (2012) 15–25. doi:10.1109/MIC.2012.65.

[14] C. Lagoze, H. Van de Sompel, The open archives initiative: building a low-barrier interoperability framework, in: Proceedings of the first ACM/IEEE-CS Joint Conference on Digital Libraries, ACM Press, 2001, pp. 54–62.

[15] SDMX Development Core Team, SDMX: Statistical Data and Metadata Exchange.

[16] M. Assante, L. Candela, D. Castelli, L. Frosini, L. Lelii, P. Manghi, A. Manzi, P. Pagano, M. Simi, An Extensible Virtual Digital Libraries Generator, in: 12th European Conference on Research and Advanced Technology for Digital Libraries (ECDL), 2008, pp. 122–134.

[17] M. Assante, P. Pagano, L. Candela, F. De Faveri, 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), Springer, 2011, pp. 101–109.

[18] Y. L. Simmhan, B. Plale, D. Gannon, A survey of data provenance in e-science, SIGMOD Rec. 34 (3) (2005) 31–36. doi:10.1145/1084805.1084812.

[19] R. Cattell, Scalable SQL and NoSQL data stores, SIGMOD Rec. 39 (4) (2011) 12–27. doi:10.1145/1978915.1978919.

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

[21] F. Simeoni, L. Candela, G. Kakaletris, M. Sibeko, P. Pagano, G. Papanikos, P. Polydoras, Y. E. Ioannidis, D. Aarvaag, F. Crestani, A Grid-Based Infrastructure for Distributed Retrieval, in: 11th European Conference on Research and Advanced Technology for Digital Libraries, 2007, pp. 161–173.

[22] M. Selamat, M. S. Othman, N. H. M. Shamsuddin, N. I. M. Zukepli, A. F. Hassan, A review on open source architecture in geographical information systems, in: Computer Information Science (ICCIS), 2012 International Conference on, Vol. 2, 2012, pp. 962–966. doi:10.1109/ICCISci.2012.6297165.

[23] L. Candela, D. Castelli, G. Coro, L. Lelii, F. Mangiacrapa, V. Marioli, P. Pagano, An infrastructure-oriented approach for supporting biodiversity research, Ecological Informatics 26 (2014) 162–172. doi:10.1016/j.ecoinf.2014.07.006.

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

[25] E. Laure, S. M. Fisher, A. Frohner, C. Grandi, P. Kunszt, A. Krenek, O. Mulmo, F. Pacini, F. Prelz, J. White, M. Barroso, P. Buncic, F. Hemmer, A. Di Meglio, A. Edlund, Programming the grid with glite, Computational Methods in Science and Technology 12 (1) (2006) 33–45.

[26] G. Coro, L. Candela, P. Pagano, A. Italiano, L. Liccardo, Parallelizing the execution of native data mining algorithms for computational biology, Concurrency and Computation: Practice and Experience n/a (n/a) (2014) n/a. doi:10.1002/cpe.3435.

[27] G. Coro, G. Panichi, P. Pagano, A web application to publish r scripts as-a-service on a cloud computing platform, Bollettino di Geofisica Teorica e Applicata: an International Journal of Earth Sciences 57 (2016) 51–53.

[28] M. Assante, L. Candela, D. Castelli, P. Manghi, P. Pagano, Science 2.0 repositories: Time for a change in scholarly communication, D-Lib Magazine 21 (1/2). doi:10.1045/january2015-assante.

[29] L. Candela, D. Castelli, A. Manzi, P. Pagano, Realising Virtual Research Environments by Hybrid Data Infrastructures: the D4Science Experience, in: International Symposium on Grids and Clouds (ISGC) 2014 23-28 March 2014, Academia Sinica, Taipei, Taiwan, PoS(ISGC2014)022, Proceedings of Science, 2014.

[30] G. Coro, L. Gonzalez Vilas, C. Magliozzi, A. Ellenbroek, P. Scarponi, P. Pagano, Forecasting the ongoing invasion of lagocephalus sceleratus in the mediterranean sea, Ecological Modelling 371 (2018) 37 – 49. doi:10.1016/j.ecolmodel.2018.01.007.

[31] R. Froese, J. T. Thorson, R. B. J. Reyes, A bayesian approach for estimating length-weight relationships in fishes, Journal of Applied Ichthiology 30 (1) (2014) 78–85. doi:10.1111/jai.12299.

[32] Y. Tzitzikas, Y. Marketakis, N. Minadakis, M. Mountantonakis, L. Candela, F. Mangiacrapa, P. Pagano, C. Perciante, D. Castelli, M. Taconet, A. Gentile, G. Gorelli, Towards a global record of stocks and fisheries, in: M. Salampasis, A. Theodoridis, T. Bournaris (Eds.), Proceedings of the 8th International Conference on Information and Communication Technologies in Agriculture, Food and Environment (HAICTA 2017),

Chania, Crete Island, Greece, September 21-24, 2017, Vol. 2030, 2017, pp. 328–340.

[33] G. Coro, M. Palma, A. Ellenbroek, G. Panichi, T. Nair, P. Pagano, Reconstructing 3d virtual environments within a collaborative e-infrastructure, Concurrency and Computation: Practice and Experience 0 (0) e5028. doi:10.1002/cpe.5028.

Massimiliano Assante is a researcher of the "Istituto di Scienza e Tecnologie dell'Informazione A. Faedo" (ISTI), an institute of the Italian National Research Council (CNR). He holds a Ph.D. on Information Engineering and a master degree (M.Sc.) on Information Technologies, both received from the University of Pisa. His research interests include e-Infrastructures, Scientific Repositories, Data Publishing, Virtual Research Environments and NoSQL Data Stores. Massimiliano joined ISTI in 2007, he worked for several EU Projects such as iMarine, EUBrazilOpenBio, D4Science II, D4Science and DILIGENT. Within these projects, he progressively covered different positions, ranging from software engineer (web services and front-end web applications) to analyst, system designer, system integrator, researcher. Currently, he is working in several EU projects (BlueBRIDGE, SoBigData, PARTHENOS, AGINFRA+) and leads the Work Package responsible for Data Access, Discovery, Storage, Analysis and Publishing for the (EU H2020) BlueBRIDGE Project.

Leonardo Candela is Researcher at the Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, Italy. He has relevant expertise in the area of Virtual Research Environments development. He was involved in various EU-funded projects including CYCLADES, Open Archives Forum, DILIGENT, DRIVER, DELOS, D4Science, D4Science-II, DL.org, EUBrazilOpenBio, iMarine, ENVRI. He was a member of the DELOS Reference Model Technical Committee and of the OAI-ORE Liaison Group. Currently, he is the Project Manager of the BlueBRIDGE Project and CNR lead person in the ENVRIPlus one. His research interests include Data Infrastructures, Virtual Research Environments, Data Publication, Open Science, Digital Library [Management] Systems and Architectures, Digital Libraries Models, Distributed Information Retrieval, and Grid and Cloud Computing.

Donatella Castelli is Senior Researcher at the Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, Italy. He leads the InfraScience research group. Her research interests include digital libraries and data infrastructure content modeling, interoperability and architectures. She is author of several research papers in these fields. Under her supervision, the InfraScience team coordinated and participated in several EU and nationally funded projects on Digital Libraries and Research Data Infrastructures. In particular, she has been the co- ordinator of the EU projects that have developed the D4Science infrastructure and technical coordinator of those that have developed the OpenAIRE one. She has also been the coordinator of the FP7 DL.org project dedicated to the interoperability of digital library systems and the CNR leading person for FP7 GRDI2020 focussed topics related to the realization of global research data infrastructures. Currently she is the Technological Coordinator of the H2020 OpenAIRE2020 project and the Scientific coordinator of the H2020 VRE BlueBRIDGE one. She is also a member of the RDA Europe Expert Group that promotes research and cross-infrastructure coordination at global level.

Roberto Cirillo Roberto Cirillo is Researcher at the Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, Italy. His scientific and professional activity involves the research and development on Data Infrastructures. His research interests include e-Infrastructures, Cloud-based technologies, Virtual Research Environments and NoSQL Data Stores. He is currently member of the BlueBRIDGE EU Project. He was involved in various EU-funded projects including iMARINE, EUBrazil-OpenBio, ENVRI, EGI-ENGAGE. In the past, he has been working on Language Technologies.

Gianpaolo Coro is a Physicist with a Ph.D. in Computer Science. His research focuses on Artificial Intelligence, Data Mining and e-Infrastructures. He has been working for more than ten years on Machine Learning and Signal Processing with applications to Computational Biology, Brain Computer Interfaces, Language Technologies and Cognitive Sciences. The aim of his research is the study and experimentation of models and methodologies to process biological data and to apply the results to fields in Ecological Modelling, Vessel Monitoring Systems and Ecological Niche Modelling with an approach oriented to Science 2.0. His approach relies on distributed e-Infrastructures and uses parallel and distributed computing via Grid- and Cloud-based technologies.

Luca Frosini is researcher at the Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, Italy. He has relevant expertise in the area of Virtual Research Environments development. He was involved in various EU-funded projects including DILIGENT, D4Science, EAGLE, PARTHENOS, SoBigData and BlueBRIDGE. Currently, he is Taks Leader of Federated Resources Management in BlueBRIDGE Project. His research interests include Data Infrastructures, Virtual Research Environments, Information Systems, Accounting Systems, and Grid and Cloud Computing.

Lucio Lelii is Researcher at the Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, Italy. His scientific and professional activity involves the Research and Development on Data Infrastructures. He is currently member of the BlueBRIDGE EU Project.

Francesco Mangiacrapa is a computer scientist and researcher at the Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, Italy. He has background on geospatial data, technologies, models and standard OGC (like WMS, WFS and so on) for spatial data representation and exchange. His scientific and professional activity includes study and research on Virtual Research Environments and Data Infrastructure, Data Publication, GeoSpatial Data and Open Science. Moreover, his work involve design and development of (Web-)GUI based on several framework (like GWT, Material, Bootstrap and so on) to support his research activity and able to improve community collaboration and exchange of scientific data. Currently, he is working in several EU projects (BlueBRIDGE, SoBigData, PARTHENOS, AGINFRA+) and is responsible for: Data Access and Exchange (Workspace Area), Data Catalogue and Publishing (Catalogue Area) of BlueBRIDGE Project.

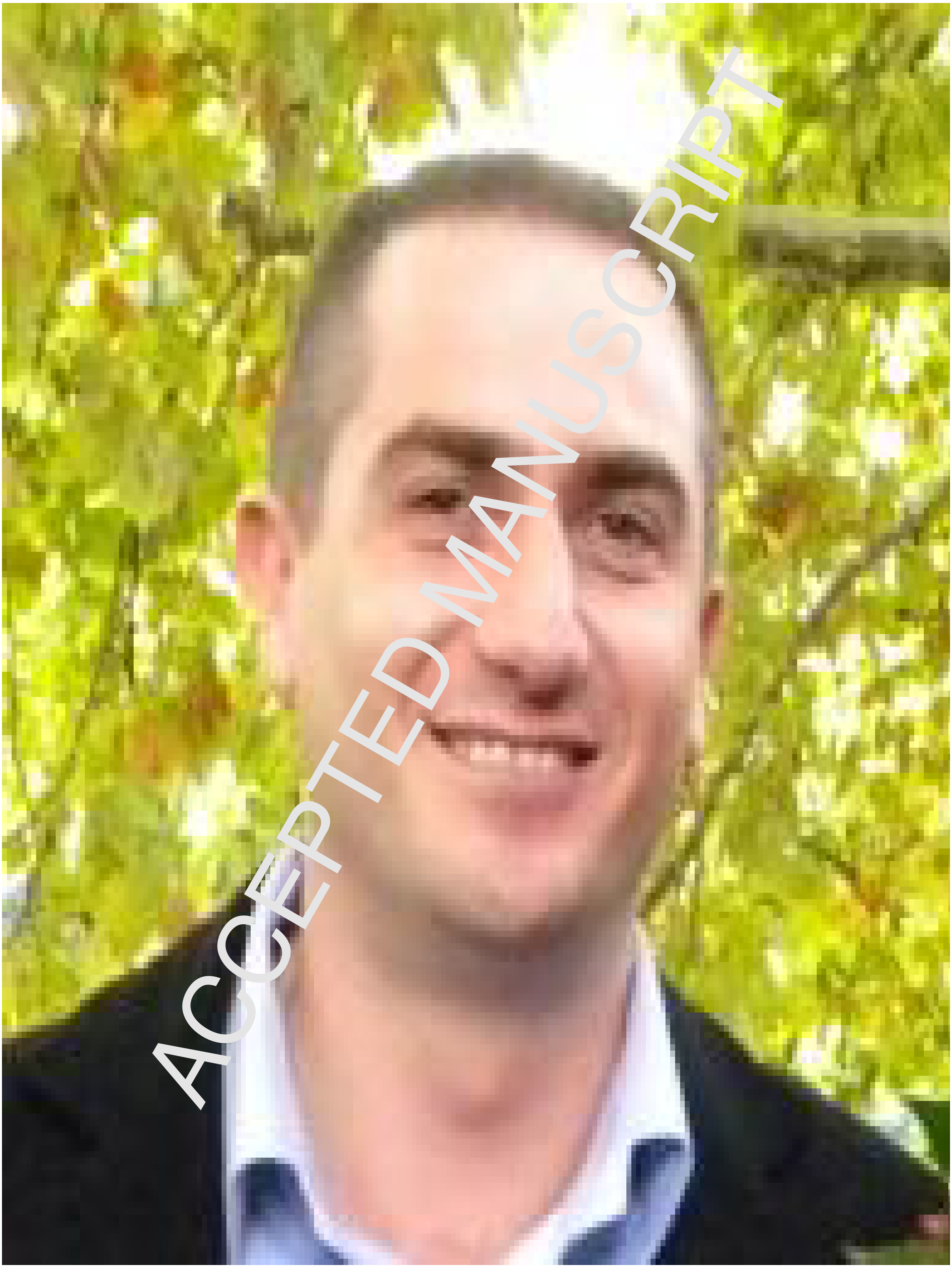
Valentina Marioli is research fellow at the Istituto di Scienza e Tecnologie dell'informazione, Consiglio Nazionale delle Ricerche, Pisa, Italy. She earned her Master's degree in Computer Science from Pisa University, Italy. She worked at her thesis at the Swedish Institute of Computer Science in Stockholm, Sweden, where she focused on data consistency in Wireless Sensor Networks. The design, implementation and evaluation of a broadcast protocol using Virtual Synchrony paradigm were the main points of her work. Currently, her research focuses on data integration in the biodiversity and environmental domains, and storage and data management. Her specialist areas include configuration and management of content repositories and the exchange of biodiversity information using Standards and Protocols.

Pasquale Pagano is Senior Researcher at CNR-ISTI. He has a strong background and experience on models, methodologies and techniques for the design and development of distributed virtual research environments (VREs) which require the handling of heterogeneous computational and storage resources, provided by Grid and Cloud based e-Infrastructures, and management of heterogeneous data sources. He participated in the design of the most relevant distributed systems and e- Infrastructure enabling middleware developed by ISTI - CNR. He is currently the Technical Director of the D4Science Data Infrastructure, Technical Director of H2020 BlueBRIDGE project and CNR lead person for the EGI-ENGAGE one. In the past, he has been involved in the iMarine, EUBrazilOpenBio, ENVRI, Venus-C, GRDI2020, D4Science-II, D4Science, Diligent, DRIVER, DRIVER II, BELIEF, BELIEF II, Scholnet, Cyclades, and ARCA European projects.

Giancarlo Panichi is a member of the Technical Staff at the "Istituto di Scienza e Tecnologie dell'Informazione A. Faedo" (ISTI), an institute of the Italian National Research Council (CNR). His skills concern e-Infrastructures, Web Processing Services, Virtual Research Environments, Data Management, Data Analytics, Web Services, Web Applications and Mobile Applications. Giancarlo joined ISTI in 2013, he worked for several EU Projects such as iMarine, EUBrazilOpenBio and ENVRI, currently, he is working in BlueBRIDGE Project (EU H2020).

Costantino Perciante is a first-year PhD student at the University of Pisa. He is also a Research Fellow at the Institute of Information Science and Technologies (ISTI) - CNR in Pisa, where he is currently involved in several activities within the D4Science Research Infrastructure. Costantino holds a Bachelor's degree and a Master's degree both in Computer Engineering from the University of Pisa.

Fabio Sinibaldi is a Researcher at CNR ISTI. He holds a degree in computer science engineering with specialization in business management technologies received from the University of Pisa. In his research studies he worked on the design and development of distributed environments' services aimed to manage scientific data, with special attention to Ecological Niche Modelling approaches. These studies involved exploitation of federated Grid and Cloud e-Infrastructures along with Digital Libraries oriented workflow analysis and design, leading to the development of D4Science's Spatial Data Infrastructure. He currently works as Spatial Data Infrastructure designer for D4Science Data Infrastructure under H2020 BlueBRIDGE project and as technology integration manager for EGI-ENGAGE one. In the past he has been involved in the iMarine, EAGLE, EUBrazilOpenBio, ENVRI, Venus-C, D4Science-II, D4Science projects.




ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT


ACCEPTED MANUSCRIPT

Highlights

- A software system enacting the delivery of Virtual Research Environments as a Service;
- Rich array of data management services;
- Data processing and analytics solutions;