Future Generation Computer Systems 101 (2019) 555–563

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Enacting open science by D4Science

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], P. Pagano [a], G. Panichi [a], 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

HIGHLIGHTS

- A platform supporting the implementation of open science practices.
- A shared workspace to store and organise any version of a research artefact.
- A social networking area to have discussions and be informed on happenings.
- A data analytics platform to share and execute processing methods.
- A catalogue-based publishing platform to communicate the existence of an artefact.

ARTICLE INFO

Article history:
Received 19 June 2018
Received in revised form 23 May 2019
Accepted 26 May 2019
Available online 30 May 2019

Keywords:
Open science
gCube
D4Science
Social networking
Analytics
Publishing

ABSTRACT

The open science movement is promising to revolutionise the way science is conducted with the goal to make it more fair, solid and democratic. This revolution is destined to remain just a wish if it is not supported by changes in culture and practices as well as in enabling technologies. This paper describes the D4Science offerings to enact open science-friendly Virtual Research Environments. In particular, the paper describes how complete solutions suitable for realising open science practices can be achieved by integrating a social networking collaborative environment with a shared workspace, an open data analytics platform, and a catalogue enabling to effectively find, access and reuse every research artefact.

1. Introduction

Open science is promising to revolutionise the entire science enterprise by envisioning and proposing new practices aiming at "improving" science [1–3]. The promised benefits include (*i*) better interpretation, understanding and reproducibility of research activities and results; (*ii*) enhanced transparency in science lifecycle and improvement in "scientific fraud" detection; (*iii*) reduction of the overall cost of research, by promoting re-use of results; (*iv*) introduction of comprehensive and fair scientific reward criteria capturing all facets and contributions in research life-cycle; and (*v*) better identification and assessment of scientific results within the "tsunami of scientific literature" witnessed by researchers.

These benefits are destined to remain a wish if the research community as a whole (funding agencies, research performing organisations, publishers, research infrastructures, researchers, as well as citizens) does not put in place efforts and initiatives aiming at making open science the norm. The good news is that the movement is gaining momentum and consensus. In fact, funding agencies are developing policies supporting open science implementation as well as are supporting the development of infrastructures and services. Research infrastructures start developing and using services and facilities aiming at supporting the implementation of open science practices. Researchers try to overcome the limitations of scholarly communication practices by relying on services and technologies to "publish" non-traditional research artefacts (datasets, workflows, software). The bad news is that the implementation of open science practices is confronting with a number of barriers including: (*i*) cultural factors, e.g. the

* Corresponding author.
E-mail 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), pasquale.pagano@isti.cnr.it (P. Pagano), giancarlo.panichi@isti.cnr.it (G. Panichi), fabio.sinibaldi@isti.cnr.it (F. Sinibaldi).

https://doi.org/10.1016/j.future.2019.05.063
0167-739X/

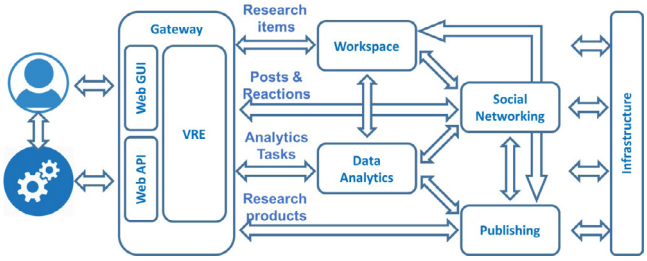


Fig. 1. D4Science open science services framework.

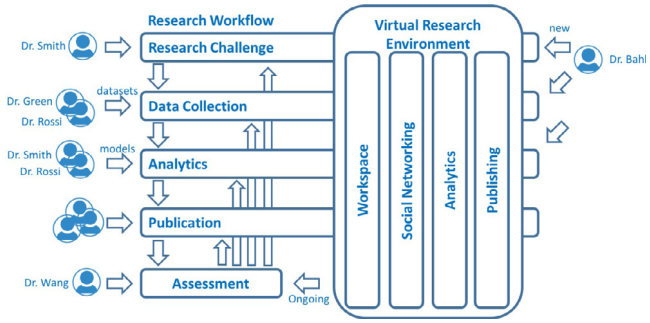


Fig. 2. D4Science-based open science workflow.

fear to lose the control and exploitation of datasets; (*ii*) cost-based factors, e.g. the extra-effort needed to make a research-artefact exploitable by a user other than the initial owner; and (*iii*) disincentive factors, e.g. the effort spent in "publishing" research artefacts going beyond traditional papers is receiving little or no value in researchers' assessment and career success.

This paper describes the solution proposed by the D4Science Infrastructure to overcome some of the above mentioned barriers — in particular those tied to cost-based arguments. D4Science provides researchers and practitioners with a working environment where open science practices are transparently promoted. This infrastructure is built and operated by relying on gCube technology [4], a software system specifically conceived to enable the construction and development of *Virtual Research Environments* (VREs) [5], i.e. web-based working environments tailored to support the needs of their designated communities, each working on a research question. Beside providing users with the domain-specific facilities, i.e. datasets and services suitable for the specific research question, each VRE is equipped with basic services supporting collaboration and cooperation among its users, namely: (*i*) a shared workspace to store and organise any version of a research artefact; (*ii*) a social networking area to have discussions on any topic (including working version and released artefacts) and be informed on happenings; (*iii*) a data analytics platform to execute processing tasks either natively provided by VRE users or borrowed from other VREs to be applied to VRE users' cases and datasets; and (*iv*) a catalogue-based publishing platform to make the existence of a certain artefact public and disseminated. These facilities are at the fingerprint of VRE users. They continuously and transparently capture research activities, authors and contributors, as well as every by-product resulting from every phase of a typical research lifecycle, thus reducing the issues related with open science and its communication [6,7].

The remainder of this paper is organised as follows. Section 2 overviews related works, technologies and initiatives supporting open science. Section 3 describes the D4Science solution by reporting the components' architecture and their open access and FAIRness features. Section 4 discusses the proposed solution by giving examples of how it has been used in concrete scenarios. Finally, Section 5 concludes the paper by describing some future works.

2. Related works

There are plenty of tools and approaches supporting open science [8]. These include (*i*) repositories maintaining different versions of datasets and software to promote their citation and reuse, e.g. Dataverse, Dryad, GitHub, Zenodo, figshare; (*ii*) tools aiming at promoting and enacting the production of new forms of publications to make the release of research results more effective and comprehensive, e.g. interactive notebooks and enhanced publications [9,10]; (*iii*) tools aiming at making the research assessment process more open, transparent, holistic and participative, e.g. open peer review, post-publication review, annotation and commenting tools like Hypothesis [11], social networks for scientists like ResearchGate [12]. One of the major barriers preventing the systematic exploitation and uptake of these tools by scientific communities and application contexts is related to their "fragmentation". Scientists have to jump across several platforms to get a complete picture of a research activity and its current and future results. Whenever possible, the "pieces" resulting from a research activity are linking each other, either by explicit links or by derived links, e.g. [13]. However, this link-based mechanism is quite fragile and costly to keep healthy and up to date. To overcome this issue, research packaging formats are now being designed and developed [14].

Scientific workflow technologies [15] are adopted by an increasing number of communities to automate scientific methods and procedures. Environments enabling publishing and sharing workflows exist, e.g. [16], yet the guarantees that the method captured by the workflow seamlessly works in settings other than the originator ones are limited. Moreover, the act of publishing workflows is not systematic across communities and contexts.

HUBZero [17] is a platform for building websites (also known as *science gateways* [18]) aiming at providing analytical tools and facilities to publish data, share resources, collaborate and build communities in a single web-based ecosystem. A key component is a content management system enacting to create and share the specific "hub" content ranging from blog entries to datasets and computational tools. This platform is focusing more on the facilities for supporting the development of "hubs" than on promoting collaboration and sharing across hubs contrarily to gCube/D4Science that promotes both, i.e. specific VREs development and cross VRE collaboration by offering services across VREs [4]. In fact, specific solutions are built by it to promote open science practices like research data sharing, e.g. [19].

The Open Science Framework (OSF)[1] is a web-based service enabling users to keep files, data, and protocols pertaining to any user defined project in a single, shared place. It provides for credits, citation and versioning as well as for carefully deciding what is going to be shared and with whom. This platform shares commonalities with the gCube based set of facilities described in this paper. Every OSF project is like a D4Science VRE, yet every VRE provides users with an aggregated working environment offering the entire set of facilities and services users need to perform their research. Thus, D4Science VREs go beyond the pure sharing of material and aim at offering complete working environments.

3. D4Science open science services framework for VRE-based scientific workflows

Fig. 1 depicts the main components and facilities offered by D4Science to support collaborative activities and enable open

[1] Open Science Framework webpage https://osf.io.

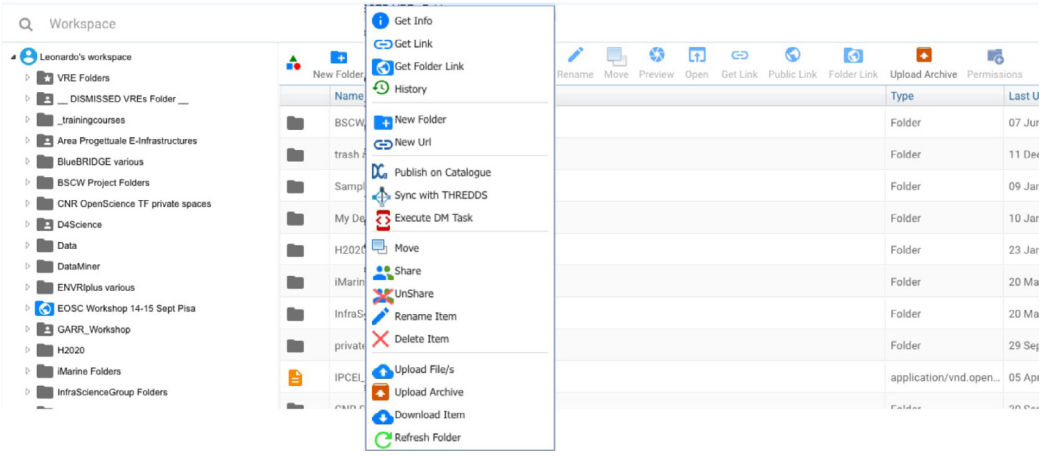


Fig. 3. D4Science workspace screenshot.

science practices, namely a shared workspace (cf. Section 3.1), a social networking area (cf. Section 3.2), a data analytics platform (cf. Section 3.3), and a publishing platform (cf. Section 3.4). These components are all correlated each other and realise a "system" where (*i*) research artefacts seamlessly flow across the various components to be "managed" according to the components' purpose, e.g. being openly discussed by social networking practices, (*ii*) research artefacts are continuously enriched and enhanced with metadata capturing their entire lifecycle, their versions, and the detailed list of authors and tasks performed leading to the current development status and shapes.

All these components (*i*) are conceived to operate in a well defined application context corresponding to the Virtual Research Environment they are serving, i.e. the VRE members are the primary researchers and practitioners expected to have fully-fledged access to the artefact shared by a VRE; (*ii*) are conceived to open up research artefacts and contribute to their FAIRness [20], independently of their maturity level, beyond VRE boundaries (no lock-in) yet according to artefact owner's policies, i.e. it is up to artefact owners to decide when a certain item resulting from a research activity is "ready" to be released and how (only metadata, role-based access to payloads, usage license); (*iii*) are operated by relying on an infrastructure that guarantees a known quality of service thus promoting community uptake, i.e. scientists might be reluctant to migrate their working environment towards innovative and "cloud"-based ones [21], the proposed facilities should be as much as possible easy to use, protect consolidated practices, and guarantee that scientists continue to get on with their daily job.

A prototypical and simplified scientific workflow enacted by these components is (cf. Fig. 2): (*i*) Dr. Smith is willing to investigate the impact of a certain alien species in the Mediterranean sea and announces this willingness by a post (social networking); (*ii*) Dr. Green and Dr. Rossi start collaborating with Dr. Smith by organising and populating a shared folder with suitable material, e.g. datasets, notes, papers (workspace); (*iii*) Dr. Smith and Dr. Rossi propose two diverse models aiming at capturing the effects of the selected species on Mediterranean sea ecosystem. They implement such models and make these implemented versions available for others (data analytics); (*iv*) the availability of these early-results suggest Dr. Bahl to start a study on another species he developed a model for in the past, and leads Dr. Bahl to create another workspace folder with specific material and to produce another version of Dr. Rossi's model; (*vi*) Dr. Smith, Dr. Green and Dr. Rossi execute a large set of concurrent experiments, make available every dataset resulting from them (workspace, publishing), and announce their findings by also preparing a paper. Meanwhile, Dr. Wang starts re-using the model(s) produced by Dr. Smith et al. as well as Dr. Bahl's one to analyse certain datasets she owns, spot a potential implementation issue affecting all of the models, produces and publishes corrected versions, and "annotate" the initial models with her findings; (*vii*) being alerted by Dr. Wang annotation, Dr. Smith et al. decide to re-execute their experiments on other datasets by using both their version of the model as well as Dr. Wang's one. As a result, they realise that Dr. Wang's model better suits with their initial hypothesis. All of this happens well before their paper is published. This representative workflow can only be implemented easily and effectively by relying on a suite of facilities like those offered by D4Science. In fact, in D4Science VRES the "place" where research activity is conducted and the "place" where the activity is published and immediately communicated are the same. In other settings where there is a decoupling of the "place" where research is performed (the scientists workbench) from the place where research is communicated, e.g. papers containing links to supporting material, the implementation of this scenario is more challenging and expensive, if at all feasible.

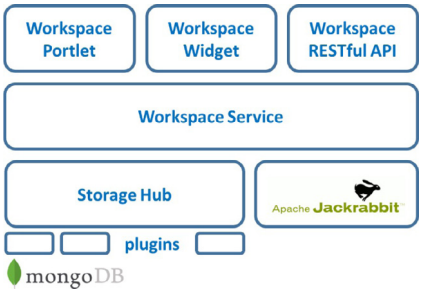


Fig. 4. D4Science workspace platform architecture.

In the following sub-sections, a per service description of the D4Science offerings is given including a paragraph highlighting how the service contributes and promotes open science practices and FAIRness of research artefacts.

3.1. The workspace platform

Fig. 3 depicts the user interface of the workspace facility, i.e. the area VRE users rely on to organise their material and have access to the material shared with others. It resembles a typical file system with files organised in folders, yet it supports an open-ended set of items that are (*i*) equipped with rich and

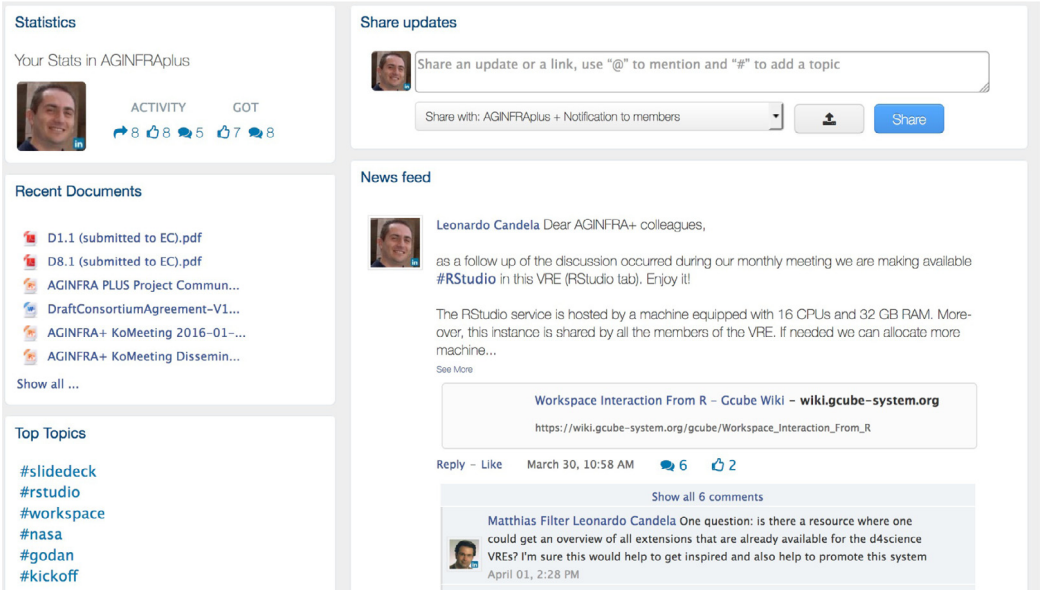


Fig. 5. D4Science social networking screenshot.

extensible metadata and (*ii*) actually stored by an array of storage solutions [4].

Fig. 4 depicts the software architecture characterising the workspace facility. This facility relies on Apache Jackrabbit for storing and managing workspace items – actually their metadata – by means of specific node types and attributes as "key, value" pairs. Items payload is stored by relying on a hybrid storage solution [4] that, by means of ad-hoc plugins, exploits various storage solutions suitable for diverse typologies of content, e.g. MongoDB for binary files, GeoServer and THREDDS Data Servers for geospatial data, RDB for tabular data. In addition to the portlet previously discussed, the workspace facility is offered by (*i*) a *widget* suitable for integrating the workspace facility in other applications (e.g. it is exploited by the Analytics Platform Portlet to give seamless access to workspace items), and (*ii*) a *RESTful API* suitable for any web-based programmatic access.

Service features enabling open science & FAIRness. (*i*) every workspace item is equipped with an actionable unique identifier (namely an URL with some metadata) that can be used for citation and access purposes; (*ii*) every workspace item is versioned and a new version is automatically produced whenever the item is explicitly changed by a user or any application/service of the VRE on behalf of an authorised user; (*iii*) every item, be it a single item or a folder, is equipped with rich and extensible metadata ("key, value" pairs) that capture descriptive features as well as lineage features; (*iv*) three typologies of folders are supported: *private*, content is available only for its owner; *shared*, content is available for selected users (decided by the owner); and, *VRE folder*, content is available to VRE members[2]; (*v*) the workspace is tightly integrated with both the social networking and the catalogue for easing the dissemination of its artefacts (either single items or groups of items).

2 There are several procedures for becoming a VRE Member: (*i*) if the VRE is "open" the user can simply apply for it (after having discovered it) and the request is automatically approved; (*ii*) if the VRE is "restricted", a user can apply for it, the request should be approved by one of the VRE managers; (*iii*) if the VRE is "private", a user is not allowed to apply for it, he/she should be explicitly invited; (*iv*) independently on the membership policy, authorised VRE members can invite other members.

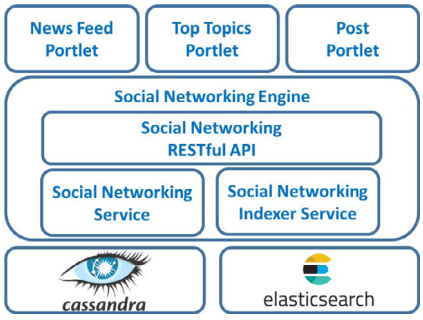


Fig. 6. D4Science social networking collaborative platform architecture.

3.2. The social networking collaborative platform

Fig. 5 depicts the user interface of the social networking area, i.e. the area VRE users rely on to communicate with their VRE co-workers and be informed on others achievements, discussions and opinions. It resembles a social networking environment with posts, tags, mentions, comments and reactions, yet its integration with the rest makes it a powerful and flexible communication channel for researchers.

Fig. 6 depicts the software architecture characterising the social networking collaborative platform. This facility relies on the *Social Networking Engine*, a Cassandra database [22] for storing social networking related data and on Elasticsearch [23] for the retrieval of social networking data. The Engine exposes its facilities by an HTTP REST Interface and comprises two services: (*i*) the *Social Networking Service* that efficiently stores and accesses social networking data (Posts, Comments, Notifications, etc.) in the underlying Cassandra Cluster, and (*ii*) the *Social Networking Indexer Service* that builds Elasticsearch indices to perform search operations over the social networking data.

Service features enabling open science & FAIRness. (*i*) every item is equipped with an actionable unique identifier (namely an URL with some metadata) that can be used for citation and access purposes; (*ii*) the enabled discussion patterns are transparent and open; every (re)action performed by a user – be it a new post, a reply to a post, or the rating of a certain post or post reply – is

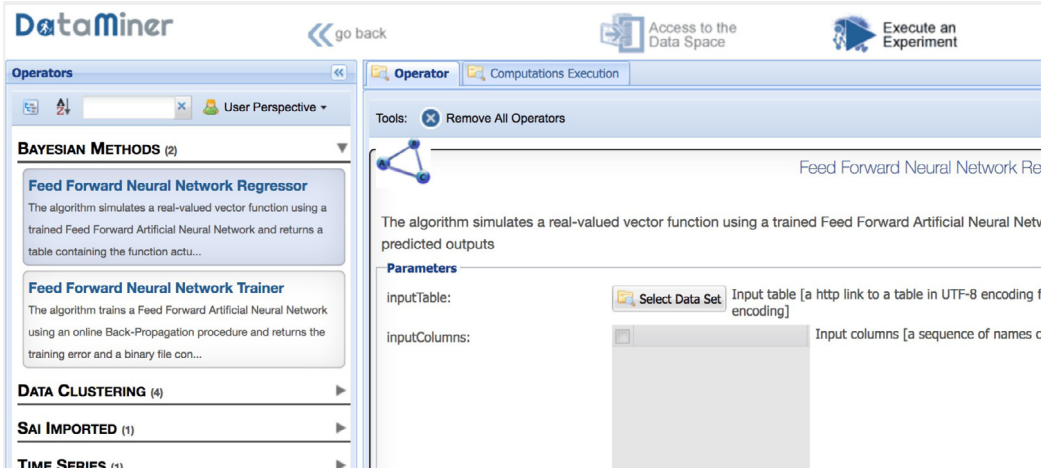


Fig. 7. D4Science data analytics platform screenshot.

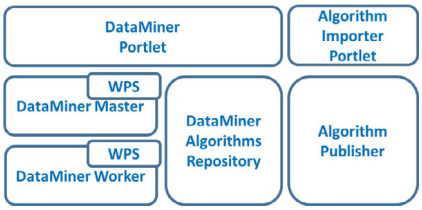


Fig. 8. D4Science analytics platform architecture.

captured and documented; (*iii*) there is no pre-defined way to structure a discussion; users can start new discussion threads, annotate them with tags for easing the cataloguing and discovery, refer to other threads and material both internally stored and available on the web.

3.3. The data analytics platform

Fig. 7 depicts the user interface of the data analytics area, i.e. the area VRE users rely on to perform their analytics tasks. It resembles a stand-alone analytics platform, e.g. Weka [24], with a collection of ready-to-use algorithms and methods, yet it relies on a distributed and heterogeneous computing infrastructure enacting the execution of complex tasks.

Fig. 8 depicts the software architecture characterising the analytics platform. The DataMiner Master is a web service in charge to accept requests for executing processes and performing the processes, either locally (in the case of processes based on local algorithms) or by relying on the DataMiner Worker(s) (in the case of processes based on distributed algorithms). The service is conceived to work in a cluster of replica services operating behind a proxy acting as load balancer. It is offered by a standard web-based protocol, i.e. OGC WPS [25]; The DataMiner Worker is a web service in charge of executing the processes it is assigned to by a Master. The service is conceived to work in a cluster of replica services and is offered by a standard web-based protocol, i.e. OGC WPS. Both the services are conceived to be deployed and operated by relying on various providers, e.g. Master and Worker instances can be deployed on private or public cloud providers. DataMiner Master and Worker instances execute processes based on an open set of algorithms hosted by a dedicated repository, the DataMiner Algorithms Repository. Two kinds of algorithms are hosted: "local" and "distributed" algorithms. Local algorithms are directly executed on a DataMiner Master instance and possibly use parallel processing on several cores and a large amount of memory. Distributed algorithms use distributed computing with a Map-Reduce approach and rely on the DataMiner Worker instances in the Worker cluster.[3] The Algorithm Importer portlet and the Algorithm Publisher service enable users to inject new algorithms into the platform by using various programming languages [26].

Service features enabling open science & FAIRness. (*i*) every process hosted by the platform is equipped with an actionable unique identifier (namely an URL with some metadata) that can be used for citation and access purposes; (*ii*) the offering and publication of user provided processes (e.g. scripts, compiled programs) by an as-a-Service standard-based approach (processes are described and exposed by the OGC WPS); (*iii*) the ability to manage and support processes produced by using several programming languages (e.g. R, Java, Fortran, Python); (*iv*) the automatic production of a detailed provenance record for every analytics task executed by the platform, i.e. the overall input/output data, parameters, and metadata that would allow a user to reproduce and repeat the task are stored into the workspace and documented by a PROV-O-based accompanying record; (*v*) integration with the shared workspace to implement collaborative experimental spaces, e.g. users can easily share datasets, methods, code; (*vi*) support for Cloud computing using a Map-Reduce approach for computing and data intensive processing; (*vii*) extensibility of the platform to quasi-transparently rely on and adapt to a distributed, heterogeneous and elastically provided array of workers to execute the processing tasks.

3.4. The publishing platform

Fig. 9 depicts the user interface of the publishing platform, i.e. the facility VRE users rely on to announce and be informed on the availability of certain artefacts at diverse maturity levels. It resembles a catalogue of artefacts with search and browse, yet the openness with respect to the typologies of products published, the metadata to document them as well as the integration with the rest make it a flexible environment. Every published item in the catalogue is characterised by (*i*) a *type*, which highlights its features and allows an easier search, (*ii*) an open ended set of metadata which carefully describe the item, and (*iii*) optional resource(s) representing the actual payload of the item.

[3] For performance reasons, it is recommended that the machines hosting the DataMiner Worker instances of a cluster reside in a data centre connecting them with a shared storage by a high-speed network.

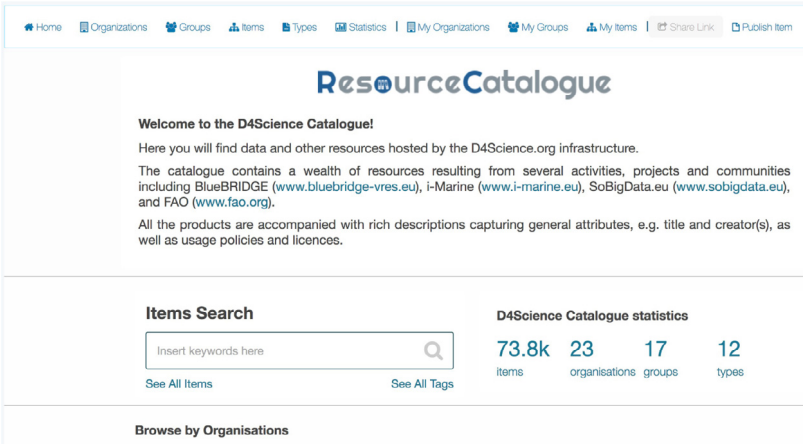


Fig. 9. D4Science data publishing platform screenshot.

Fig. 10. D4Science publishing platform architecture.

Fig. 10 depicts the software architecture characterising the publishing platform. This platform primarily relies on CKAN technology, i.e. an open source software enabling users to build and operate open data portals/catalogues.[4] This core technology has been wrapped and extended by means of the *Catalogue Service*, a component realising the business logic of the publishing platform. The Catalogue Service enacts the management of Catalogue Item Types, i.e. diverse typologies of items to be supported can be specified. Each catalogue item type carefully defines the metadata elements characterising the item typology by specifying the names of the attributes, the possible values, whether an attribute is a single instance or a repeatable one. In addition to that, each item type contains directives on how to exploit attributes for items organisational purposes, e.g. automatically transform values in tags or exploit the values for creating collections or groups of items. On top of this Catalogue Service, gCube offers several components to make publication of items easier for VRE users and services. A Catalogue Portlet, accessible in each VRE, allows navigating the catalogue content as well as publishing content by exploiting the Publishing Widget. This widget is also embedded into the Workspace portlet, so users can publish folders and/or files directly from there. External services can access the catalogue content and publish new items via the gCube Catalogue RESTful APIs. The Catalogue Service relies on the Workspace and Storage Hub (cf. Section 3.1) for storing the payload of the published items.

Service features enabling open science & FAIRness. (*i*) every catalogue item is equipped with an actionable, persistent, unique identifier (namely a PURL) that can be used for citation and access purposes; (*ii*) whenever a catalogue item is published, the associated payload(s) is stored in a persistent storage area to guarantee its long-term availability; (*iii*) every catalogue item is equipped with a license carefully characterising the possible (re-)uses; (*iv*) every publication of an item leads to the automatic production of a post in the social networking area of the VRE to inform its members; (*v*) every catalogue item is equipped with rich and open metadata, i.e. it is possible to carefully customise the typologies of products and the accompanying metadata to the community needs; (*vi*) catalogue contents (item's metadata and resources) is made available for consumption by clients by the RESTful API as well as by several approaches aiming at facilitating the interoperability (data transfer) between the catalogue and other systems, e.g. DCAT and OAI-PMH.

4. Discussion

A quantitative assessment of the effectiveness and efficiency of the facilities just described to support collaborative open science is challenging for several reasons. No benchmark exists (and will never exist) to compare with. User studies must be massive, multifaceted and continuous to avoid biases resulting from the great variety of scenarios and contexts characterising the application domain. Collaboration and open science practices are very diverse and continuously evolving across disciplines, very often they are different across homologous teams. The tools and services that can be exploited to support a certain task are many (cf. Section 2), the features homologous tools and services offer are diverse and the importance users attribute to specific features are various. Any quantitative assessment must be accompanied by qualitative discussions. These reasons somehow suggest that "the" tool or service for collaborative open science will never exist. Rather, there will be a proliferation of instruments communities can (are requested to) use and combine when facing specific needs. The interoperability among diverse instruments and their openness represent one of the most important requirements to consider when selecting the supporting tools. Regarding interoperability, D4Science services have been designed to promote their exploitation from "third-party clients and tools", e.g. the catalogue content can be acquired by OAI-PMH, analytics methods can be executed by relying on WPS, there are RESTful APIs for all the components discussed in the paper. It is also possible to populate the presented services with content residing in "third-party services", e.g. the source code of analytics methods can be acquired by Git, StorageHub plugins can be developed to populate the workspace with content residing in other tools. However, interoperability with existing tools is an almost open ended domain since new tools of interest for the communities will emerge. The key requirement is to cater for their handling, e.g. by plug-ins.

The facilities presented in this work are the result of a continuous dialogue with diverse communities, stakeholders and

[4] CKAN technology website https://ckan.org/.

use cases. In fact, they have been designed and developed in the context of several projects and initiatives supporting real communities, often multidisciplinary ones, involved in research investigations in fields including biological sciences, earth and environmental sciences, agricultural sciences, social sciences and humanities. Concrete use cases have been discussed in previous works, e.g. [27–31].

For the sake of this paper we briefly report one recent experience in exploiting the just described offering for a computational biology case. This case is an instance of the prototypical workflow previously discussed (cf. Section 3 and Fig. 2). Coro et al. [32] 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. Their method generated potential geographical reachability maps for today and 2050 by combining observations of the puffer fish in its native area, with environmental parameters and recent observations in the invaded area. The maps were generated through the combination of seven models. The authors report a risk estimate of the invasion in different subdivisions of the Mediterranean Sea related to marine resources exploitation. They highlight westward patterns and high risk zones (e.g. the Bosporus) to support management strategies and advice decision makers. The D4Science e-Infrastructure has been exploited to support every step of the workflow, from data retrieval to models' training and projection, and results publication by promoting collaboration. Given the potential impact of the results, every step of the workflow had to be reproducible. Coro et al. method required to combine native habitat observations from large repositories (e.g. GBIF and OBIS) with environmental information from other sources (e.g. AquaMaps, NASA-NEX, Argo). The e-Infrastructure services made this possible in very short time with respect to a "manual" approach, which would have required months [33]. Usage of standards to represent both observations and environmental data coming from sources already connected to the infrastructures was key in this context. The models combined by Coro et al. into the new workflow were previously implemented and made available by the analytics platform (cf. Section 3.3) by other researchers. The entire set of artefacts resulting from the investigation has been published as a *research object* via the Catalogue (cf. Section 3.4) making the access and re-use of every single part effective. For example, the maps produced in NetCDF format were reused in ecological modelling courses in European Universities,[5] where one VRE was created for each course and data and algorithms were shared with the students. Overall, the work of Coro et al. relied on the D4Science e-Infrastructure as a means to implement an open science approach, which was crucial to address and convince the target audience of the authors. Indeed, the results became part of the WWF and FAO advices about climate change implications for Mediterranean and Black Sea fisheries [34].

Overall, D4Science is currently serving more than thousands of users (more than 7000 in Feb. 2019) by more than 100 active VREs. In the period Mar. 2018–Feb. 2019 the users served by this infrastructure and its VREs performed: a total of 62,683 working sessions, with an average of circa 5223 sessions per month; a total of 7579 social interactions, with an average of circa 631 interactions per month; a total of 110 million of analytics tasks, with an average of circa 9.3 million tasks per month.

5. Conclusion and future work

This paper described a suite of tools overall realising open science-friendly working environments. These tools support all the phases of typical research lifecycles and transparently inject practices aiming at making the entire process leading to a certain version of a research artefact more transparent and repeatable without posing additional requirements for scientists. They are conceived to make the "publishing" act an easy, dynamic, comprehensive, lossless and holistic task where owners retain the control of and credit for every published artefact that, being interlinked with other artefacts and the working environment exploited for their production, cater for their effective understanding and reuse. Open publishing is the beginning of a research task rather than the concluding one.

These tools are offered with the *as-a-Service* delivery model by the D4Science infrastructure.[6] This model makes it possible (*i*) for the service provider (D4Science), to leverage economies of scale to keep developments and operational costs low; (*ii*) for the service consumer (VRE users), to acquire the services and the capacity needed in an elastic way. However, the enabling technology (gCube [4]) is freely available and can be used by institutions and organisations to set up and operate their own set of services.

Future work include the integration with recommender systems [35,36], scientific workflows [15], and research objects [14] to enlarge the possible exploitation cases and scenarios.

CRediT authorship contribution statement

M. Assante: Writing - original draft, Software. **L. Candela:** Writing - original draft, Writing - review & editing, Funding acquisition. **D. Castelli:** Writing - original draft, Funding acquisition. **R. Cirillo:** Writing - original draft, Software. **G. Coro:** Writing - original draft, Software. **L. Frosini:** Writing - original draft, Software. **L. Lelii:** Writing - original draft, Software. **F. Mangiacrapa:** Writing - original draft, Software. **P. Pagano:** Writing - original draft, Funding acquisition. **G. Panichi:** Writing - original draft, Software. **F. Sinibaldi:** Writing - original draft, Software.

Acknowledgements

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).

Declaration of competing interest

The authors declare that they have no conflict of interest.

References

[1] B. Fecher, S. Friesike, Open science: One term, five schools of thought, in: S. Bartling, S. Friesike (Eds.), Opening Science, Springer International Publishing, 2014, pp. 17–47, http://dx.doi.org/10.1007/978-3-319-00026-8_2.

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

[3] B.A. Nosek, G. Alter, G.C. Banks, D. Borsboom, S.D. Bowman, S.J. Breckler, S. Buck, C.D. Chambers, G. Chin, G. Christensen, M. Contestabile, A. Dafoe, E. Eich, J. Freese, R. Glennerster, D. Goroff, D.P. Green, B. Hesse, M. Humphreys, J. Ishiyama, D. Karlan, A. Kraut, A. Lupia, P. Mabry, T. Madon, N. Malhotra, E. Mayo-Wilson, M. McNutt, E. Miguel, E. Levy Paluck, U. Simonsohn, C. Soderberg, B.A. Spellman, J. Turitto, G. VandenBos, S. Vazire, E.J. Wagenmakers, R. Wilson, T. Yarkoni, Promoting an open research culture, Science 348 (6242) (2015) 1422–1425, http://dx.doi.org/10.1126/science.aab2374.

[5] Training courses organised by the BlueBRIDGE project http://www.bluebridge-vres.eu/training-courses

[6] D4Science Infrastructure Gateway https://services.d4science.org/

[4] M. Assante, L. Candela, D. Castelli, R. Cirilllo, G. Coro, L. Frosini, L. Lelii, F. Mangiacrapa, V. Marioli, P. Pagano, G. Panichi, C. Perciante, F. Sinibaldi, The gcube system: Delivering virtual research environments as-a-service, Future Gener. Comput. Syst. 95 (2019) 445–453, http://dx.doi.org/10.1016/j.future.2018.10.035, URL http://www.sciencedirect.com/science/article/pii/S0167739X17328364.
[5] L. Candela, D. Castelli, P. Pagano, Virtual research environments: an overview and a research agenda, Data Sci. J. 12 (2013) GRDI75–GRDI81, http://dx.doi.org/10.2481/dsj.GRDI-013.
[6] B.A. Nosek, Y. Bar-Anan, Scientific utopia: I opening scientific communication, Psychol. Inq. 23 (3) (2012) 217–243, http://dx.doi.org/10.1080/1047840X.2012.692215.
[7] M. Assante, L. Candela, D. Castelli, P. Manghi, P. Pagano, Science 2.0 repositories: Time for a change in scholarly communication, D-Lib Mag. 21 (1/2) (2015) http://dx.doi.org/10.1045/january2015-assante.
[8] B. Kramer, J. Bosman, Innovations in scholarly communication - global survey on research tool usage [version 1; referees: 2 approved], F1000Research 5 (692) (2016) http://dx.doi.org/10.12688/f1000research.8414.1.
[9] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, P. Ivanov, D. Avila, S. Abdalla, C. Willing, Jupyter Development Team, Jupyter notebooks - a publishing format for reproducible computational workflows, in: F. Loizides, B. Schmidt (Eds.), in: Positioning and Power in Academic Publishing: Players, Agents and Agendas, 20th International Conference on Electronic Publishing, 2016, pp. 87–90, http://dx.doi.org/10.3233/978-1-61499-649-1-87, Göttingen, Germany, June (2016) 7–9.
[10] A. Bardi, P. Manghi, Enhanced publications: Data models and information systems, LIBER Q. 23 (4) (2014) 240–273, http://dx.doi.org/10.18352/lq.8445.
[11] H. Staines, Digital open annotation with hypothesis: Supplying the missing capability of the web, J. Sch. Publ. 49 (3) (2018) 345–365, http://dx.doi.org/10.3138/jsp.49.3.04.
[12] M. Thelwall, K. Kousha, Researchgate: Disseminating, communicating, and measuring scholarship? J. Assoc. Inf. Sci. Technol. 66 (5) (2015) 876–889, http://dx.doi.org/10.1002/asi.23236.
[13] A. Burton, H. Koers, P. Manghi, M. Stocker, M. Fenner, A. Aryani, S. La Bruzzo, M. Diepenbroek, U. Schindler, The scholix framework for interoperability in data-literature information exchange, D-Lib Mag. 23 (1/2) (2017) http://dx.doi.org/10.1045/january2017-burton.
[14] S. Bechhofer, I. Buchan, D. De Roure, P. Missier, J. Ainsworth, J. Bhagat, P. Couch, D. Cruickshank, M. Delderfield, I. Dunlop, M. Gamble, D. Michaelides, S. Owen, D. Newman, S. Sufi, C. Goble, Why linked data is not enough for scientists, Future Gener. Comput. Syst. 29 (2) (2013) 599–611, http://dx.doi.org/10.1016/j.future.2011.08.004.
[15] C.S. Liew, M.P. Atkinson, M. Galea, T.F. Ang, P. Martin, J.I.V. Hemert, Scientific workflows: Moving across paradigms, ACM Comput. Surv. 49 (4) (2016) 66:1–66:39, http://dx.doi.org/10.1145/3012429.
[16] D. De Roure, C. Goble, R. Stevens, The design and realisation of the myexperiment virtual research environment for social sharing of workflows, Future Gener. Comput. Syst. 25 (5) (2009) 561–567, http://dx.doi.org/10.1016/j.future.2008.06.010.
[17] M. McLennan, R. Kennell, Hubzero: A platform for dissemination and collaboration in computational science and engineering, Comput. Sci. Eng. 12 (2) (2010) 48–53, http://dx.doi.org/10.1109/MCSE.2010.41.
[18] M. Barker, S.D. Olabarriaga, N. Wilkins-Diehr, S. Gesing, D.S. Katz, S. Shahand, S. Henwood, T. Glatard, K. Jeffery, B. Corrie, A. Treloar, H. Glaves, L. Wyborn, N.P. Chue Hong, A. Costa, The global impact of science gateways, virtual research environments and virtual laboratories, Future Gener. Comput. Syst. 95 (2019) 240–248, http://dx.doi.org/10.1016/j.future.2018.12.026.
[19] A.C. Catlin, C. HewaNadungodage, S. Pujol, L. Laughery, C. Sim, A. Puranam, A. Bejarano, A cyberplatform for sharing scientific research data at datacenterhub, Comput. Sci. Eng. 20 (3) (2018) 49–70, http://dx.doi.org/10.1109/MCSE.2017.3301213.
[20] M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L.B. da Silva Santos, P.E. Bourne, J. Bouwman, A.J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran, A.J.G. Gray, P. Groth, C. Goble, J.S. Grethe, J. Heringa, P.A.C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S.J. Lusher, M.E. Martone, A. Mons, A.L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M.A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The FAIR guiding principles for scientific data management and stewardship, Sci. Data 3 (2016) http://dx.doi.org/10.1038/sdata.2016.18, 160018 EP. URL http://dx.doi.org/101038/sdata.201618.
[21] 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, Commun. ACM 53 (4) (2010) 50–58, http://dx.doi.org/10.1145/1721654.1721672.
[22] J. Carpenter, E. Hewitt, Cassandra: The Definitive Guide, second ed., O'Reilly Media, 2016.
[23] C. Gormley, Z. Tong, Elasticsearch: The Definitive Guide, O'Reilly Media, 2015.
[24] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software: An update, ACM SIGKDD Explor. Newsl. 11 (1) (2009) 10–18, http://dx.doi.org/10.1145/1656274.1656278.
[25] G. Coro, G. Panichi, P. Scarponi, P. Pagano, Cloud computing in a distributed e-infrastructure using the web processing service standard, Concurr. Comput.: Pract. Exper. 29 (18) (2017) e4219, http://dx.doi.org/10.1002/cpe.4219.
[26] G. Coro, L. Candela, P. Pagano, A. Italiano, L. Liccardo, Parallelizing the execution of native data mining algorithms for computational biology, Concurr. Comput.: Pract. Exper. 27 (17) (2015) 4630–4644, http://dx.doi.org/10.1002/cpe.3435.
[27] R. Froese, J.T. Thorson, R.B.J. Reyes, A Bayesian approach for estimating length-weight relationships in fishes, J. Appl. Ichthyol. 30 (1) (2014) 78–85, http://dx.doi.org/10.1111/jai.12299.
[28] G. Coro, C. Magliozzi, A. Ellenbroek, P. Pagano, Improving data quality to build a robust distribution model for architeuthis dux, Ecol. Model. 305 (2015) 29–39, http://dx.doi.org/10.1016/j.ecolmodel.2015.03.011.
[29] G. Coro, S. Large, C. Magliozzi, P. Pagano, Analysing and forecasting fisheries time series: purse seine in Indian Ocean as a case study, ICES J. Mar. Sci. (2016) fsw131, http://dx.doi.org/10.1093/icesjms/fsw131.
[30] G. Coro, P. Pagano, A. Ellenbroek, Combining simulated expert knowledge with neural networks to produce ecological niche models for *Latimeria chalumnae*, Ecol. Model. 268 (2013) 55–63.
[31] G. Coro, T.J. Webb, W. Appeltans, N. Bailly, A. Cattrijsse, P. Pagano, Classifying degrees of species commonness: North sea fish as a case study, Ecol. Model. 312 (2015) 272–280.
[32] G. Coro, L. Gonzalez Vilas, C. Magliozzi, A. Ellenbroek, P. Scarponi, P. Pagano, Forecasting the ongoing invasion of *Lagocephalus sceleratus* in the Mediterranean sea, Ecol. Model. 371 (2018) 37–49, http://dx.doi.org/10.1016/j.ecolmodel.2018.01.007.
[33] L. Candela, D. Castelli, G. Coro, L. Lelii, F. Mangiacrapa, V. Marioli, P. Pagano, An infrastructure-oriented approach for supporting biodiversity research, Ecol. Inform. 26 (2015) 162–172, http://dx.doi.org/10.1016/j.ecoinf.2014.07.006, Information and Decision Support Systems for Agriculture and Environment.
[34] FAO, Report of the Expert Meeting on Climate Change Implications for Mediterranean and Black Sea Fisheries, FAO Fisheries and Aquaculture Report 1233, Food and Agriculture Organizations of the United Nations, Rome, Italy, 2018, URL http://www.fao.org/3/I9528EN/i9528en.pdf.
[35] H. Avancini, L. Candela, U. Straccia, Recommenders in a personalized, collaborative digital library environment, J. Intell. Inf. Syst. 28 (3) (2007) 253–283.
[36] J. Bobadilla, F. Ortega, A. Hernando, A. Gutiérrez, Recommender systems survey, Knowl.-Based Syst. 46 (2013) 109–132, http://dx.doi.org/10.1016/j.knosys.2013.03.012.

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 several EU-funded projects dealing with the design and development of innovative soluitions for research data management. He was the Project Manager of the BlueBRIDGE Project and currently he is the CNR lead person in the AGINFRA+ and ENVRIPlus projects. His research interests include Data Infrastructures, Virtual Research Environments, Data Publishing, Open Science, Scholarly Communication, Digital Library [Management] Systems and Architectures, Digital Libraries Models, Distributed Information Retrieval, and Grid and Cloud Computing.

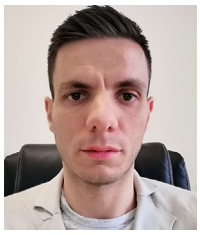
Donatella Castelli is a Research Director at Istituto di Scienza e Tecnologie dellInformazione "A. Faedo" of the Italian National Research Council (CNR-ISTI). She has been the principal investigator of several European and International funded projects on digital libraries and data e-Infrastructures acquiring considerable experience in these domains. Currently, she is the technical coordinator of OpenAIRE, the Open Access infrastructure for Research in Europe. She has lead the definition of the Architecture of the European Open Science Cloud in the EU EOSCPilot project. She also participated in starting-up the D4Science data Infrastructure and in supporting its evolution since now. Her scientific interests are centered on data modeling, data interoperability and data infrastructures.

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 Open Science. 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 Task 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.

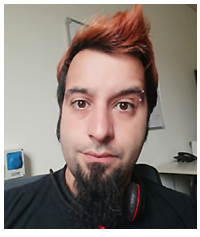
Lucio Lelii is Researcher at the Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, Italy. His scientific and professional activity involves research and development of services for Data Infrastructures. He is currently member of the BlueBRIDGE EU Project.



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.

Pasquale Pagano is a Senior Researcher at the Networked Multimedia Information Systems Laboratory of the Istituto di Scienza e Tecnologie della Informazione "A. Faedo" (ISTI) of the Italian National Research Council (CNR). He received the M.Sc. in Information Systems Technologies from the Department of Computer Science of the University of Pisa (1998), and the Ph.D. degree in Information Engineering from the Department of Information Engineering: Electronics, Information Theory, Telecommunications of the same university (2006). The aim of his research is the study and experimentation of 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, for the management of heterogenous data sources. He has a strong background on distributed architectures. He is currently the Technical Director of D4Science, the Hybrid Data Infrastructure serving scientists in more than 40 countries world-wide, and chief manager of gCube software, the open-source platform for the management and operation of scientific data infrastructures.

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 Service, 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).



Fabio Sinibaldi is a Researcher at CNR-ISTI. He holds a degree in computer science engineering with specialisation 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.