Enacting Open Science by gCube

Massimiliano Assante, Leonardo Candela, Donatella Castelli, Gianpaolo Coro,
Francesco Mangiacrapa, Pasquale Pagano, Costantino Perciante
Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" – Consiglio Nazionale delle Ricerche
via G. Moruzzi, 1 – 56124, Pisa, Italy
Email: {name.surname}@isti.cnr.it

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 gCube offering to enact Open Science-friendly Virtual Research Environments. In particular, the paper describes how a complete solution suitable for realising Open Science practices is achieved by a social networking collaborative environment in conjunction with a shared workspace, an open data analytics platform, and a catalogue enabling FAIR principles on every research artefact.**

Keywords—**Open Science; gCube; D4Science; Social Networking; Analytics; Publishing;**

I. Introduction

Open Science is promising to revolutionise the entire science enterprise by envisioning and proposing new practices aiming at making science better [1]–[3]. The promised benefits include (*i*) better interpretation, understanding and re-producibility 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 scientists.

These benefits are destined to remain a wish if the research community as a whole (funding agencies, research performing organizations, publishers, research infrastructures, scientists, as well as citizens) does not fully embrace Open Science and put in place efforts and initiatives aiming at making it the norm. The good news is that the movement is gaining momentum and consensus. In fact, funding agencies are developing policies supporting its implementation as well as are supporting the development of infrastructures and services. Moreover, research infrastructures start offering services and facilities going in the direction of Open Science. Scientists 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 this movement is confronting with a number of barriers including: (*a*) cultural factors, e.g. the fear to lose the control and exploitation of datasets; (*b*) cost-based factors, e.g. the extra-effort needed to make a research-artefact exploitable by a user other than

the initial owner; and (*c*) disincentive factors, e.g. the effort spent in "publishing" research artefacts going beyond papers is receiving little or no value in researchers assessment and career success.

This paper describes the solution proposed by gCube [4] to overcome some of the above mentioned barriers – in particular those tied to cost-based arguments – by providing researchers and practitioners with a working environment where Open Science practices are transparently promoted. gCube is 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 community working on a research question. Beside providing their users with the domain-specific facilities, i.e. datasets and services suitable for the 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 provided by a user or provided by others to be applied to 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].

II. Related Works

There are plenty of tools and approaches supporting Open Science [8]. They include (*a*) repositories maintaining different versions of datasets and software to promote their citation and reuse, e.g. Dryad, GitHub, Zenodo, figshare; (*b*) 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]; (*c*) tools aiming at making more open, transparent, holistic and participative the research assessment process, e.g. open peer review, post-publication review, annotation and commenting tools, social networks for scientists like ResearchGate [10]. One of the major barrier
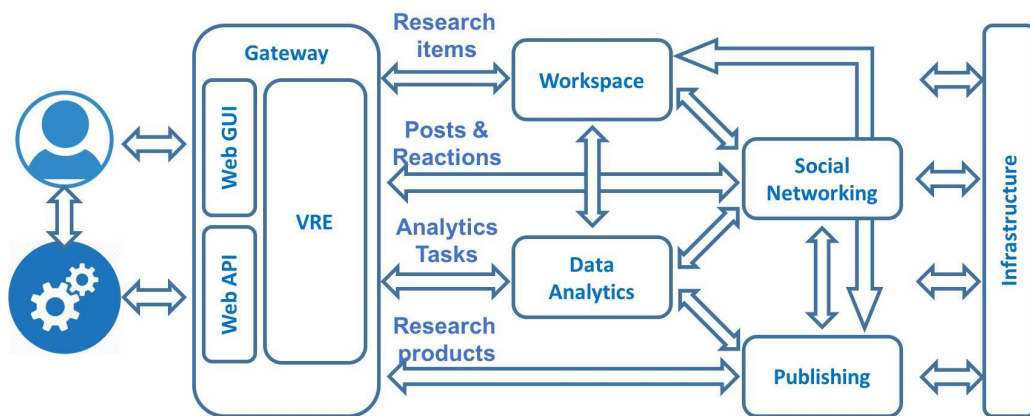
Fig. 1. gCube-based Open Science Framework

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. [11]. However this link-based mechanism is quite fragile and costly to keep healthy and up to date, and this is leading to research packaging formats [12].

Scientific workflows technologies [13] are adopted by an increasing number of communities to automate scientific methods and procedures. Environments enabling to publish and share workflows exist, yet the guarantees that the method captured by the workflow seamlessly works in settings other then the originator ones are limited. Moreover, the act of publishing workflows is not systematic across communities and contexts.

The Open Science Framework is a web-based service enabling 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 with whom. This platform share commonalities with the gCube based set of facilities described in this paper, e.g. a project is like a VRE, yet the VRE based approach make available to its users in one place the entire set of facilities and services they need to perform their research, thus going beyond the pure sharing of material.

III. VIRTUAL RESEARCH ENVIRONMENTS BASED SCIENTIFIC WORKFLOWS

Figure 1 depicts the main components and facilities offered by gCube to support collaborative activities and enable Open Science practices, namely a shared workspace (cf. Sec. III-A), a social networking area (cf. Sec. III-B), a data analytics platform (cf. Sec. III-C), and a publishing platform (cf. Sec. III-D). These components are all correlated each other and realize 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 (a) 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; (b) are conceived to open up research artefacts, independently of their maturity level, beyond VRE boundaries (no lock-in) yet according to artefact owner 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); (c) 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 [14], 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.
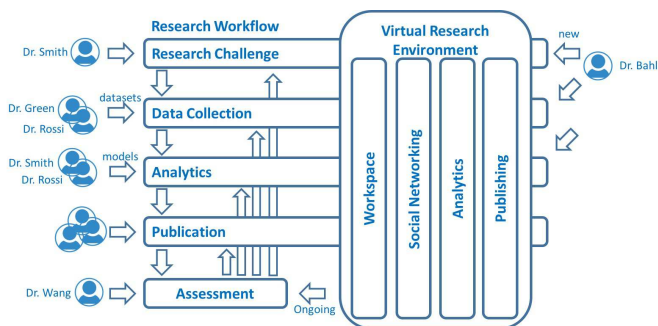

Fig. 2. gCube-based Open Science Workflow

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 and make them available (data analytics); (*iv*) the availability of these early-results suggests 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 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 start 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 and Dr. Wang's one to realise that Dr. Wang model better suits with their initial hypothesis (all of this happen well before their paper being published). This representative workflow can be easily and effectively implemented only by relying on a suite of facilities like those offered by gCube where the "place" where research activity is conducted and the "place" 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 feasible at all.

A. The Workspace Platform

Figure 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 (*a*) equipped with rich and extensible metadata and (*b*) actually stored by an array of storage solutions [4].

Figure 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
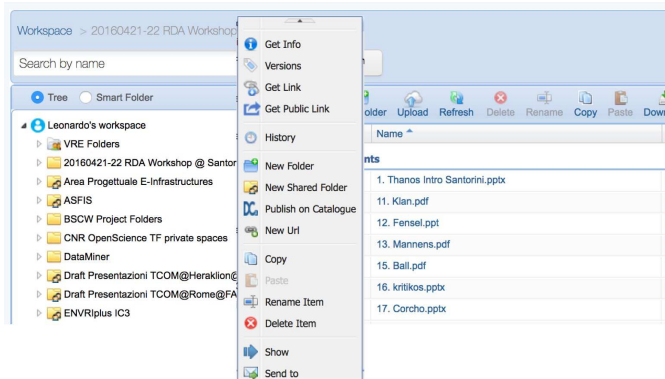

Fig. 3. gCube Workspace screenshot

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.
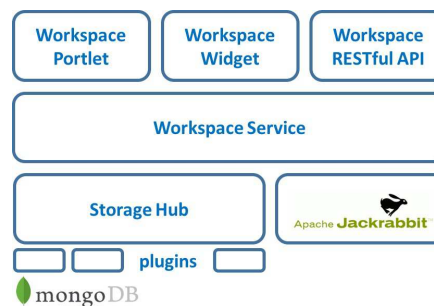

Fig. 4. gCube Workspace Platform Architecture

The distinguishing features of this platform for Open Science are the following: (*i*) every workspace item is equipped with an actionable unique identifier 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 the 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 for the owner only; *shared*, content is available for selected users decided by the owner; and *VRE folder*, content is available to VRE members; (*v*) the workspace is tightly integrated with both the social networking and catalogue for easing the dissemination of its artefact (either single items or groups of items).

B. The Social Networking Collaborative Platform

Figure 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 network with posts, tags, mentions, comments and reactions, yet its

integration with the rest makes it a powerful and flexible communication channel for scientists.
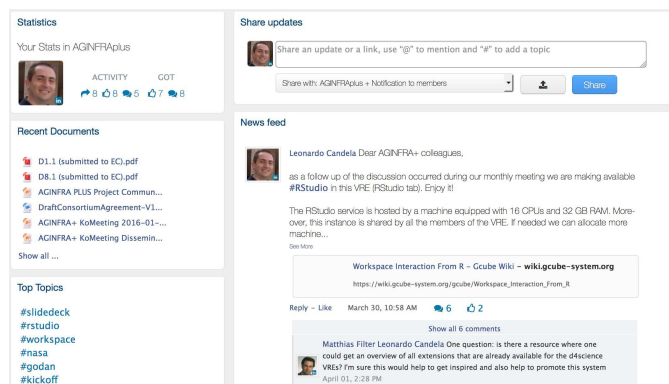

Fig. 5. gCube Social Networking screenshot

Figure 6 depicts the software architecture characterising the social networking collaborative platform. This facility relies on the Social Networking Engine, a Cassandra database [15] for storing social networking related data and Elasticsearch [16] 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 store 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.
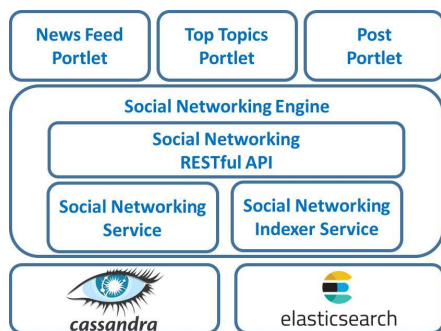

Fig. 6. gCube Social Networking Collaborative Platform Architecture

The distinguishing features of this platform for Open Science are the following: (*i*) every item is equipped with an actionable unique identifier that can be used for citation and access purposes; (*ii*) the discussion patterns enabled are really 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 carefully 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.

C. The Data Analytics Platform

Figure 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, with a collection of ready-to-use algorithms and methods, yet it relies on a distributed and heterogeneous computing infrastructure enacting to executed complex tasks.
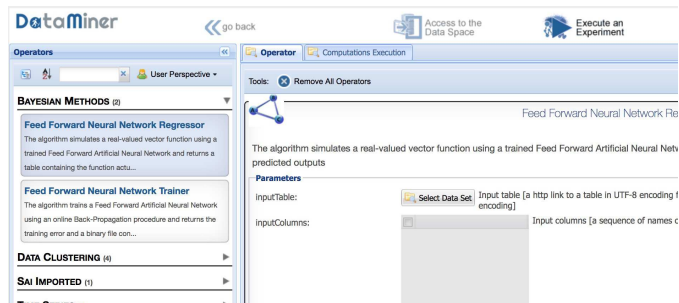

Fig. 7. gCube Data Analytics Platform screenshot

Figure 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 executing them, either locally or by relying on the DataMiner Worker(s) depending from the specific process. 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[1]; The DataMiner Worker is a web service in charge to execute 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. The Algorithm Importer portlet and the Algorithm Publisher service enable users to inject new algorithms into the platform by using various programming languages [17].

The distinguishing features of this platform for Open Science are the following: (*i*) every process hosted by the platform is equipped with an actionable unique identifier 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 (pro-

[1]OGC Web Processing Servicehttp://www.opengeospatial.org/standards/wps
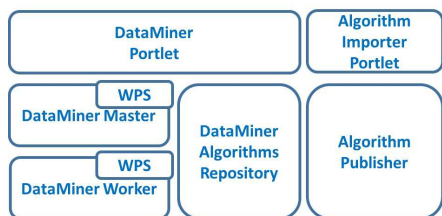
Fig. 8. gCube Analytics Platform Architecture

cesses are described and exposed by the OGC Web Processing Service standard); (*iii*) the ability to manage and support processes produced by using several programming languages (e.g. R, Java, Fortran, Phyton); (*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 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.

D. The Publishing Platform

Figure 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.

Figure 10 depicts the software architecture characterising the publishing platform. This platform primarily relies on CKAN technology, i.e. an open source software enabling to build and operate open data portals / catalogues [2]. 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 enact the management of Catalogue Item Types, i.e. specifications of diverse typologies of items supported. 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 single instance or repeatable. In addition to that, each item type contains

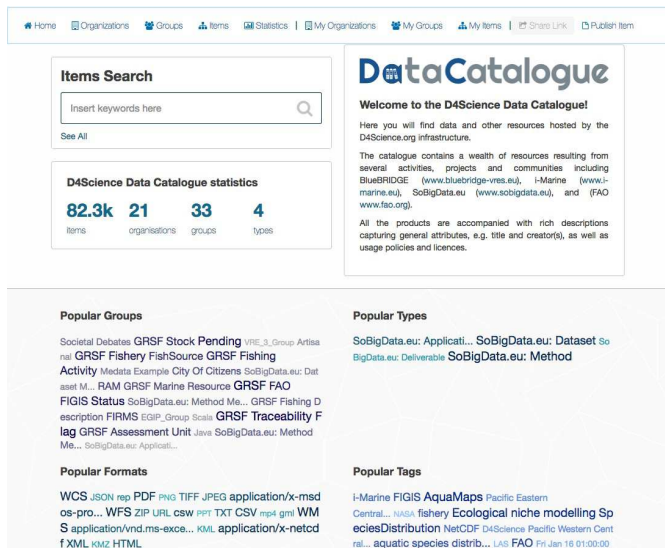[2]CKAN technology website https://ckan.org/

Fig. 9. gCube Data Publishing Platform screenshot

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 to navigate the catalogue content as well as to publish 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. Sec. III-A) for storing the payload of the published items.

Fig. 10. gCube Publishing Platform Architecture

The distinguishing features of this platform for Open Science are the following: (*i*) every catalogue item is equipped with an actionable, persistent, unique identifier 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.

IV. 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 ones.

These tools are an essential part of the gCube Open Source technology [4]. They are offered as-a-Service by means of the D4Science.org infrastructure [18]. Concrete exploitation cases and experiences demonstrate their effectiveness, e.g. [19]–[23].

Future work include the integration with recommender systems [24], [25], scientific workflows [13], and research objects [12] to enlarge the possible exploitation cases and scenarios.

References

[1] B. Fecher and S. Friesike, "Open science: One term, five schools of thought," in *Opening Science*, S. Bartling and S. Friesike, Eds. Springer International Publishing, 2014, pp. 17–47.

[2] S. Bartling and S. Friesike, "Towards another scientific revolution," in *Opening Science*. Springer International Publishing, 2014, pp. 3–15.

[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, and T. Yarkoni, "Promoting an open research culture," *Science*, vol. 348, no. 6242, 2015.

[4] M. Assante, L. Candela, D. Castelli, G. Coro, L. Lelii, and P. Pagano, "Virtual research environments as-a-service by gcube," *PeerJ Preprints*, 2016.

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

[6] B. A. Nosek and Y. Bar-Anan, "Scientific utopia: I. opening scientific communication," *Psychological Inquiry*, vol. 23, no. 3, pp. 217–243, 2012.

[7] M. Assante, L. Candela, D. Castelli, P. Manghi, and P. Pagano, "Science 2.0 repositories: Time for a change in scholarly communication," *D-Lib Magazine*, vol. 21, no. 1/2, 2015.

[8] B. Kramer and J. Bosman, "Innovations in scholarly communication - global survey on research tool usage [version 1; referees: 2 approved]," *F1000Research*, vol. 5, no. 692, 2016.

[9] A. Bardi and P. Manghi, "Enhanced publications: Data models and information systems," *LIBER Quarterly*, vol. 23, no. 4, pp. 240–273, 2014.

[10] M. Thelwall and K. Kousha, "ResearchGate: Disseminating, communicating, and measuring scholarship?" *Journal of the Association for Information Science and Technology*, vol. 66, no. 5, pp. 876–889, 2015.

[11] A. Burton, H. Koers, P. Manghi, M. Stocker, M. Fenner, A. Aryani, S. La Bruzzo, M. Diepenbroek, and U. Schindler, "The scholix framework for interoperability in data-literature information exchange," *D-Lib Magazine*, vol. 23, no. 1/2, 2017.

[12] 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, and C. Goble, "Why linked data is not enough for scientists," *Future Generation Computer Systems*, vol. 29, no. 2, pp. 599–611, 2013.

[13] C. S. Liew, M. P. Atkinson, M. Galea, T. F. Ang, P. Martin, and J. I. V. Hemert, "Scientific workflows: Moving across paradigms," *ACM Computing Surveys*, vol. 49, no. 4, 2016.

[14] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "A view of cloud computing," *Communications of the ACM*, vol. 53, no. 4, pp. 50–58, Apr. 2010.

[15] J. Carpenter and E. Hewitt, *Cassandra: The Definitive Guide, 2nd Edition*. O'Reilly Media, 2016.

[16] C. Gormley and Z. Tong, *Elasticsearch: The Definitive Guide*. O'Reilly Media, 2015.

[17] G. Coro, L. Candela, P. Pagano, A. Italiano, and L. Liccardo, "Parallelizing the execution of native data mining algorithms for computational biology," *Concurrency and Computation: Practice and Experience*, vol. 27, no. 17, pp. 4630–4644, 2015.

[18] L. Candela, D. Castelli, A. Manzi, and 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*, ser. Proceedings of Science, 2014.

[19] R. Froese, J. T. Thorson, and R. B. J. Reyes, "A bayesian approach for estimating length-weight relationships in fishes," *Journal of Applied Ichthiology*, vol. 30, no. 1, pp. 78–85, 2014.

[20] G. Coro, C. Magliozzi, A. Ellenbroek, and P. Pagano, "Improving data quality to build a robust distribution model for architeuthis dux," *Ecological Modelling*, vol. 305, pp. 29–39, 2015.

[21] G. Coro, S. Large, C. Magliozzi, and P. Pagano, "Analysing and forecasting fisheries time series: purse seine in indian ocean as a case study," *ICES Journal of Marine Science: Journal du Conseil*, p. fsw131, 2016.

[22] G. Coro, P. Pagano, and A. Ellenbroek, "Combining simulated expert knowledge with neural networks to produce ecological niche models for latimeria chalumnae," *Ecological modelling*, vol. 268, pp. 55–63, 2013.

[23] G. Coro, T. J. Webb, W. Appeltans, N. Bailly, A. Cattrijsse, and P. Pagano, "Classifying degrees of species commonness: North sea fish as a case study," *Ecological Modelling*, vol. 312, pp. 272–280, 2015.

[24] H. Avancini, L. Candela, and U. Straccia, "Recommenders in a Personalized, Collaborative Digital Library Environment," *Journal of Intelligent Information Systems*, vol. 28, no. 3, pp. 253–283, June 2007.

[25] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez, "Recommender systems survey," *Knowledge-Based Systems*, vol. 46, pp. 109–132, 2013.