Data Use – Virtual Research Environments, by Leonardo Candela

STATE OF THE ART

A recent study promoted by The Royal Society in cooperation with Elsevier reviewed the changing patterns of science and scientific collaborations and confirmed that science is increasingly global, multipolar and networked (Llewellyn Smith, et al., 2011). This trend calls for innovative, dynamic and ubiquitous research supporting environments where scattered scientists can seamlessly access data, software, and processing resources managed by diverse systems in separate administration domains through their web browser.

Dependent on context, these environments are commonly referred to as either *Virtual Research Environments* (Carusi & Reimer, 2010), *Science Gateways* (Wilkins-Diehr, 2007), *Collaboratories* (Wulf, 1993), *Digital Libraries* (Candela, Castelli, & Pagano, 2011) or *Inhabited Information Spaces* (Snowdon, Churchill, & Frécon, 2004). These environments are among the goals that *e-Infrastructures* (e-Infrastructure Reflection Group, 2010) and *cyberinfrastructures* (Cyberinfrastructure Council, 2007) are going to realise. A variety of systems and services fall within the scope of these definitions, from ad-hoc portals with minimal access services to content resources held in external repositories (lightweight integration to promote resource discovery) to general-purpose management systems with advanced services defined over a wide range of resources (strong integration to promote resource exploitation). In some cases, motivations and design (sharing, on-demand resource provision, economies of scale) align with the principles of *grid computing* and its ecology of *virtual organizations* (Foster & Kesselman, 1998) as well as with *cloud computing* (Foster, Zhao, Raicu, & Lu, 2008).

For the purposes of this paper, the term Virtual Research Environment (VRE) is used with a comprehensive scope, i.e. it represents a concept overarching all the environments cited above, and identifies a system with the following distinguishing features: (*i*) it is a web-based working environment; (*ii*) it is tailored to serve the needs of a *community of practice* (Lave & Wenger, 1991); (*iii*) it is expected to provide a community of practice with the whole array of commodities needed to accomplish the community's goal(s); (*iv*) it is open and flexible with respect to the overall service offering and lifetime; and (*v*) it promotes fine-grained controlled sharing of both intermediate and final research results by guaranteeing ownership, provenance and attribution.

The VREs' characteristic of being a web based working environment is the most common one and usually that which contributes to misuse of the term "VRE" itself. In many cases, ad-hoc portals implementing simple catalogue facilities and completely missing the collaborative, dynamicity and openness features discussed

above have been tagged with the "VRE" term (Allan, 2009). Allan explains how Web-based services should be loosely combined into portals to provide a comprehensive infrastructure for the support of research across all academic disciplines. He feels that "VRE" portals should not only provide an environment for housing, indexing, and retrieving large data sets but also leverage Web 2.0 technologies (O'Reilly, 2005) and social networking solutions to give researchers a comprehensive environment for collaboration and resource discovery.

The VREs' characteristic of being the framework expected to support communities of practice is what makes VREs definition very heterogeneous and VREs implementation a challenging activity. "Community of practice" is a term coined to capture an "activity system" that includes individuals who are united in action and in the meaning that "action" has for them and for the larger collective. The communities of practice are "virtual", i.e. they are not formal structures, such as departments or project teams. Instead, these communities exist in the minds of their members, are glued together by the connections they have with each other, as well as by their specific shared problems or areas of interest. The generation of knowledge in communities of practice occurs when people participate in problem solving and share the knowledge necessary to solve the problems (Wenger, 1998). Creating and supporting communities of practice as a strong alternative to building teams was an early observation (Nirenberg, 1994). This is particularly true in science and scientific collaborations as confirmed by the Royal Society study previously cited. It is evident that realising working environments for communities with features for communities of practice is a challenging task: the service has to be guaranteed at a level of quality of service although the requirements and needs are highly evolving and the membership is volatile.

The VREs' characteristic of being the system which offers the whole array of needed commodities is another aspect concurring to difficulties in defining VREs' scope boundaries and enlarging realization challenges. The larger the pool of expected commodities (both quantitatively and qualitatively) the bigger the effort needed to implement the related VRE. It is quite common to describe VRE's commodities by decoupling the resources managed through the VRE from the VRE's services facilitating resources management. Resources range from data sets, collections, storage facilities and computing power to services realising specific utilities and research objects. Research objects themselves evolve from traditional research outputs, like papers and experimental data, to *living reports* (Candela, Castelli, Pagano, & Simi, 2005; Candela, et al., 2007), *executable research papers* (Van Gorp & Mazanek, 2011; Nowakowski, et al., 2011), *scientific workflows* (De Roure, Goble, & Stevens, 2009). In addition to that, a VRE is required to offer a unified and sometimes virtualised view on a pool of resources that might come from different "providers".

The VREs' characteristic of being open and flexible with respect to the overall service offering brings development approaches into question. Traditional approaches, mainly based on from scratch development of ad-hoc portals, are not sustainable in community-of-practice-oriented scenarios. There is a need for innovative approaches aimed at promoting and maximising sharing and reuse of existing commodities to build and operate a number of VREs. In the context of the DILIGENT, D4Science, D4Science-II triplet of EU projects an approach has been developed and deployed based on: (i) an infrastructure making available a rich pool of resources including datasets, computing power and hosting machines; (ii) a software framework offering re-

source management facilities for a rich array of resources including software packages; (iii) a wizard-based mechanism allowing users to characterize the VRE they are interested in; and (iv) automatic VRE deployment facilities that acquire the constituents needed to satisfy the VRE specification by relying on the infrastructure and software framework offering (Assante, et al., 2008). In the three projects, the intended communities of practice ranged from humanities research to biodiversity. Moreover, this initiative is among the first to rely on cloud technologies to implement VREs (Candela, Castelli, & Pagano, 2010).

Finally, the VREs' characteristic of supporting fine-grained controlled sharing of both intermediate and final research results while guaranteeing ownership, provenance and attribution is somehow a consequence of the scenarios VREs are going to serve. Many science users will not be willing to contribute unless mechanisms guaranteeing their work are in place (De Roure, Goble, & Stevens, 2009). These mechanisms can be either *explicit*, e.g. the visibility of a resource is defined by its creator/owner through a set of policies, or *implicit*, e.g. it is the framework implementing the VRE which injects provenance metadata in the research outputs.

10-YEAR VISION

In ten years, it is expected that the trend characterizing science and scientific collaborations discussed above continues, thus becoming the "default" approach for scientific investigations as well as for any societal collaboration-based activity. Virtual Research Environments will be integrated into standard practices and tools used by communities of practice, thus becoming the "enabler" working environments for implementing investigation and collaboration activities efficiently and effectively.

The creation and management of Virtual Research Environments will be a very straightforward process that relies on specific services – VRE Management Services – built atop a "global virtual infrastructure" resulting from the aggregation and interoperation of a number of existing infrastructures and systems. The VRE Management Services will support the phases of VRE definition, deployment and monitoring / maintenance.

The VRE definition phase will guide an authorised actor of an application domain in characterizing the expected VRE service in very abstract terms, e.g. defining the policies and procedures governing the VRE community building, defining the policies governing the VRE operation, identifying the datasets the VRE community is willing to play with, describing the data types the VRE community is going to manage, identifying the facilities the VRE is requested to support. Which characterizations are allowed depends on the current offering of the "global virtual infrastructure", i.e. the "global virtual infrastructure" is actually playing the role of "resources provider" and a VRE will be an application built by dynamically acquiring the needed constituents from the overall offering. The quality of service of the resulting VRE is declared in its specification, thus it is known a-priori and depends on the amount of resources spent to acquire the resources needed to realize the VRE.

The VRE deployment phase will be almost automatic. The Management Services will crunch the specification of the expected VRE service including (i) the directives on the quality of service and (ii) the available budget to identify the "optimal" set of resources to be acquired from the "global virtual infrastructure". The Manage-

ment Services will take care of creating the application context changing this set of resources from a complex whole into an integrated system.

The VRE monitoring / maintenance phase will require little direct human control. The Management Services will take care of checking the state of the set of resources allocated to implement the VRE service. When needed, they will perform corrective actions aiming at guaranteeing that the VRE service specification is satisfied, e.g. by dynamically acquiring new resources from the "global virtual infrastructure".

The resulting Virtual Research Environment will be very flexible and customizable. Every single user can simply define its own workflows – workflows realizing a scientific investigation – by combining existing facilities without taking care of implementation details and computational resources acquisition. The computational resources as well as the workflow constituents will be dynamically acquired and combined by the Management Service, in accordance with the VRE specification.

Thus Virtual Research Environments creation and management will become a societal and organisational process rather than a technological one.

CURRENT CHALLENGES

There are three major issues to be resolved to realise the above vision as well as to implement sustainable Virtual Research Environments: *large scale integration and interoperability*, *sustainability* and *adoption*.

Because of their intrinsic nature, any Virtual Research Environment is built as a "collection" of existing systems and resources, thus their developers have to deal with the entire stack of issues that go under the interoperability umbrella. Interoperability is actually a multi-layered and context-specific concept, which encompasses different levels along a multi-dimensional spectrum ranging from *organisational* to *semantic* and *technological* aspects. From the VRE developers' point of view it is fundamental to rely on a rich array of systems and resources – both in terms of variety and size – that can be seamlessly accessed and combined in innovative ways to satisfy the evolving needs of the community of practice. Part of the resources can be acquired and put in place from scratch for specific purposes while other resources have necessarily to be acquired from existing systems either because they are produced by those systems or for opportunistic reasons, e.g. economic ones. However, the challenges affecting Virtual Research Environments are actually very broad and include those characterising every aspect of a data infrastructure. In fact, Virtual Research Environments are at the higher level in a conceptually layered architecture of a virtual and scattered system, as they represent the application layer that is built on top of one or more layers offering at least (i) raw resources (e.g. computing, storage, network and software resources), (ii) communication and authentication protocols, (iii) protocols for publication, discovery, negotiation, monitoring, accounting and payment of resources usage, and (iv) protocols allowing the definition and management of groups of resources. In the context of a (global) research data infrastructure, the majority of these challenges are expected to be assigned to the infrastructure itself, i.e. the infrastructure should take care of put in place a rich array of mechanisms enabling interoperability with existing systems – conceptually acting as resource providers – to build a unified space of

resources – ranging from data sets, collections, storage facilities and computing power to services realising specific utilities and research objects. The richer the array of interoperability mechanisms the infrastructure is equipped with, the larger the resources space and, consequently, the domain of "VREs" that can be built.

Sustainability is definitely one of the major challenges affecting Virtual Research Environments development. VREs require effort and money to be built and maintained according to the communities of practice needs. It is a waste of effort and money building them without having a long term support, although costs can be mitigated by devising innovative development approaches eventually based on "global virtual infrastructures". As proposed in (Carusi & Reimer, 2010), there are three key strategies for sustainability which might be put in place either singly or in combinations: (i) acquire further funding from diverse research bodies; (ii) develop business models aiming at self-sustainability; and (iii) relying on community support. However, given the volatile nature of communities of practice the sustainability issue remains a challenging problem.

Although several Virtual Research Environments have been developed in various application domains and a plethora of communities of practice are in action, the majority of these systems are not yet fully integrated into standard practices, tools and research protocols used by real life communities of practice. This reluctance to migrate from traditional and consolidated research practices and facilities to the innovative ones promoted by VREs is among the most difficult barriers affecting the entire VRE domain. As recognised by (Carusi & Reimer, 2010), among the factors causing this issue there are: (i) the lack of support of both technical (e.g., bug fixing and further development of the VRE service) and instructional (e.g., training – especially in early stages) nature; (ii) the gap between the community of practice needs and the actual service implemented by the VRE; (iii) the reliability of the technology (very often VREs are based on cutting edge and evolving technologies); (iv) legal, ethical and cultural issues (the willingness to "share" research outputs and participate in web based research investigations might be nullified by fear for ownership and attribution); and (v) interdisciplinarity (differences in "languages" and working practices are a need, a potentiality and an issue as well). The lack of community uptake has cascading effects on the entire VRE research domain, in particular it impacts on the sustainability.

RECOMMENDATIONS

Virtual Research Environments represent innovative working environments that aim at enhancing the cooperation and collaboration among researchers in all modern research scenarios. They promote novel approaches and facilitate global and timely sharing of research findings, expertise and any research supporting "asset" across organizational and operational boundaries and barriers. Because of these potentialities, their development should be guided by a number of principles and best practices aiming at promoting efficiency and effectiveness of the resulting services.

A rich array of resources and systems have been developed and a lot of effort is currently spent in build-

ing infrastructures all over the world including Internet infrastructures (e.g. GÉANT[42] and Internet2[43]), grid infrastructures (e.g. European Grid Infrastructure[44] and Open Science Grid[45]), data infrastructures (e.g. DataONE[46], Data Conservancy[47], OpenAIRE[48], and D4Science[49]). Moreover, a lot of momentum has been gained by cloud technologies (Foster, Zhao, Raicu, & Lu, 2008). All these efforts should be considered as building blocks for realising Virtual Research Environments. However, to make this possible, services and resources that are aggregated and offered by such infrastructures should, as much as possible, be independent of a specific application domain and "*designed for reuse*". ***From scratch and 'self-sustained' approaches***, e.g. approaches aimed at building the entire spectrum of the needed resources without 'outside' assistance, ***should be discouraged and prevented*** because of their intrinsic development costs and difficulties to deal with evolving scenarios. Actually, Virtual Research Environments should be linked to existing infrastructures with both roles of *consumer*, i.e. VREs should benefit from the services offered by these infrastructures, and *provider*, i.e. the resources produced in the context of the VRE operation should contribute to the infrastructures offering.

Virtual Research Environments should be designed, since the beginning, to ***promote uptake, ensure usability and guarantee sustainability***. These three aspects form a virtuous circle that, if properly managed, ensure the success of a specific VRE. In reference to *uptake*, it is fundamental that the community served by the specific VRE, although virtual and aggregated by the VRE itself, is provided with tools and facilities for managing and maintaining the VRE services that have limited requirements with respect to community expertise. Moreover, the conceivers of the VRE should plan how to engage the broader community of

42 GÉANT2 is the high-bandwidth Internet serving Europe research and education community. Website www.geant2.net.

43 Internet2 is the network designed to serve the US research and education community. Website www.internet2.edu.

44 European Grid Infrastructure is building a grid infrastructure by federating a number of mainly European providers. Website www.egi.eu.

45 Open Science Grid is building a grid infrastructure by bringing together computing and storage resources from computers and research communities in the US. Website www.opensciencegrid.org.

46 DataONE is building an infrastructure for supporting Earth observational data mainly in US. Website www.dataone.org.

47 Data Conservancy is building an infrastructure promoting scientific data curation. Website dataconservancy.org.

48 OpenAIRE is building an infrastructure promoting the dissemination and sharing on open access artifacts including data. Website www.openaire.eu.

49 D4Science-II is the third project in a series dedicated to build an infrastructure that (i) serves a number of communities by providing them with diverse Virtual Research Environments and (ii) by interoperating with other infrastructures creates the core of an ecosystem of infrastructures. Website www.d4science.eu. Its enabling technology, *gCube*, will be the core technology in a forthcoming project implementing an infrastructure for the community of practice involved in Fisheries Management and Conservation of Marine Living Resources. Website www.i-marine.eu.

practice that can be served by the VRE, e.g. it might be possible to build a core team that sustains the VRE itself in the medium and long term by awareness raising, targeted training and other engagement events tailored to attract and convince key representatives of the community of practice. As regards *usability*, Virtual Research Environments building should be mainly a community building process rather than a technology development process. This implies that the focus should be primarily on using technology to identify and rationalise workflows, procedures and processes characterising a certain research scenario rather than having technology invading the research scenario and distracting effort from its real needs. As far as *sustainability* is concerned, it is fundamental that the resulting VRE service is conceived as a vital tool in the community of practice it is dedicated to. Moreover, sustainability is further enhanced whenever the VRE is perceived as a useful tool in the context of larger research initiatives and communities so to benefit from *economies of scale*, i.e. savings gained by an incremental level of production, and *economies of scope* i.e. savings gained by producing two or more distinct goods when the costs of doing so is less than that of producing each of them separately.

BIBLIOGRAPHY

[1] Allan, R. (2009). *Virtual Research Environments: From Portals to Science Gateways*. Oxford: Chandos Publishing.

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

[3] Blanke, T., Candela, L., Hedges, M., Priddy, M., & Simeoni, F. (2010). Deploying general-purpose virtual research environments for humanities research. *Phil. Trans. R. Soc. A , 368*, 3813-3828.

[4] Candela, L., Akal, F., Avancini, H., Castelli, D., Fusco, L., Guidetti, V., et al. (2007). DILIGENT: integrating Digital Library and Grid Technologies for a new Earth Observation Research Infrastructure. *International Journal on Digital Libraries , 7* (1-2), 59-80.

[5] Candela, L., Castelli, D., & Pagano, P. (2011). History, Evolution and Impact of Digital Libraries. In I. Iglezakis, T.-E. Synodinou, & S. Kapidakis, *E-Publishing and Digital Libraries: Legal and Organizational Issues* (pp. 1-30). IGI Global.

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

[7] Candela, L., Castelli, D., Pagano, P., & Simi, M. (2005). From Heterogeneous Information Spaces to Virtual Documents. *Digital Libraries: Implementing Strategies and Sharing Experiences, 8th International Conference on Asian Digital Libraries, ICADL 2005, Bangkok, Thailand, December 12-15, 2005, Proceedings*. Springer.

[8] Carusi, A., & Reimer, T. (2010). *Virtual Research Environment Collaborative Landscape Study.* JISC.

[9] Cyberinfrastructure Council. (2007). *Cyberinfrastructure Vision for the 21st Century Discovery.* National Science Foundation.

[10] Davies, S. (2011). Still Building the Memex. *Communications of the ACM , 54* (2), 80-88.

[11] De Roure, D., Goble, C., & Stevens, R. (2009). The design and realisation of the myExperiment Virtual Research Environment for social sharing of workflows. *Future Generation Computer Systems* (25), 561-567.

[12] e-Infrastructure Reflection Group. (2010). *Blue Paper.* E-IRG.

[13] Foster, I., & Kesselman, C. (1998). *The Grid: Blueprint for a New Computing Infrastructure.* Morgan Kaufmann.

[14] Foster, I., Zhao, Y., Raicu, I., & Lu, S. (2008). Cloud Computing and Grid Computing 360-Degree Compared. *In Grid Computing Environments Workshop, 2008. GCE '08* .

[15] Hey, T., Tansley, S., & Tolle, K. (2009). *The Fourth Paradigm - Data-intensive Scientific Discovery.* Microsoft Research.

[16] Lave, J., & Wenger, E. (1991). *Situated Learning: Legitimate Peripheral Participation.* New York, NY: Cambridge University Press.

[17] Llewellyn Smith, C., Borysiewicz, L., Casselton, L., Conway, G., Hassan, M., Leach, M., et al. (2011). *Knowledge, Networks and Nations: Global Scientific Collaboration in the 21st Century.* The Royal Society. The Royal Society.

[18] Nirenberg, J. (1994). From team building to community building. *National Productivity Review , 14* (1), 51-62.

[19] Nowakowski, P., Ciepiela, E., Harężlak, D., Kocot, J., Kasztelnik, M., Bartyoski, T., et al. (2011). The Collage Authoring Environment . *Procedia Computer Science , 4*, 608-617.

[20] O'Reilly, T. (2005). *What Is Web 2.0 - Design Patterns and Business Models for the Next Generation of Software.* O'Reilly.

[21] Snowdon, D. N., Churchill, E. F., & Frécon, E. (2004). *Inhabited Information Spaces: Living with your Data.* Springer.

[22] Van Gorp, P., & Mazanek, S. (2011). SHARE: a web portal for creating and sharing executable research papers. *Procedia Computer Science , 4*, 589-597.

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

[24] Wilkins-Diehr, N. (2007). Special Issue: Science Gateways - Common Community Interfaces to Grid Resources. *Concurrency and Computation: Practice and Experience , 19* (6), 743-749.

[25] Wulf, A. (1993). The collaboratory oppurtunity. *Science , 261*, 854-855.