# A Reference Architecture for Virtual Research Environments

Keith G Jeffery
Keith G Jeffery Consultants
keith.jeffery@keithgjefferyconsultants.co.uk

Carlo Meghini
ISTI - CNR, Italy
carlo.meghini@isti.cnr.it

Cesare Concordia
ISTI - CNR, Italy
cesare.concordia@isti.cnr.it

Theodore Patkos
Information Systems Laboratory, FORTH, Greece
patkos@ics.forth.gr

Valerie Brasse
euroCRIS / IS4RI, France
vbrasse@is4ri.com

Jacco van Ossenbruck
CWI, Amsterdam
Jacco.van.Ossenbrugge@cwi.nl

Yannis Marketakis
Information Systems Laboratory, FORTH, Greece
marketak@ics.forth.gr

Nikos Minadakis
Information Systems Laboratory, FORTH, Greece
minadakn@ics.forth.gr

Eda Marchetti
ISTI - CNR, Italy
eda.marchetti@cnr.it

## Abstract

This paper describes the Reference Architecture of the *enhanced VRE* (e-VRE), a Virtual Research Environment defined in the context of the VRE4EIC Project, funded by EU H2020 e-Infrastructure program. e-VRE is designed to overcome limits of existing VREs with respect to a number of orthogonal dimensions: improving the quality of VRE user experience by providing user centered, secure, privacy compliant, sustainable environments for accessing data, composing workflows and tracking data publications; increasing VRE usage in multidisciplinary research domains by abstracting and reusing building blocks and workflows from existing VRE initiatives; improving the interoperability of heterogeneous discovery, contextual and detailed metadata across all layers of the VRE.

**Keywords:** Virtual Research Environments; multidisciplinary; interoperability; innovation; collaboration; distributed systems architecture; use cases.

## Introduction

The goal of a VRE system is to decouple Science from ICT complexity, by providing researchers with a facility that takes care of ICT so allowing them to focus on their work. In this sense, a VRE is a fundamental component running on top of an e-Research Infrastructure (e-RI) (Carusi and Torseten, 2010) as it purports at making the resources of the e-RI easily accessible and reusable to the community of researchers that owns the e-RI[1]. Here, by e-RI we mean "facilities, resources and related services used by the scientific community to conduct top-level research in their respective fields"[2], while *resource* indicates any ICT entity that is of interest in an e-science community. Typically, a resource is owned by an e-RI that provides an identity for the resource and manages it, making it accessible and reusable. Examples of resources are: datasets, workflows, algorithms, Web Services, computational or storage facilities, cloud endpoints etc.

In general, a VRE is expected to:

- allow researchers to communicate with each other and to use and share the resources available in the community's e-RI;
- allow researchers to advance the state of the art by building new resources as the result of processing existing resources with the available tools. Such processing may be the application of an individual piece of software to a dataset, such as the extraction of certain knowledge from a single file; or, it may result from the execution of a complex workflow combining available services, including other workflows, on a number of data resources;
- allow research managers to apply economy of scale models to access and manage resources that researchers or single organizations alone could not afford.

Moreover, a VRE can offer all of the above on top of an individual e-RI or on top of several e-RIs, the latter option clearly requiring a level of interoperability that would empower researchers and managers in ways that are only imaginable today.

The most advanced e-RIs have developed their own VRE, showing awareness of the crucial role that a VRE can play for their researchers. Others are currently designing their VRE. However, the number of currently existing or designed VREs is very limited; more importantly, these VREs show a great heterogeneity in scope, features, underlying protocols and technologies, partially defeating the interoperability goal that lies at the very heart of a VRE.

---

[1] https://www.jisc.ac.uk/full-guide/implementing-a-virtual-research-environment-vre

[2] http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=what

VRE4EIC (*A Europe-wide Interoperable Virtual Research Environment to Empower Multidisciplinary Research Communities and Accelerate Innovation and Collaboration*) is a three-year project funded by the European Commission in the context of the H2020 Program, under the topic EINFRA-9-2015 *e-Infrastructures for virtual research environments*. One of the major goals of the VRE4EIC project is to overcome the above described issues by building an e*nhanced VRE* (e-VRE) system whose main features are:

- Increase the quality of VRE User Experiences (UX) by providing user centered, secure, privacy compliant, sustainable environments on searching data, composing workflows and tracking data publications.
- Increase the deployment of the VRE on different clusters of research infrastructures by abstracting and reusing building blocks and workflows from existing VREs, infrastructures and projects.
- Improve the contextual awareness and interoperability of the metadata across all layers of the resources in the VRE.
- Promote the exploitation and standardisation of the VRE4EIC solution to different research domains and communities.
- Provide interoperation across 'silo' e-RIs.

The main step to implement the above features is to create a Reference Architecture that can serve as a guide for the development of enhanced VREs. This paper describes the Reference Architecture produced by the VRE4EIC project after its first year of activity.

The paper is structured as follows: next Section presents related work in the same area, while the following Section outlines the methodology that has inspired our work and the way this methodology has been adapted to the present context. Next, the paper presents the vision of a VRE that lies at the basis of our Reference Architecture, presented in the following Section. The paper then concludes after briefly alluding at the further development of the Architecture within the VRE4EIC Project.

## Related Work

In designing the reference Architecture of e-VRE we have taken cognizance of past and ongoing work on:

- Science Gateways (SG). In general SGs are portals to datasets. Some have analytical, visualisation and simulation capabilities. Some provide access to –and steering of – equipment. Some provide access to specialist computing resources and some provide collaboration tools. However – in contrast to VRE4EIC - in general they are constructed on top of one or more – e-RIs in a particular domain.

- Virtual Laboratories (VL). VLs have taken a different approach, they are constructed from a pool of general software modules, available datasets and user groups. Again each VL tends to be domain specific and linked with one or a small number of e-RIs.

The booklet produced by DG-CNECT on Research Infrastructures is a useful reference resource[3]. Additionally the VRE4EIC team is carefully checking and tracking activities related to other EU projects on VRE. Indeed, the VRE4EIC project is one of the four H2020 RIA projects concerned with Virtual Research Environments; other projects are: EVER-EST[4] (geoscience); BlueBridge[5] (dominantly marine) and West-LIFE[6] (Bio). All projects are currently at early stage of development, so our considerations have relied on (a) information from the project websites (b) personal contacts especially with EVER-EST and Blue Bridge (where the major partner is the same organization, but a different group, as the architecture developers in VRE4EIC). The significant differences in approaches are:

1. VRE4EIC is producing a reference architecture (and prototype demonstrator) that can bridge across e-RIs (and hence underlying e-Is) in a multidisciplinary manner; the other projects are restricted to particular domains;
2. BlueBridge produces a VRE that is tightly coupled to the underlying e-RIs;
3. EVER-EST is using research objects; this binds data and code in a particular way that restricts openness and interoperability, which is unacceptable for the objectives of VRE4EIC. Nonetheless we are working together and believe there are opportunities for co-development.
4. WEST-LIFE is built in context of the European Grid Initiative and its modularized components for self-assembly by the e-RIs to form a VRE.

Outside of these H2020 projects for VRE4EIC participants, the scientific coordinator has initiated a RDA IG (Research Data Alliance Interest Group) jointly with EVER-EST.

[3] https://ec.europa.eu/digital-single-market/en/news/e-infrastructures-making-europe-best-place-research-and-innovation
[4] http://www.ever-est.eu
[5] http://www.bluebridge-vres.eu
[6] https://portal.west-life.eu

## Methodology and approach

In the development of the Reference Architecture, an incremental software development process has been adopted, largely inspired by the RUP process[7] (Jacobson, Booch, Rumbaugh & Booc 1999). Architectural components are derived based on an analysis of the requirements collected in the project. Fig. below presents a UML class diagram outlining the entities involved in this analysis and their relationships. In particular:

- The analysis started from the Requirements (yellow box in Figure). Each requirement has been considered individually, and the functions (green box) required for its implementation have been derived.
- In order to ease the specification of functions, a set of generalised functions has also been derived, which are included or specialised by functions, or which may be used as preconditions by functions.
- The components that are required for the implementation of functions have finally been derived .

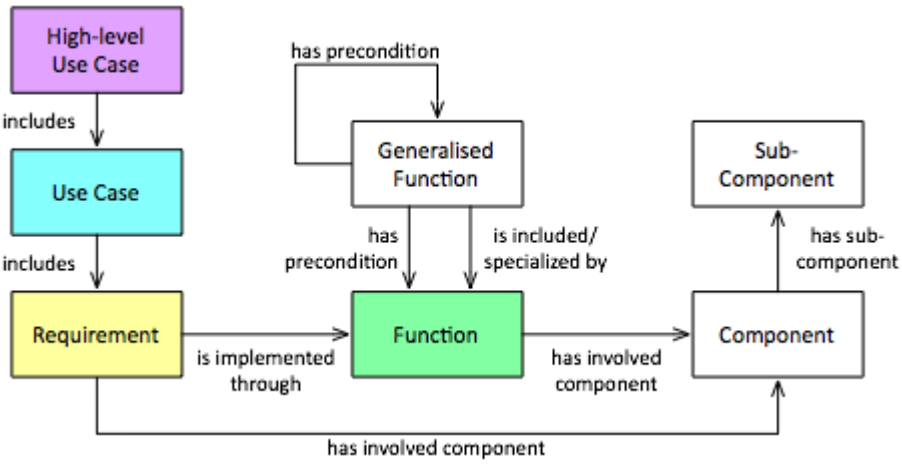


Figure 1 Entities and relationships involved in the analysis of requirements

As the figure also shows, this analysis of requirements into functions and components connects to the use cases (turquoise box) and the high-level use cases (purple box). The connection is realized through requirements, and is used for the assessment of the Reference Architecture.

[7]https://www.ibm.com/developerworks/rational/library/content/03July/1000/1251/1251_bestpractices_TP026B.pdf

Overall, the schema shown in Figure above allows us to maintain the relationship between use cases, requirements and components, thereby realizing the traceability of the Reference Architecture.

## A Vision for the e-VRE Architecture

In order to enable a VRE to make its resources available to the researchers that use the VRE, each e-RI that *participates* to a VRE must provide *descriptions* of its resources, and such descriptions must be rich enough in information to support the VRE services. This information may include the protocol that must be used to interact with an e-RI service; the schema, size and operations allowed on a e-RI dataset; the permission framework that must be adopted for authentication/authorization of the e-RI users, and so on. This process is naturally divided into two steps, as depicted by Figure 2: the e-RI resources are given at the bottom level since they are the basic assets that both e-RIs and VREs operate on; at the next level up, the descriptions of these resources used by the e-RI services (*e-RI Resource descriptions*) are given, next to the e-RI services using them; at the top level, the descriptions of the e-RI resources used by the VRE services (*VRE Resource descriptions*) are given, next to the VRE services using them. In both cases, the services mentioned are purely exemplificative.

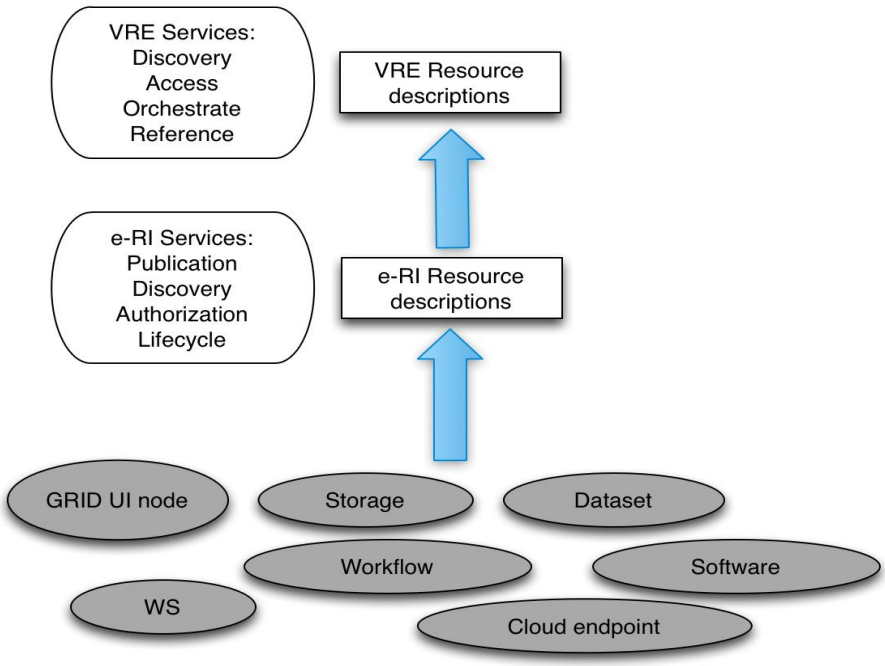


Figure 2 Resources, e-RIs Services and VRE Services

Resource descriptions are typically collected in *Catalogues*. So, the VRE needs to access the catalogues of the participating e-RIs in order to discover the existing resources in the e-RI and obtain enough information on these resources to create its own descriptions of them in its own catalogue. In order to simplify the creation of the VRE catalogue, one could employ the same data model for both the e-RI and the VRE descriptions. In fact, VREs that live within e-RIs follow this approach. In this case, e-RIs and VRE Resource descriptions only differ for the type of information they contain, while sharing the identity and the basic attributes of resource descriptions. However, this approach is not feasible for VREs with *many* participating e-RIs, due to the fact that in general different e-RIs use different data models to structure their catalogues. In this case, there are two main approaches to create and maintain the VRE catalogue:

- The *centralized* approach (see Figure 3, left), in which there exists a VRE Catalogue used by the VRE services for carrying out their own operations.
- The *distributed* approach (see Figure 3, right), in which there is no VRE Catalogue, but the VRE accesses the e-RIs catalogues when the information is needed.

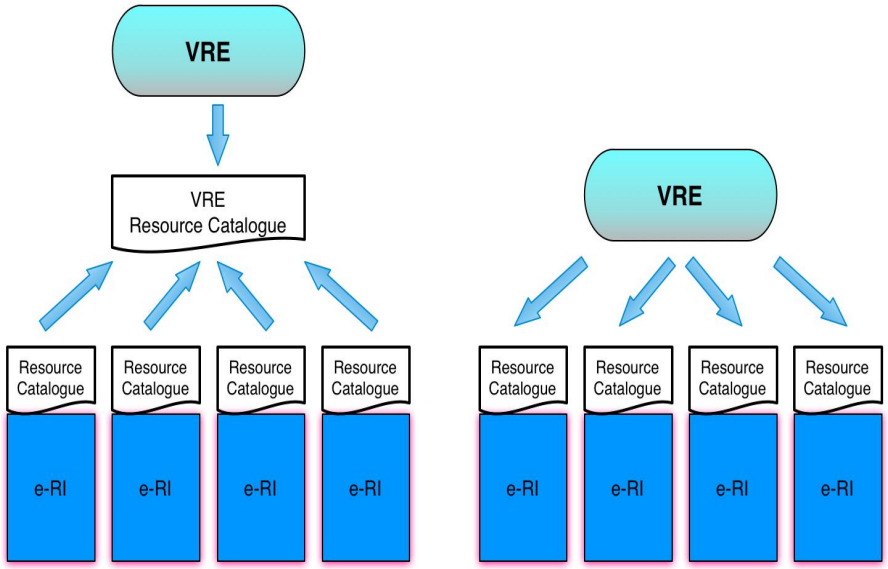


Figure 3 Alternative approaches to the cooperation on e-RIs and a VRE

Each approach has its own pros and cons, as it is well known in distributed system design. In fact, the availability of a VRE Catalogue facilitates all VRE operations that rely exclusively on resource descriptions, such as

resource discovery. For operations that require data access, such as data discovery, the centralized approach can only alleviate the problem, by offering information for executing part of the operation. On the other hand, the distributed approach makes it easier to have complete information in real time, since it does not require propagation of updates to the Catalogue.

Our Reference Architecture chooses the centralized approach, because it facilitates one important service, namely the construction of workflows across one or more RIs. The construction of workflows requires numerous access to resource descriptions, followed by optimisation for parallel/distributed operations; the centralized approach makes it possible to implement this access in the most efficient way possible.

## A Reference Architecture for e-VRE

At a more general level, the Reference Architecture conforms to the multi-tiers view paradigm used in the design of distributed information systems. Following this paradigm, we can individuate three logical tiers in e-VRE, as shown in Figure 4:

- The *Application* tier, which provides functionalities to manage the system, to operate on it, and to *expand* it, by enabling administrators to plug new tools and services into the e-VRE.
- The *Interoperability* tier, which deals with interoperability aspects by providing functionalities for: i) enabling application components to discover, access and use e-VRE resources independently from their location, data model and interaction protocol; ii) publishing e-VRE functionalities via a Web Service API; and iii) enabling e-VRE applications to interact each others.
- The *Resource Access* tier, which implements functionalities that enable e-VRE components to interact with eRIs resources. It provides synchronous and asynchronous communication facilities.

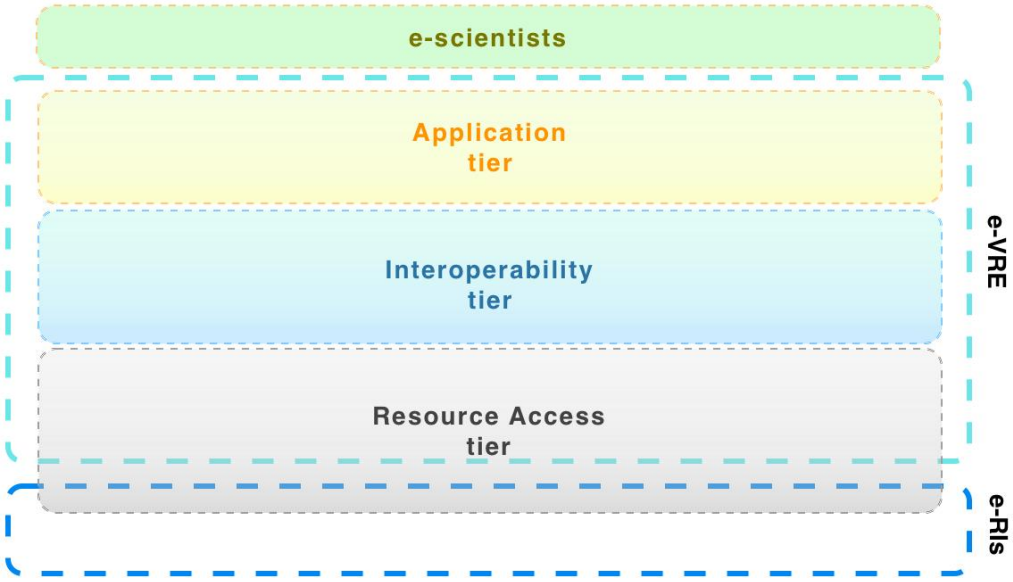


Figure 4 Architectural tiers in a VRE

Figure 4 depicts the logical tiers of e-VRE and shows their placement in an ideal space between the e-scientists that use the e-VRE and the e-RIs that provide the basic resources to the e-VRE.

Generally speaking a VRE system can be viewed as a dynamic framework; it "is the result of joining together new and existing components to support as much of the research process as appropriate for any given activity or role"[8]. To implement this fundamental non-functional requirement the e-VRE system has been designed following a *component-oriented* approach.

According to this approach a system is composed by an integration infrastructure where a set of software components can be deployed, these components implement the system functionalities and potentially can be specified, developed and deployed independently of one another.

Based on these considerations and on the analysis of the requirements, for the basic integration infrastructure of e-VRE we have individuated a set of basic functionalities grouped into six *conceptual components*:

- The e-VRE management is implemented in the **System Manager** component. The System Manager can be viewed as the component enabling Users to use the *core* functionalities of the e-VRE: access, create and manage resource descriptions, query the e-VRE

[8]Fraser M. "Virtual Research Environments: Overview and Activity" 30-July-2005, http://www.ariadne.ac.uk/issue44/fraser/

information space, configure the e-VRE, plug and deploy new tools in the e-VRE and more.

- The **Workflow Manager** enables users to create, execute and store business processes and scientific workflows.
- The **Linked Data (LD) Manager** is the component that uses the LOD (Linked Open Data) paradigm, based on the RDF (Resource Description Framework) data model, to publish the e-VRE information space - i.e. the metadata concerning the e-VRE and the e-RIs in a form suitable for end-user browsing in a SM (Semantic Web)-enabled ecosystem.
- The **Metadata Manager** (MM) is the component responsible for storing and managing resource catalogues, user profiles, provenance information, preservation metadata used by all the components. All these entities and their relations are captured in the CERIF data model [9] (Jeffery et al., 2002), using extended entity-relational conceptual and object-relational logical representation for efficiency.
- The **Interoperability Manager** provides functionalities to implement interactions with e-RIs resources in a transparent way. It can be viewed as the interface of e-VRE towards e-RIs. It implements services and algorithms to enable e-VRE to: communicate synchronously or asynchronously with e-RIs resources, query the e-RIs catalogues and storages, map the data models.
- The **Authentication, Authorization, Accounting Infrastructure (AAAI)** component is the responsible for managing the security issues of the e-VRE system. It provides user authentication for the VRE and connected e-RIs, authorisation and accounting services, and data encryption layers for components that are accessible over potentially insecure networks.

Figure 5 shows how these six components are distributed on the 3-tier space introduced above.

[9] http://eurocris.org/ontologies/semcerif/1.3/

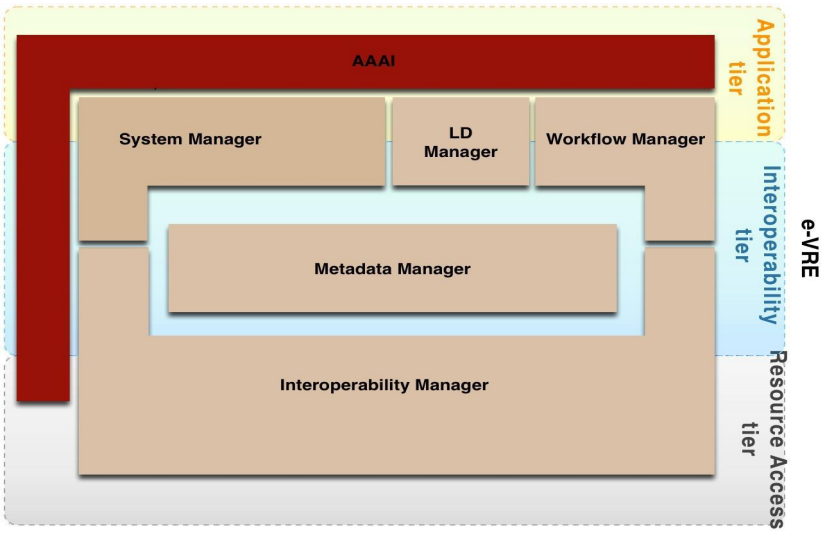


Figure 5 Conceptual components and logical tiers

## Outlook

A first version of the Reference Architecture has been released at the end of the first year of the project, that is end of September 2016. The Reference Architecture has been provided in terms of two kinds of UML diagrams: a component diagram highlighting the components of the architecture, their provided and used interfaces; and a number of sequence diagrams highlighting the interactions occurring in the execution of the main methods. A revised version of the first release will be produced by the end of the second year of the project (end of September 2017), for internal usage only. This revision has been deemed as necessary by the project management as a consequence of the fact that the list of user requirements will be in turn refined three times during the second year of the project. These refinements have been planned in order to accommodate the collection of a large amount of requirements, along with the on-going characterization of existing e-RIs. Each time the requirements and the e-RI characterizations will be updated, the Reference Architecture will consequently be revised.

In parallel to the revision of the architecture, a Gap Analysis work will be performed, to determine the components that will be implemented. The development process will rely on the re-use of existing technologies and standards, which in turn may lead to the revision of some interfaces of the Reference Architecture, for instance to align an interface with the selected standard or technology; this alignment may be propagated into the Reference Architecture, if the standard or technology provoking it are important enough.

## Conclusions

The Reference Architecture for Virtual Research Environments developed in the context of the VRE4EIC Project has been presented, based on a vision of VREs and of the relationships between VREs and their close relatives, i.e. e-Research Infrastructures, e-Infrastructures, Science Gateways and Virtual Laboratories. The Architecture has been described in terms of its main components, organized in a three-layered approach that closely resembles classic three-tiers architectures. These components have been derived by analysing a large set of requirements, collected in more than 40 interviews, and by characterising 5 existing e-Research Infrastructures [D2.1]. A complete specification of the Reference Architecture, as well as a detailed documentation of the process that led to its derivation, can be found in the deliverable D3.1 of the VRE4EIC Project [D3.1], which presents the main UML component diagram of the architecture, highlighting the interfaces provided and used by each component. Such interfaces are further specified in terms of the signatures of the methods that are included in each of them. UML sequence diagrams are also provided for the main methods and use cases. The flow of work within the VRE4EIC Project that will lead to the refinement, validation, and partial implementation of the Reference Architecture has also been presented, putting in context the present Architecture and indicating its evolution in the next two years of work.

## Acknowledgements

The research leading to these results has received funding from the European Union Horizon 2020 Programme, Topic: e-Infrastructures for virtual research environments, Research and Innovation action VRE4EIC (A Europe-wide Interoperable Virtual Research Environment to Empower Multidisciplinary Research Communities and Accelerate Innovation and Collaboration), grant agreement n 676247.

## References

Jacobson, I., Booch, G., Rumbaugh, J. and Booch, G. (1999). The unified software development process (Vol. 1). Reading: Addison-Wesley.

Carusi , Annamaria and Reimer , Torseten (2010). Virtual Research Environment .Collaborative Landscape Study. http://www.webarchive.org.uk/wayback/archive/20140614205957/http://www.jisc.ac.uk/publications/reports/2010/vrelandscapestudy.aspx

Jeffery , Keith G, Lopatenko, Andrei, and Asserson, Anne (2002). Comparative Study of Metadata for Scientific Information: The place of CERIF in CRISs and Scientific Repositories. 6th International Conference on Current Research Information Systems, Kassel, Germany.