PROCEEDINGS OF SCIENCE

The CHAIN-REDS Knowledge Base and Semantic Search Engine

Roberto Barbera[1]
Department of Physics and Astronomy of the University of Catania and INFN
Via S. Sofia 64, 95123 Catania, Italy
E-mail: roberto.barbera@ct.infn.it

Carla Carrubba
INFN Division of Catania
Via S. Sofia 64, 95123 Catania, Italy
E-mail: carla.carrubba@ct.infn.it

Giuseppina Inserra
INFN Division of Catania
Via S. Sofia 64, 95123 Catania, Italy
E-mail: giuseppina.inserra@ct.infn.it

Rita Ricceri
INFN Division of Catania
Via S. Sofia 64, 95123 Catania, Italy
E-mail: rita.ricceri@ct.infn.it

e-Infrastructures, and in particular Open Access Data Infrastructures, are essential platforms for e-Science and e-Research and are being built since several years both in Europe and the rest of the world to support diverse multi/inter-disciplinary Virtual Research Communities. So far, however, it is difficult for scientists to correlate papers to datasets used to produce them and to discover data and documents in an easy way. In this paper we present the CHAIN-REDS project's Knowledge Base and we introduce its Semantic Search Engine which attempts to address those drawbacks and contribute to the reproducibility of science.

1 Speaker

PoS(eIeS2013)016

PROCEEDINGS OF SCIENCE

e-Infrastructures for e-Sciences 2013 A CHAIN-REDS Workshop organised under the aegis of the European Commission (eIeS 2013)
October 22, 2013
Beijing, P.R. of China

PoS(eIeS2013)016

1. Introduction

In the last 30 years or so, scientific computing has steadily evolved from mainframe-based centralized solutions to a really distributed environment. This has been possible thanks to the concurrent availability of powerful "Commercial Of The Shelf" (COTS) computers and decrease of costs of Local Area Networks. In the first half of 90's the emergence of cluster computing for High Throughput Computing (HTC) applications was confirmed and "farms" of computers with many-core processors, interconnected by very low latency networks, have become the norm also in the domain of High Performance Computing (HPC) at a point that in the last five years about 80% of the Top500 machines are based on a distributed architecture.

Furthermore, the steep decrease of costs of large/huge-bandwidth Wide Area Net-works has fostered in the recent years the spread and the uptake of the Grid Compu-ting paradigm and the distributed computing ecosystem has become even more com-plex with the recent emergence of Cloud Computing.

Indeed, e-Infrastructures are being built since several years both in Europe and the rest of the world to support diverse multi/inter-disciplinary Virtual Research Commu-nities (VRCs) [1] and a shared vision for 2020 is that e-Infrastructures will allow scientists across the world to do better (and faster) research, independently of where they are deployed and of the paradigm(s) adopted to build them.

E-Infrastructure components can be key platforms to support the Scientific Method [2], the "knowledge path" followed every day by scientists since Galileo Galilei, in many aspects. Distributed Computing and Storage Infra-structures (local HPC/HTC resources, Grids, Clouds, long term data preservation services) are ideal both for the creation of new datasets and the analysis of existing ones while Data Infrastructures (including Open Access Document Repositories – OADRs – and Data Repositories – DRs) are essential also to evaluate existing data and annotate them with results of the analysis of new data produced by experiments and/or simulations. Last but not least, Semantic Web based enrichment of data is key to correlate document and data, allowing scientists to discover new knowledge in an easy way.

2. The CHAIN-REDS Knowledge Base

CHAIN-REDS [3] is a project co-funded by the European Commission within its Seventh Framework Program. CHAIN-REDS started on the 1st of December 2012 and will last for 30 months. The project consortium [4] is made of nine renowned organisations in the field of e-Infrastructures, representing Europe and the following world regions: i) China, ii) India, iii) Latin America and the Caribbean, iv) Mediterranean, Middle-Eastern and Gulf Region Arab Countries, and v) Sub-Saharan Africa.

CHAIN-REDS vision is to promote and support technological and scientific collaboration across different e-Infrastructures established and operated in various continents in order to facilitate their uptake and use by established and emerging VRCs but also by single researchers, promoting instruments and practices that can facilitate their inclusion in the global e-Science and e-Research.

In order to reach its objectives, CHAIN-REDS has devised a work plan based on four pillars: Awareness, Information, Access and Inclusion which are deemed key to reach the long term sustainability of e-Infrastructures.

In the present paper we will mainly deal with Information which constitutes the first part of the Scientific Method.

In order to "inform" specialised researchers, "citizen scientists" and the general public about existing e-Infrastructure sites, services and applications as well as open access documents and freely-accessible data available on Data Infrastructures relying on those e-Infrastructures, CHAIN-REDS (and CHAIN [5], its predecessor) has built a knowledge base The CHAIN-REDS Knowledge Base [6] is one of the largest existing e-Infrastructure-related digital information systems. It currently contains information, gathered both from dedicated surveys and other web and documental sources, for largely more than half of the countries in the world.

Information is presented to visitors through geographic maps and tables. Users can choose a continent in the map and, for each country where a marker is displayed, get the information about the Regional Research & Education Network(s) and the Grid Regional Operation Centre(s) (ROCs) the country belongs to as well as the National Research & Education Network, the National Grid Initiative, the Certification Authority, and the Identity Federation available in the country, down to the Grid site(s) running in the country and the scientific application(s) developed by researchers of the country and running on those sites.

Besides e-Infrastructure sites, services and applications, the CHAIN-REDS Knowledge Base publishes information about Open Access Document Repositories and Data Repositories. The OADR site view is shown in Figure 1.

Figure 1: The CHAIN-REDS Knowledge Base: the OADR site view.

Red markers in the map correspond to the almost 2,500 repositories of DRIVER [7], OpenAIRE [8] and OpenDOAR [9] while yellow ones refer to the new repositories that have been added thanks to the CHAIN-REDS outreach activity. Clicking on a marker, one gets the direct link to the corresponding repository in order to search inside it. Globally, the CHAIN-REDS Knowledge Base implicitly contains links to more than 33 million documents.

The DR site view is shown in Figure 2.

Figure 2: The CHAIN-REDS Knowledge Base: the DR site view.

Red markers in the map correspond to the location of at least one of the organizations owning the more than 500 Data Repositories included in Databib [10] and DataCite [11]. Yellow markers refer to other DRs added thanks to the outreach work done by CHAIN-REDS. Clicking on a marker, one gets the direct link(s) to the corresponding repository(ies) in order to search inside it(them).

3. The Semantic Search Engine

3.1 Generalities

Although it is quite useful to have a central access point to thousands of repositories and millions of documents and datasets, with both geographic and tabular infor-mation, the OADR and DR part of the CHAIN-REDS Knowledge Base is only a demonstrator with limited impact on scientists' day-by-day life. In order to find a document or a dataset, users should know beforehand what they are looking for and there is no way to correlate documents and data which would actually be of the most important facilitators of the Scientific Method.

In order to overcome these limitations and turn the Knowledge Base into a powerful research tool, the CHAIN-REDS consortium has decided to semantically enrich OADRs and

DRs and build a search engine on the related linked data. The architecture and the current implementation of the CHAIN-REDS Semantic Search Engine [12] are presented in Section 3.2 and 3.3, respectively.

3.2 Architecture

The multi-layered architecture of the search engine is sketched in Figure 3 where both the official and the "de facto" Semantic Web standards and technologies [13] adopted are described by their corresponding logos.

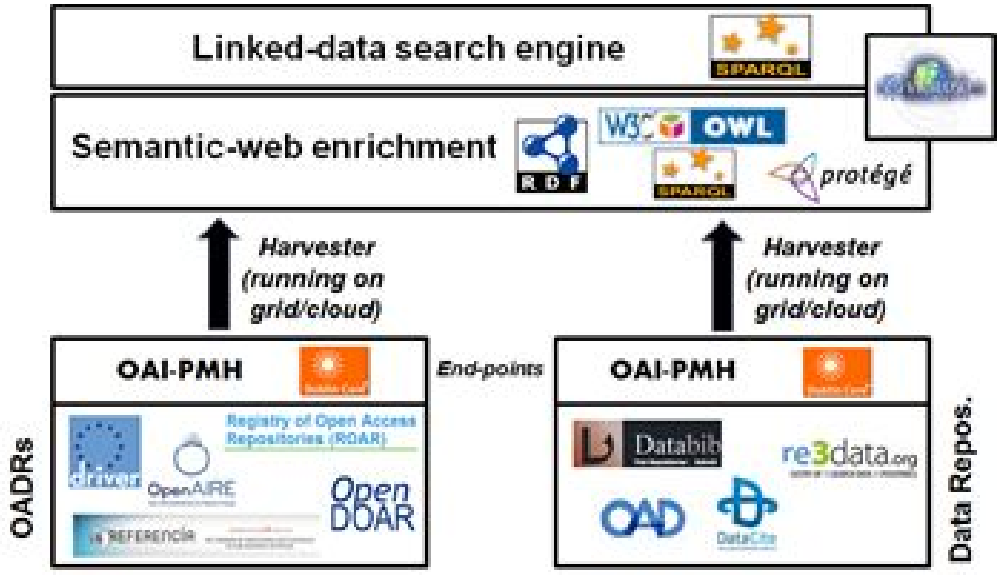

Figure 3: Architecture of the Semantic Search Engine.

Starting from the bottom of Figure 3, the first two components of the service are described below.

1. Metadata Harvester

As shown in Fig. 3, the metadata harvester is a process running either on a Grid or a Cloud infrastructure which consists of the following parts:

(a) Get the address of each repository publishing an OAI-PMH standard [14] endpoint;
(b) Retrieve, using the OAI-PMH repository address, the related Dublin Core [15] encoded metadata in XML format;
(c) Get the records from the XML files and, using the Apache Jena API [16], transform the metadata in RDF format;
(d) Save the RDF files into a Virtuoso [17] triple store according to an OWL-compliant ontology built using Protégé [18].

2. Semantic-Web Enricher

Each RDF file retrieved and saved in the Virtuoso triple store is mapped onto a Virtuoso Graph that contains the ontology expressly developed for the search engine.
The ontology, built using the Dublin Core and the FOAF standards, consists of:

- Classes that describe the general concepts of the domain: Resource, Author, Organisation, Repository and Dataset (where Resource is a given open access document);
- Object properties that describe the relationships among the ontology classes; the ontology developed for the service described in this paper has several specific properties such as *hasAuthor* (i.e., the relation between Resources and Authors) and *hasDataSe*t (i.e., the relation between Resources and Datasets);
- Data properties (or attributes) that contain the characteristics or classes' parameters.

The third, and highest-level, component is the Search Engine itself which is described in the next sub-section.

3.3 Implementation

The home page of the CHAIN-REDS Semantic Search Engine is reachable at the URL: www.chain-project.eu/semantic-search. Visitors can either enter a keyword and submit a SPARQL query to the Virtuoso triple store or select a language and get, on the left side of the page, the list of subjects available in that language with the indication, between parentheses, of the number of records available for that particular subject. So far, more than 22 million resources are searchable by the CHAIN-REDS Semantic Search Engine and its Virtuoso RDF database contains almost 600 million triples.

The results of a given query are listed in a summary view. For each record found, the title, the author(s) and a short description of the corresponding resource are provided. Clicking on the "More Info" link, visitors can also access a detailed view of the resource. In the "Dataset information" panel users get the link to the open access document and, if existing, to the corresponding dataset.

4. Summary and conclusions

Distributed Computing and Storage Infrastructures and Data Infrastructures are essential components of e-Infrastructures to support the application of the Scientific Method in the 21st-century researchers' day-by-day work.

The CHAIN-REDS Knowledge Base and its Semantic Search Engine have been conceived to demonstrate the potential of information coupled with semantic web technologies to address the issues of data discovery and correlation. The next step, with reference to the considerations made at the beginning of Section 2, is now to move from "Information" to "Inclusion" and identify/create new OADRs and DRs in those regions addressed by CHAIN-REDS to be included in the Knowledge Base and made available in the Search Engine to support several different Virtual Research Communities. Future developments also include the possibility i) to enrich the information contained in the CHAIN-REDS Knowledge Base with that included both in general-purpose (e.g., DBpedia [19] or Google Scholar [20]) and domain-specific semantic repositories (e.g., PubMed [21] or CIARD R.I.N.G. [22]), and ii) to allow the Search Engine to work on domain-specific sets of the semantic-web-enriched data.

References

[1] G. Andronico et al, *E-Infrastructures for International Cooperation*, in *Computational and Data Grids: Principles, Applications and Design* (N. Preve Editor), IGI Global 2011, DOI:

10.4018/978-1-61350-113-9; see also www.igi-global.com/book/computational-data-grids/51946.

[2] There are many equivalent definitions and depictions of the Scientific Method, both on the web and on textbooks. In this paper we refer to http://home.badc.rl.ac.uk/lawrence/blog/2009/04/16/scientific_method.

[3] The home page of the CHAIN-REDS project can be found at www.chain-project.eu (last time visited: October 2013).

[4] The list of CHAIN-REDS partners can be inspected at www.chain-project.eu/partners (last time visited: October 2013).

[5] The home page of the CHAIN project can be found at www.chain-project.eu/web/old-project (website frozen on the 30th of November 2012 and not any more subject to change).

[6] The CHAIN-REDS Knowledge Base can be browsed at www.chain-project.eu/knowledge-base (last time visited: October 2013).

[7] The home page of the DRIVER project can be found at www.driver-repository.eu (last time visited: October 2013).

[8] The home page of the OpenAIRE project can be found at www.openaire.eu (last time visited: October 2013).

[9] The home page of the OpenDOAR initiative can be found at www.opendoar.org (last time visited: October 2013).

[10] The home page of the Databib initiative can be found at www.databib.org (last time visited: October 2013).

[11] The home page of the DataCite initiative can be found at www.datacite.org (last time visited: October 2013).

[12] The CHAIN-REDS Search Engine on Linked Data can be accessed at www.chain-project.eu/semantic-search (last time visited: October 2013).

[13] The Semantic Web standards can be inspected at http://semanticweb.org/wiki/Semantic_Web_standards (last time visited: October 2013).

[14] The OAI-PMH standard home page can be found at www.openarchives.org/pmh (last time visited: October 2013).

[15] The Dublin Core Metadata Initiative home page can be found at www.dublincore.org (last time visited: October 2013).

[16] The Apache Jena API home page can be found at http://jena.apache.org (last time visited: October 2013).

[17] The Virtuoso home page can be found at http://virtuoso.openlinksw.com (last time visited: October 2013).

[18] The Protégé home page can be found at http://protege.standford.edu (last time visited: October 2013).

[19] The home page of DBpedia can be found at www.dbpedia.org (last time visited: October 2013).

[20] The home page of Google Scholar can be found at http://scholar.google.com (last time visited: October 2013).

[21] The home page of PubMed can be found at www.ncbi.nlm.nih.gov/pubmed (last time visited: October 2013).

[22] The home page of CIARD R.I.N.G. can be found at http://ring.ciard.net (last time visited: October 2013).