

The CHAIN-REDS Semantic Search Engine

Roberto BARBERA^{1,2}, Carla CARRUBBA², Giuseppina INSERRA²,

Rita RICCERI², Rafael MAYO-GARCÍA³

(On behalf of the CHAIN-REDS project)

*¹Department of Physics and Astronomy of the University of Catania,
Viale A. Doria, 6, 95125 Catania, Italy*

Email: roberto.barbera@ct.infn.it

*²Italian National Institute of Nuclear Physics, Division of Catania,
Via S. Sofia, 64, 95123 Catania, Italy*

Email: {carla.carrubba, giuseppina.inserra, rita.ricceri}@ct.infn.it

*³Centro de Investigaciones Energéticas Medioambientales y Tecnológicas (CIEMAT)
Avenida Complutense, 40, 28040 Madrid, Spain*

Email: rafael.mayo@ciemat.es

Abstract

e-Infrastructures, and in particular Data Repositories and 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, the CHAIN-REDS project's Knowledge Base and its Semantic Search Engine are presented, which attempt to address those drawbacks and contribute to the reproducibility of science.

Keywords

CHAIN-REDS, Cloud Computing, Data Repositories, e-Infrastructures, Grid Computing, Linked Data, Search Engines, Semantic Web.

1. Introduction

In the last 30 years or so, scientific computing has steadily evolved from mainframe-based centralized solutions to a really distributed environment (see Fig. 1). 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 Networks has fostered in the recent years the spread and the uptake of the Grid Computing paradigm and the distributed computing ecosystem has become even more complex with the recent emergence of Cloud Computing.

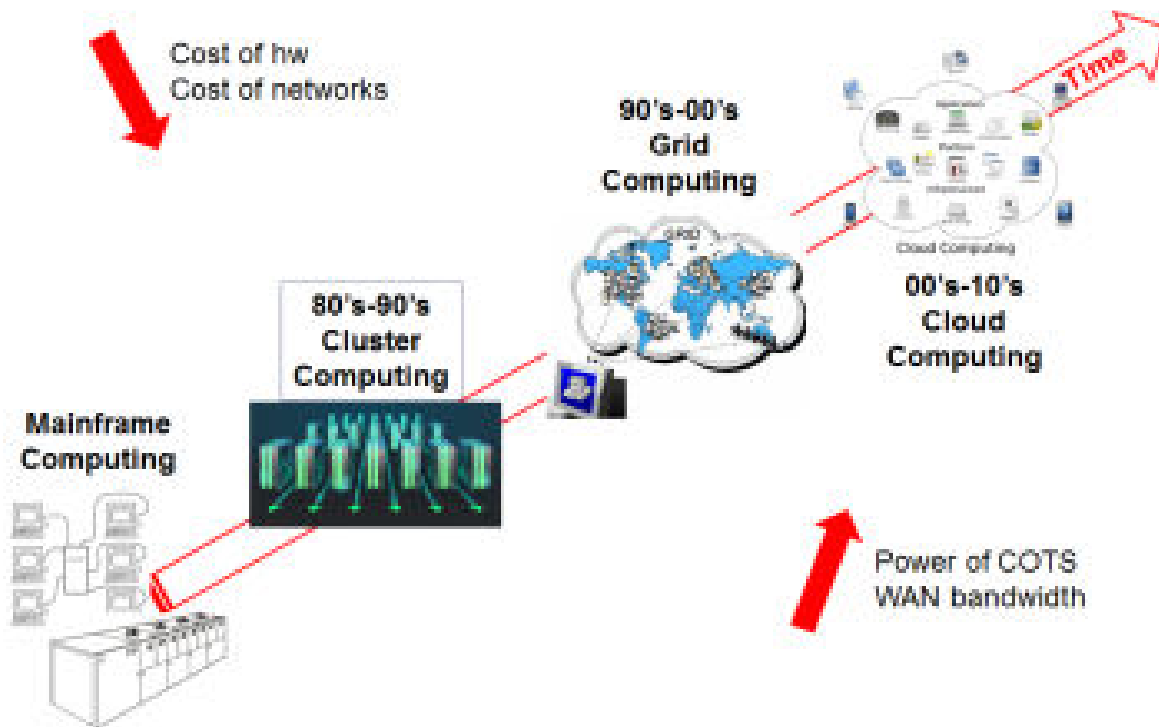


Figure 1 Evolution of scientific computing.

All these developments have triggered, at the onset of the 21st century, the new concept of e-Infrastructure (also called cyber-infrastructure, especially in the US) defined as “an environment where research resources (hardware, software and content) can be readily shared and accessed where necessary to promote better and more effective research; such environment integrate hard-, soft- and middleware components, networks, data repositories, and all sorts of support enabling virtual research collaborations to flourish globally” [1].

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 Communities (VRCs) (Andronico, G. *et al*, 2011) 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. With reference to Fig. 2, Distributed Computing and Storage Infrastructures (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

2.1 The CHAIN-REDS Project

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.



Figure 2 The Scientific Method (the figure originally comes from [2]).

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.

The specific objectives of the project are to:

- Obj1. Extend and consolidate the international cooperation of Europe with other regions of the world in the domain of e-Infrastructures for Research & Education (R&E), reinforcing the scientific collaboration and broadening the impact of the European Research Area.
- Obj2. Promote, coordinate and support the effort of a critical mass of non-European e-Infrastructures for R&E to collaborate with Europe addressing interoperability and interoperation of Grids and other Distributed Computing Infrastructures such as potential upcoming of Cloud federations and High Performance Computing (HPC) centres.
- Obj3. Study the opportunities of data sharing across different e-Infrastructures and continents widening the scope of the Knowledge Base (see next subsection) to Data Infrastructures and Cloud implementations.
- Obj4. Promote trust building towards open scientific Data Infrastructures across the world regions, including organizational, operational and technical aspects.
- Obj5. Demonstrate the relevance of intercontinental cooperation in several scientific data fields addressing existing and emerging VRCs and propose pragmatic approaches that could impact the everyday work of the single researchers, even if not structured in a VRC.
- Obj6. Provide guidance and recommendations for roadmaps for long-term global collaboration in e-Infrastructures and harmonization of existing policies. These are envisaged to act as input to policy- and decision-making mechanism, harmonized with the European Digital Agenda and Horizon 2020.

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. The four lines of action are organised in a virtuous cycle, as sketched in Fig. 3.



Figure 3 The CHAIN-REDS virtuous cycle.

In the present paper we will mainly deal with Information which mostly relates to objectives Obj3 and Obj4 listed above and, as shown in Fig. 2, 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 which is described in detail in the following sub-section.

2.2 The 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. The “country view” is shown in Fig. 4. 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.



Figure 4 The CHAIN-REDS Knowledge Base: the country view.

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 Fig. 5.

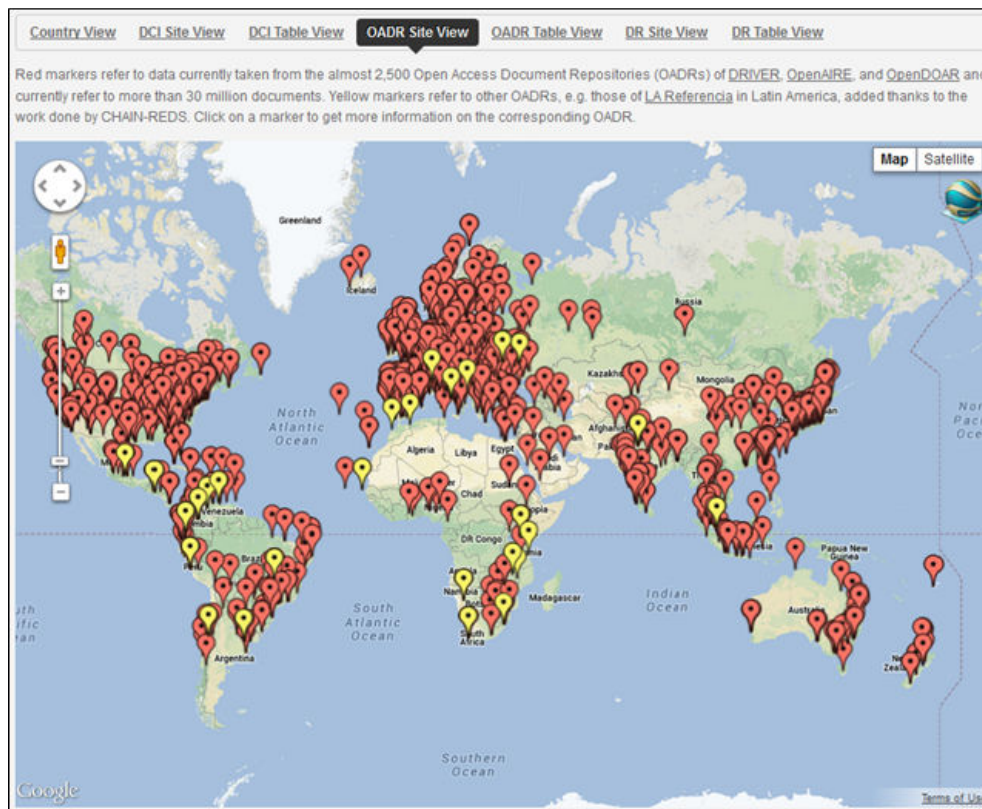


Figure 5 The CHAIN-REDS Knowledge Base: the OADR site view.

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 Fig. 6.

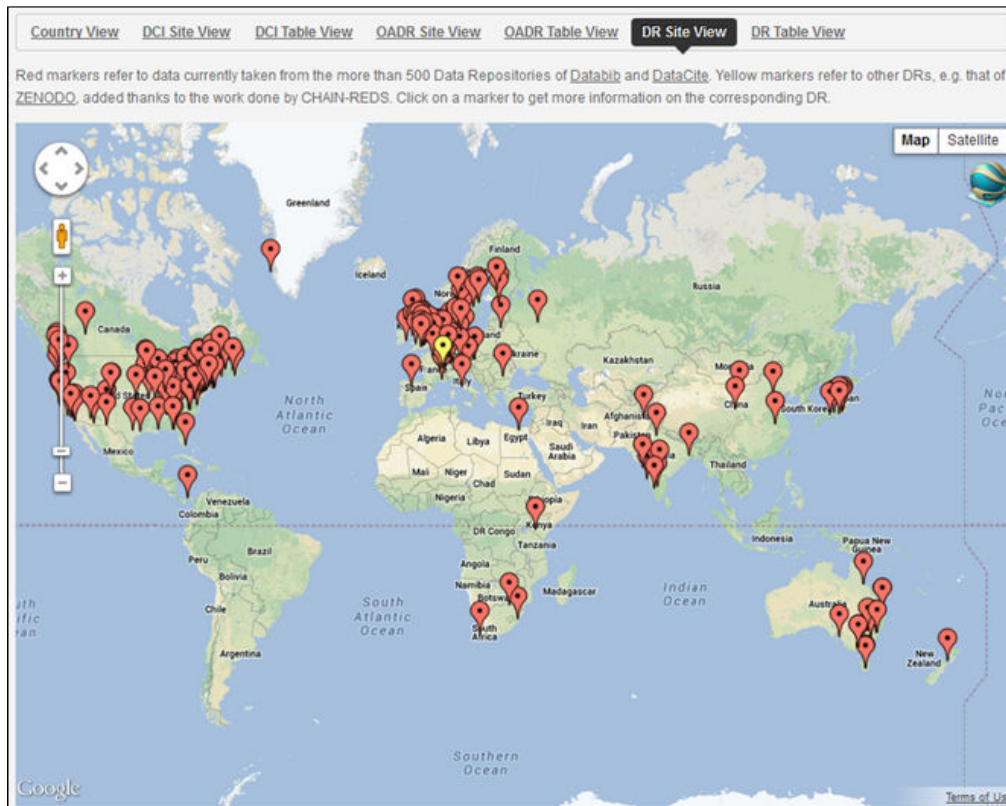


Figure 6 The CHAIN-REDS Knowledge Base: the DR site view.

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]. Clicking on a marker, one gets the direct link(s) to the corresponding repository(ies) in order to search inside it(them).

3. The CHAIN-REDS Knowledge Base

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 information, 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 (see Fig. 2).

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 OADR 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 Fig. 7 where both the official and “de facto” Semantic Web standards and technologies [13] adopted are described by small logos.

Starting from the bottom of Fig. 7, the first two components of the service are described below.

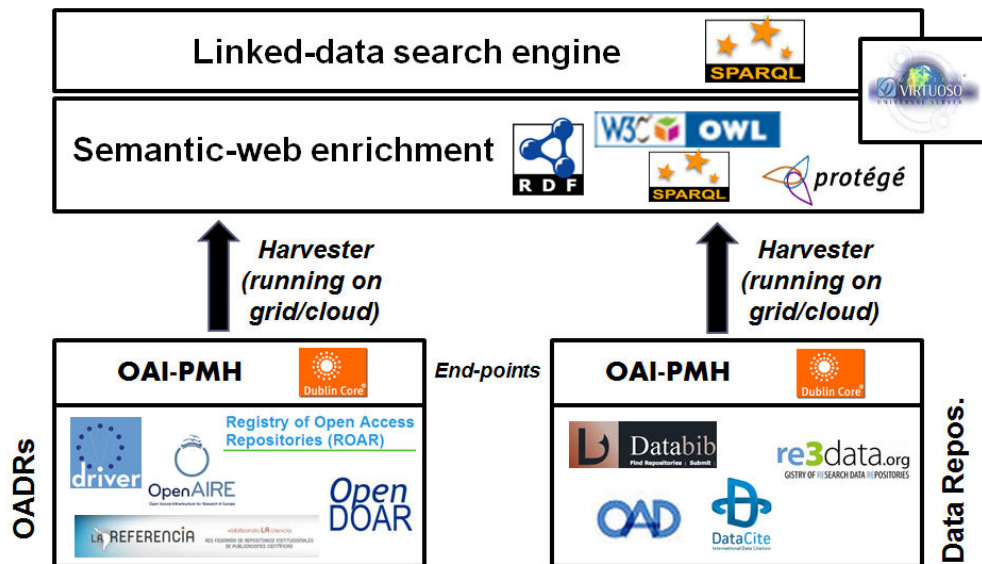


Figure 7 Architecture of the Semantic Search Engine.

3.2.1 Metadata Harvester

As shown in Fig. 7, the metadata harvester is a process running either on a Grid or a Cloud infrastructure which consists of the following parts:

- Get the address of each repository publishing an OAI-PMH standard [14] endpoint;
- 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].

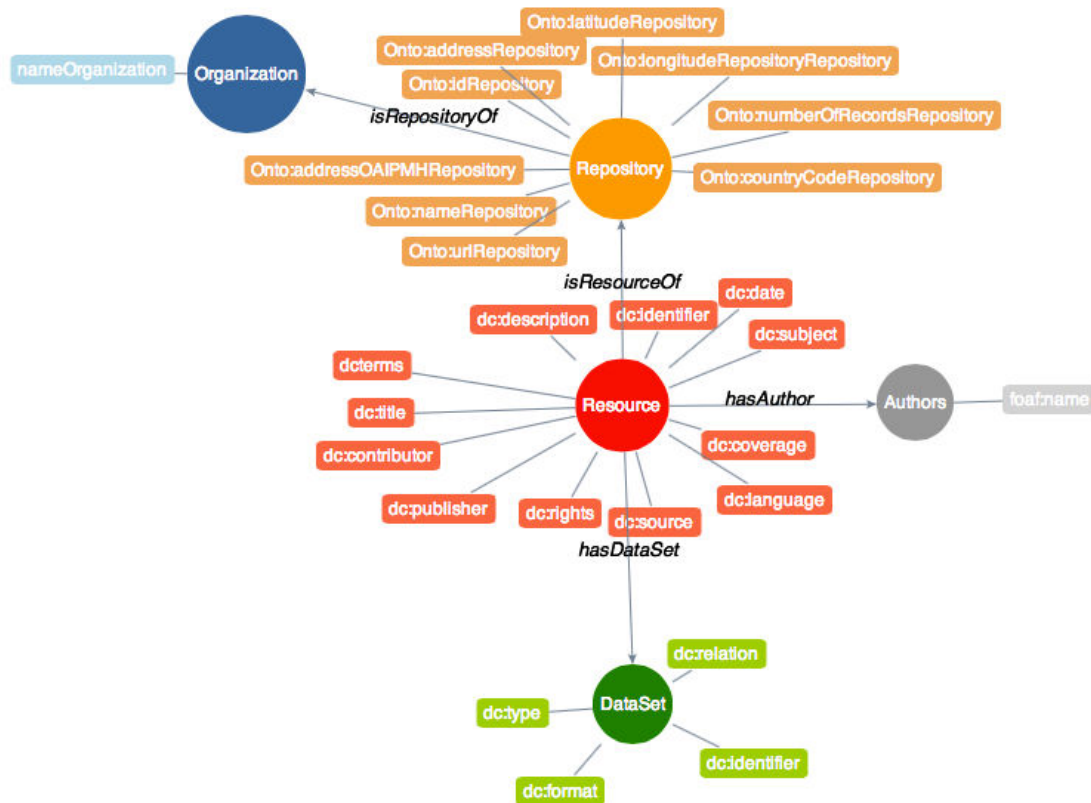


Figure 8 Schema of the ontology used for the Semantic Search Engine.

3.2.2 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, shown in Fig. 8 for the sake of completeness. The ontology, built using Dublin Core and 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 *hasDataSet* (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 detail in the next sub-section.

3.3 Implementation

The home page of the CHAIN-REDS Semantic Search Engine [13] is shown in Fig. 9. 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.

The screenshot is divided into two main sections. The top section, titled "Semantic Search on Linked Data", contains a search interface. It includes a text box for entering a keyword, a "Submit Query" button, and a language selection dropdown currently set to "English". Below the search box is a prompt: "Enter a keyword or select a language and then choose a subject".

The bottom section displays search results for a query. Each result entry includes a description, author information, and repository details. The first result is titled "Optimization of microsatellite genotyping and genetic sexing of non-invasively collected Polar Bear (Ursus maritimus) hair samples." and lists authors "Harris ChristopherMatthew". The second result is titled "(Table 3) Stable nitrogen and carbon isotope ratios, and mercury content in teeth of polar bears (Ursus maritimus)" and lists authors "Sonne Christian; Wiig Øystein; Dietz Rune; Aubail Aurore; Rigét FrankF; Caurant Florence". Both results include a "More Info" link. At the bottom of the results area, it says "Records found. Displaying 1 to 20" and a red link for "----More Resources----

Figure 9 Schema of the ontology used for the Semantic Search Engine (Up) and for the Semantic Search Engine (Down).

The results of a given query are listed in the summary view shown at the bottom part of Fig. 9.

For each record found, the title, the author(s) and a short description of the corresponding resource are provided. Clicking on “More Info”, visitors can access the detailed view of the resource, as shown in Fig. 10.

(1)The public sphere according to UK stem cell scientists

(2)The public sphere according to UK stem cell scientists

General Information

Authors: Koika Alexandra

Date stamp: 2012-07-02T13:29:59Z

Description (1) : In this thesis the concept of social representations is made relevant to the study of the 'public sphere' according to scientists. This is elaborated by the re-examination of the notion of a 'consensual' and a 'refilled universe' substantiating a more sociopsychological approach in the study of relevant phenomena. Two processes generate social representations of the public: anchoring and objectification. The empirical study investigates the scientists' views of the public sphere, in relation to public perceptions, media coverage and the regulation of cloning technology. Elite media coverage of the stem cell debate and conversations with stem cell scientists are systematically analysed with multiple methods. Findings are based on 461 news articles that appeared in Nature and Science between 1997 and 2005 and on interviews with 18 U.K based stem cell researchers conducted between February and October 2005. The analysis compares the debate before and after the 'stem cell war' of 2002, and typifies a high tension in representing the public sphere, elaborated in metaphors and prevailing arguments. Central elements of the representation assume a strong disassociation of science from the public sphere; peripheral elements operate with a degree of blurring of those same boundaries, which recognises a common project. This representation, while being expressive of its context of production, constitutes a functional response to it.

Description (2) : In this thesis the concept of social representations is made relevant to the study of the 'public sphere' according to scientists. This is elaborated by the re-examination of the notion of a 'consensual' and a 'refilled universe' substantiating a more sociopsychological approach in the study of relevant phenomena. Two processes generate social representations of the public: anchoring and objectification. The empirical study investigates the scientists' views of the public sphere, in relation to public perceptions, media coverage and the regulation of cloning technology. Elite media coverage of the stem cell debate and conversations with stem cell scientists are systematically analysed with multiple methods. Findings are based on 461 news articles that appeared in Nature and Science between 1997 and 2005 and on interviews with 18 U.K based stem cell researchers conducted between February and October 2005. The analysis compares the debate before and after the 'stem cell war' of 2002, and typifies a high tension in representing the public sphere, elaborated in metaphors and prevailing arguments. Central elements of the representation assume a strong disassociation of science from the public sphere; peripheral elements operate with a degree of blurring of those same boundaries, which recognises a common project. This representation, while being expressive of its context of production, constitutes a functional response to it.

Identifier (1) : Koika, Alexandra (2012) The public sphere according to UK stem cell scientists. PhD thesis, The London School of Economics and Political Science.

Identifier (2) : http://etheses.lse.ac.uk/391/1/Koika_The%20public%20sphere%20according%20to%20UK%20stem%20cell%20scientists.pdf

Subject (1) : HM Sociology

Subject (2) : HM Sociology

Language : en

Date : 2012-06

Dataset Information

Identifier : http://etheses.lse.ac.uk/391/1/Koika_The%20public%20sphere%20according%20to%20UK%20stem%20cell%20scientists.pdf

Type (1) : Thesis

Type (2) : NonPeerReviewed

Format : application/pdf

Relation : <http://etheses.lse.ac.uk/391/>

Repository information

Name: LSE Theses Online

URL : <http://etheses.lse.ac.uk/>

OAI-PMH address : <http://etheses.lse.ac.uk/cgi/oai2>

Country Code : GB

Address :

Longitude : -0.117

Latitude : 51.514

Domain : Multidisciplinary

Project : OpenDOAR

Organization : London School of Economics & Political Science

Figure 10 Detailed view of a record found by the Semantic Search Engine.

In the “Dataset information” panel users get the link to the open access document and, if

existing, to the corresponding dataset. Clicking on the “Graphs” tab, which appears at the top of the summary view (see bottom part in Fig.9), users can select one or more of the resources found and get a graphic view of the semantic connections among Authors, Subjects and Publishers, as shown in Fig. 11.

In this way, if new links appear, connecting different resources (as shown in the lower left corner of the figure), users can infer new relations among resources, thus discovering new knowledge. It is worth mentioning that this part of the Semantic Search Engine is at prototypal stage and is subject to changes and improvements in the coming months.

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 Fig. 3, is now to move from “Information” to “Inclusion” and identify/create new OADR 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. As an example on future integrations that can be made with African repositories, it is worth mentioning the case of Latin America, where La Referencia [19] is now part of the CHAIN-REDS Knowledge Base.

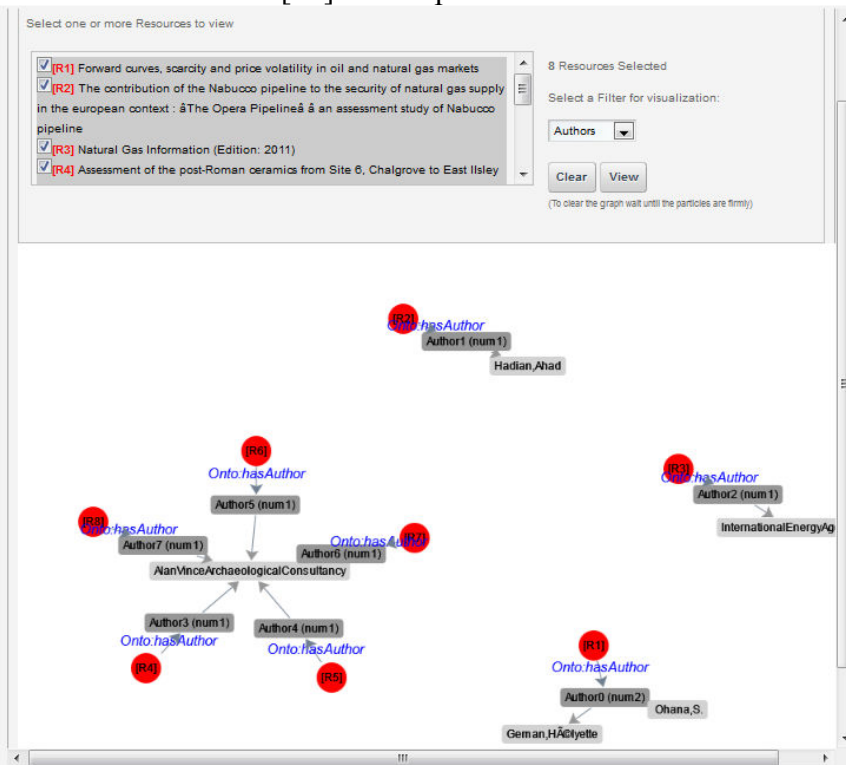


Figure 11 Graphic connections among records found the Semantic Search Engine.

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 [20] or Google Scholar [21]) and domain-specific semantic repositories (e.g., PubMed [22] or CIARD R.I.N.G. [23]), and ii) to allow the Search Engine to work on domain-specific sets of the semantic-web-enriched data.

References

Andronico, G. *et al* (2011), 'E-Infrastructures for International Cooperation' In: Preve, N. (ed), *Computational and Data Grids: Principles, Applications and Design* IGI Global, DOI: 10.4018/978-1-61350-113-9; see also www.igi-global.com/book/computational-data-grids/51946

Endnotes

[1] This definition of e-Infrastructure appears in an European Commission web page:

<http://cordis.europa.eu/ictresults/index.cfm?ID=90825§ion=news&tpl=article>

[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, from which we have re-used the picture included in Fig. 2

[3] The home page of the CHAIN-REDS project can be found at www.chain-project.eu

[4] The list of CHAIN-REDS partners can be inspected at www.chain-project.eu/partners

[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

[7] The home page of the DRIVER project can be found at www.driver-repository.eu

[8] The home page of the OpenAIRE project can be found at www.openaire.eu

[9] The home page of the OpenDOAR initiative can be found at www.opendoar.org

[10] The home page of the Databib initiative can be found at www.databib.org

[11] The home page of the DataCite initiative can be found at www.datacite.org

[12] The CHAIN-REDS Search Engine on Linked Data can be accessed at www.chain-project.eu/linked-data

[13] The Semantic Web standards can be inspected at

http://semanticweb.org/wiki/Semantic_Web_standards

[14] The OAI-PMH standard home page can be found at www.openarchives.org/pmh

- [15] The Dublin Core Metadata Initiative home page can be found at www.dublincore.org
- [16] The Apache Jena API home page can be found at <http://jena.apache.org>
- [17] The Virtuoso home page can be found at <http://virtuoso.openlinksw.com>
- [18] The Protégé home page can be found at <http://protege.stanford.edu>
- [19] La Referencia home page can be found at <http://lareferencia.redclara.net>
- [20] The home page of DBpedia can be found at www.dbpedia.org
- [21] The home page of Google Scholar can be found at <http://scholar.google.com>
- [22] The home page of PubMed can be found at www.ncbi.nlm.nih.gov/pubmed
- [23] The home page of CIARD R.I.N.G. can be found at <http://ring.ciard.net>

Biographies

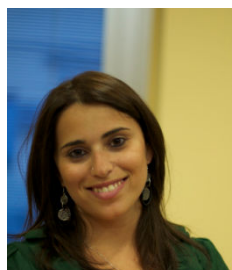


Prof. Roberto Barbera graduated in Physics "cum laude" at the University of Catania in 1986 and since 1990 he holds a Ph.D. in Physics from the same University. Since 2005 he is Associate Professor at the Department of Physics and Astronomy of the Catania University. Since his graduation his main research activity has been done in the domains of Experimental Nuclear and Particle Physics. He has been involved in many experiments in France, Russia, United States and Sweden to study nuclear matter properties in heavy ion collisions at intermediate energies. Since 1997 he has been involved in CERN experiments and he is currently one of the members of the ALICE Experiment at LHC. Within ALICE, he has been the coordinator of the Off-line software of the Inner Tracking System detector and member of the Off-line Board. Since late 1999 he is interested in Grid Computing. He has been member of the TERENA Technical Committee and he is member since 2009 of the Scientific & Technical Committee of GARR (the Italian National Research and Education Network). At European level, he has been involved with managerial duties in many EU funded projects, several of them involving South Africa. In 2004, he created the international GILDA Grid infrastructure for training and dissemination that he coordinates since the beginning and that has been used so far in more than 500 events in almost 60 countries worldwide. He is currently involved in the design and implementation of Science Gateways for various Virtual Research Communities and in the promotion of Data Infrastructures and Identity Federations in several regions of the world, including Africa.



Carla Carrubba was born in Augusta (Italy) in September 1982. She graduated in Computer Engineering "cum laude" at the University of Catania in 2011 with thesis "Analysis of Nystagmus through processing of saccadic eye movements captured with Eye Tracker Tobii T60". She worked, in 2007, as software programmer in PROIND s.r.l. and in 2011 in G.I.S. s.r.l. Since October 2011 she has worked at INFN of Catania with a two-year grant to

develop a small research project, named COGITO-MED that stands for CIOud and Grid Integration for disTRIBUTed data and applicatiON sharing in MEDicine. During this period she focused the study on the semantic search technologies.



Giuseppina Inserra was born in Augusta (Italy) on 15 December 1982. She studied at the University of Catania (Italy) where she graduated in computer engineering on January 2011 with a thesis on "Sviluppo di un software per effettuare esperimenti oculistici tramite l'utilizzo di Tobii eye tracker T60" (Development of a software to do ophthalmic experiments through the use of Tobii eye tracker T60). On September 2012 She got a master's degree in develop for mobile applications at the University of Catania. Since December 2011 she's working at the INFN of Catania, 2 year grant, and her work is focused on the study and use of the semantic web technologies and science gateway applications.



Rita Ricceri was born in Paterno' (Italy) in Luglio 1978. Since March 2005 she has worked at INFN of Catania participating in national and European projects about Grid computing, dealing with the production of dissemination tools and particularly in web development aligned with the projects specific objectives. She is currently involved in the implementation and development of theme and portlets for Science Gateways based on the Liferay platform.



Dr. Rafael Mayo García is Senior Researcher at CIEMAT and earned his PhD in Physics from the Universidad Complutense de Madrid (2004). From 2006 he has also been Adjunct Faculty and Honorary Fellow at the same University in the Physics of Materials Department. He has been involved in many experiments in Bulgaria, Sweden and Ireland (funded, among others, by the European Commission with a Marie Curie fellowship) to study plasma properties. He has also obtained a postdoctoral fellowship in the Spanish Juan de la Cierva Programme and worked on Data Networks for two years and a half. He is author of 26 scientific articles published in international JCR referenced journals (being cited 165 times) and more than 60 proceedings (being cited 288 times in Google Scholar). He has been involved in several European and National Projects working on ICT scientific developments (EGEE-III, EUFORIA, EFDA-ITM, Spanish e-Science Network) and even on managerial activities as Work Package Manager and/or member of Executive Boards (EELA, EELA-2, EPIKH, GISELA, CHAIN, CHAIN-REDS, BETTY). He also has served to several institutions as evaluator for their competitive calls, European Commission included.