Contribution ID: **45** Type: **Sessions contributions**

A CHAIN-REDS solution for data workflows

Wednesday, 21 May 2014 14:50 (20 minutes)

This work is presented on behalf of the CHAIN-REDS project (www.chain-project.eu) and proposes a solution for data workflows. The project has developed several tools that provide an easy and intuitive access to repositories hosted worldwide: the Knowledge Base (http://www.chain-project.eu/knowledge-base) displays in a two-fold basis (geo- and table-views) information about Open Access Document Repositories and Data Repositories (name, organization that manages it, country it belongs to, scientific domain it covers; direct link to access it); and, the Semantic Search Engine (http://www.chain-project.eu/linked-data) allows any user searching for a specific term in the previous repositories and also in those managed by the ENGAGE platform. In addition, CHAIN-REDS counts on a Science Gateway for executing unattendedly jobs on Grid, Cloud and local clusters.

Profiting from all these tools and the technological development underneath and by assigning Persistent Identifiers to specific datasets, a workflow will be presented that fully covers the common research cycle: search of specific term or data (repositories); retrieval of the associated raw data; use of these data as input for a specific application or code; execution of the latter; production of new data (results); and, upload and storage of the new publication and dataset into repositories already displayed in the Knowledge Base in order to start the cycle again.

Wider impact and conclusions

Big Data and their associated management and curation have become a major issue in scientific computation. As a consequence, huge initiatives such as the EUDAT and the Research Data Alliance initiatives are coordinating efforts worldwide.

Both projects have identified data workflows as an important component and have set up specific Working Groups devoted to propose useful solutions.

In this way, the one presented in this work by the CHAIN-REDS project not only has a direct impact on the final user daily work, but have also received a positive answer from both EUDAT and RDA as a valid case statement.

Description of work

In order to set up a data trust building, it is mandatory to define a set of standards that will rule the data management for facilitating its further use. Those promoted by CHAIN-REDS are OAI-PMH for metadata retrieval, Dublin Core as metadata schema, SPARQL for semantic web search, Resource Description Framework (RDF) for data interchange, and XML as potential standard for the interchange of data represented as a set of tables. To these standards, Persistent Identifier (PID) are also added as a long-lasting reference to a digital object.

By harvesting techniques based on the previous standards, the Knowledge Base retrieves the information from more than 3,100 repositories which contain more than 33 million of documents.

The metadata harvester is a process able to run both on Grid and Cloud infrastructures which consists of the following parts: get the address of each repository publishing an OAI-PMH standard endpoint; retrieve, using the OAI-PMH repository address, the related Dublin Core encoded metadata in XML format; get the records from the XML files and, using the Apache Jena API, transform the metadata in RDF format; and, save the RDF files into a Virtuoso triple store according to an OWL-compliant ontology built using Prot�g�. Each RDF file retrieved and saved in a Virtuoso-enabled triple store is mapped onto a Virtuoso Graph that contains the ontology expressly developed for the search engine.

The highest-level, component is the Search Engine itself. Using it, visitors can either enter a keyword and submit a SPARQL query to the Virtuoso triple. Thus, a user who is simply searching for a specific term or data and does not know which repository (or document) holds it, can find the required information.

Last, PID are used to assign permanent identifiers to specific repositories, allowing the retrieval of raw data that are used as input of a code and, at the same time, can be used to assign a new PID to the output of the

previous execution

URL(s) for further info

http://science-gateway.chain-project.eu
http://www.chain-project.eu/knowledge-base
http://www.chain-project.eu/linked-data
http://www.chain-project.eu

Primary authors: RODRIGUEZ-PASCUAL, Manuel (CIEMAT); MAYO-GARCIA, Rafael (CIEMAT)

Co-authors: Mr RUBIO-MONTERO, Antonio (CIEMAT); Ms CARRUBBA, Carla (INFN); KANELLOPOULOS, Christos (GRNET); SCARDACI, Diego (INFN); RUGGIERI, Federico (INFN - Roma Tre); Dr LA ROCCA, Giuseppe (INFN); Ms INSERRA, Giuseppina (INFN); PRNJAT, Ognjen; Mrs RICCERI, Rita (INFN); BARBERA, Roberto (University of Catania and INFN)

Presenters: RODRIGUEZ-PASCUAL, Manuel (CIEMAT); MAYO-GARCIA, Rafael (CIEMAT)

Session Classification: New data management solutions for EGI

Track Classification: Requirements and solutions for data management and computing (Track Leaders: B. Konya, H. Heller, S. Tarkoma)