The Catania Science Gateway for the EGI Long Tail of Science – A How-to

Last modified: 25/11/2016
Version: 1.0

1. Introduction

The Catania Science Gateway Framework[1] (CSGF) is an open source toolkit jointly developed by the Division of Catania of the Italian National Institute for Nuclear Physics (INFN) and by the Department of Physics and Astronomy of the University of Catania. CSGF allows build, in a fast and easy way, standard-based web 2.0 Science Gateways to exponentially increase the number of potential users of Distributed Computing Infrastructure (DCIs) worldwide. The CSGF is released under the Apache 2.0 license and all code is available on GitHub[2].

The framework, conceived in the context of both EU-funded and national project, has been used in the last 5-6 years to develop Science Gateways for several EU co-funded projects such as: DECIDE[3], EarthServer[4], EUMEDGRID-Support[5], GISELA[6], DCH-RP[7], INDICATE[8] and CHAIN-REDS[9], just to name a few. The framework is currently being completely re-engineered in the context of the INDIGO-DataCloud[10] to offer a rich RESTful

[1] http://www.catania-science-gateways.it/
[2] https://github.com/csgf/
[3] https://www.eu-decide.eu/
[4] http://www.earthserver.eu/
[5] http://www.eumedgrid.eu/
[6] http://www.gisela-grid.eu/
[7] http://www.dch-rp.eu/
[8] http://www.indicate-project.org/
[9] https://www.chain-project.eu/
[10] https://www.indigo-datacloud.eu/

API as well as to include additional cloud, data and workflow data management functionalities and to improve its performances and reliability.

This short guide is intended for the users of the CSGF-based Science Gateway[11] built for the EGI Long Tail of Science[12] (LToS). Within these pages, users will find an introduction on how to use the scientific applications currently available in the portal. The guide does not cover the registration and authentication of users, for which the reader can see this page[13].

Once the user is successfully registered, authenticated and authorized s(he) is then presented with the web page(s) of the application(s) s(he) is allowed to execute on the EGI LToS infrastructure.

The complete list of applications the user is entitled to run is available navigating the Applications menu located in the top bar, as shown in Figure:

2. Applications

In the next sub-sections, we will describe each of these applications in details.

2.1 R

The R Project for Statistical Computing[14] is a language for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S (see this page[15] for more information).

R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, etc.) and graphical techniques, and it is highly extensible. The S language is often the

[11] http://csgf.egi.eu
[12] https://wiki.egi.eu/wiki/Long-tail_of_science
[13] https://access.egi.eu/start
[14] https://www.r-project.org/
[15] https://www.r-project.org/about.html

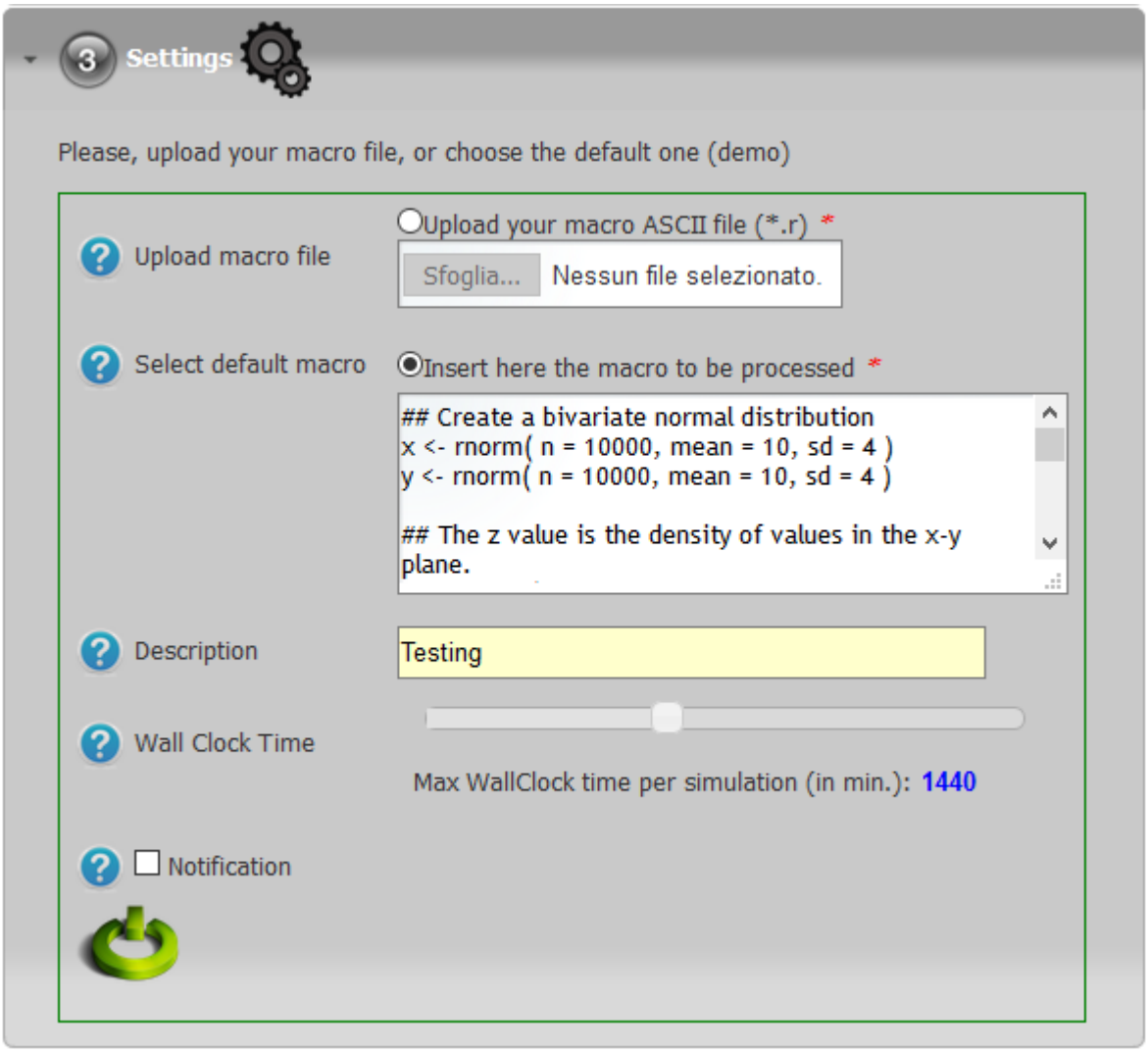
vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity.

One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. Great care has been taken over the defaults for the minor design choices in graphics, but the user retains full control. R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and MacOS.

A dedicated web page to access the service and execute the Statistical R for Computing is available at this URL[16]. The service is organized in three different accordions. The first one shows some generic information about the service, the second shows the list of computing resources entitled to run statistical analysis on High Throughput Computing (HTC) resources.

In the last accordion, the user can select to upload a local R macro file selecting the Browse button, or using the text box. By default, a macro demo is loaded when the corresponding radio button is clicked.

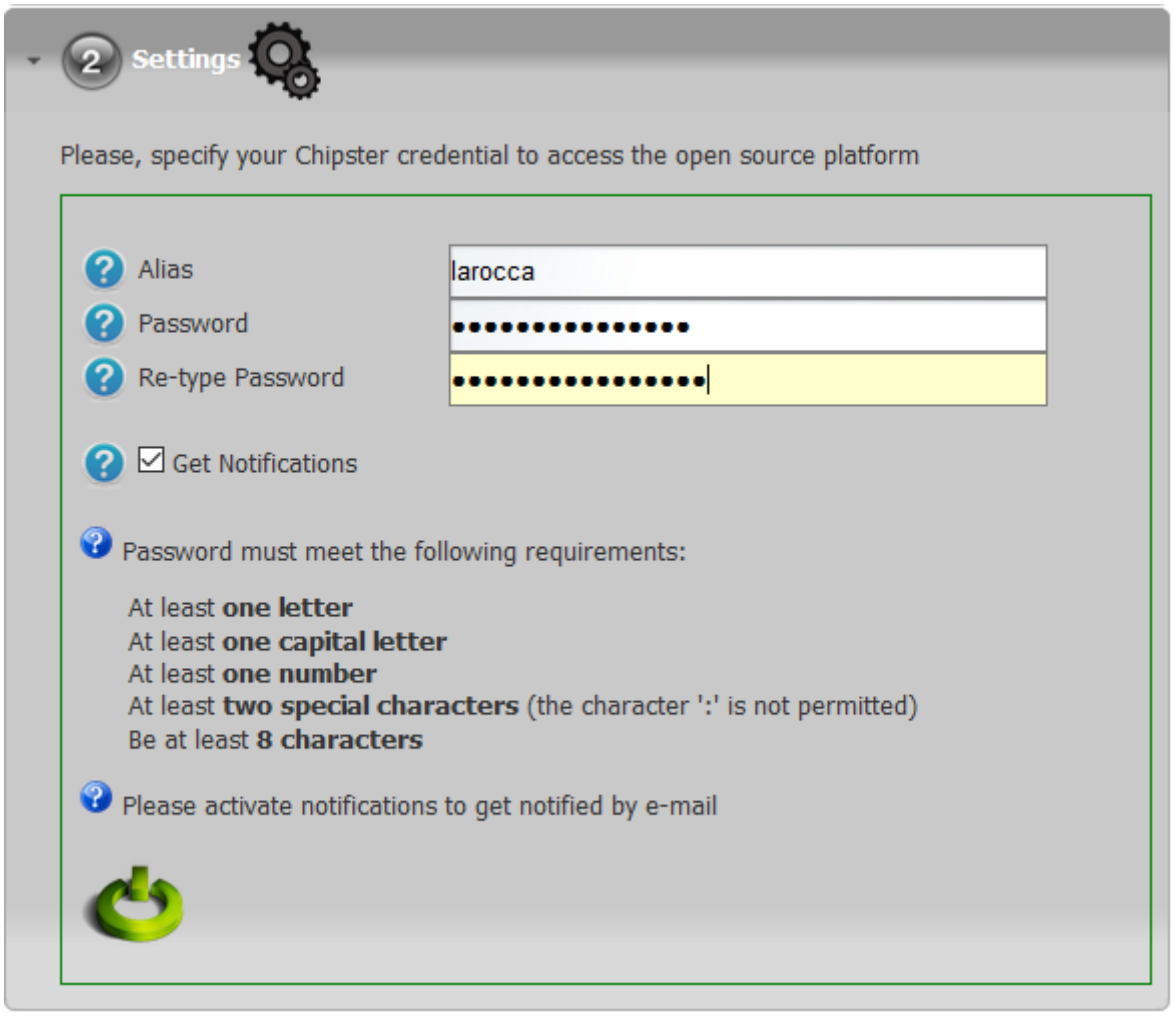
[16] https://csgf.egi.eu/run-r



2.1 Chipster

Chipster is a user-friendly analysis software for high-throughput data. It contains over 300 analysis tools for next generation sequencing (NGS), microarray, proteomics and sequence data. Users can save and share automatic analysis workflows, and visualize data interactively using a built-in genome browser and many other visualizations. Chipster's client software uses Java Web Start to install itself automatically, and it connects to computing servers for the actual analysis. Please see the Chipster main site[17] for courses, updates and other information. A dedicated web page to create a Chipster account for accessing the open source platform is available at this page[18].

[17] http://chipster.csc.fi/
[18] https://csgf.egi.eu/run-clustalw2



If ‘Get Notification’ is checked, the user will receive via e-mail instructions on how to access the Chipster Server.

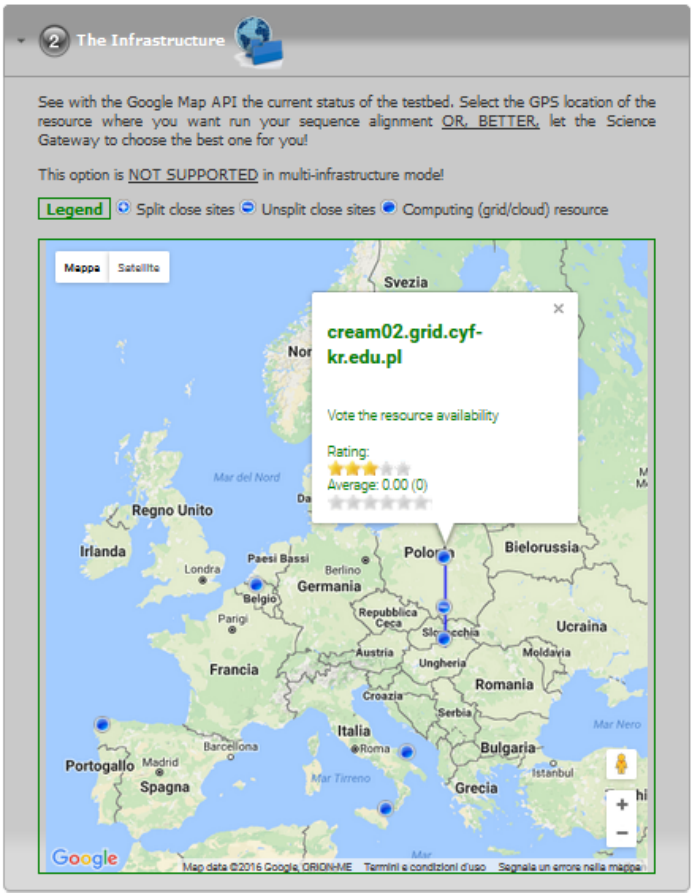
2.3 ClustalW2

ClustalW2[19] is a widely used multiple program for multiple alignment of nucleic acid and protein sequences. sequence alignment computer program.

The program accepts a wide range on input formats including: NBRF/PIR, FASTA, EMBL/Swissprot, Clustal, GCC/MSF, GCG9 RSF, and GDE, and executes the following workflow:

- Pairwise alignment;
- Creation of a phylogenetic tree (or use a user-defined tree);
- Use of the phylogenetic tree to carry out a multiple alignment.

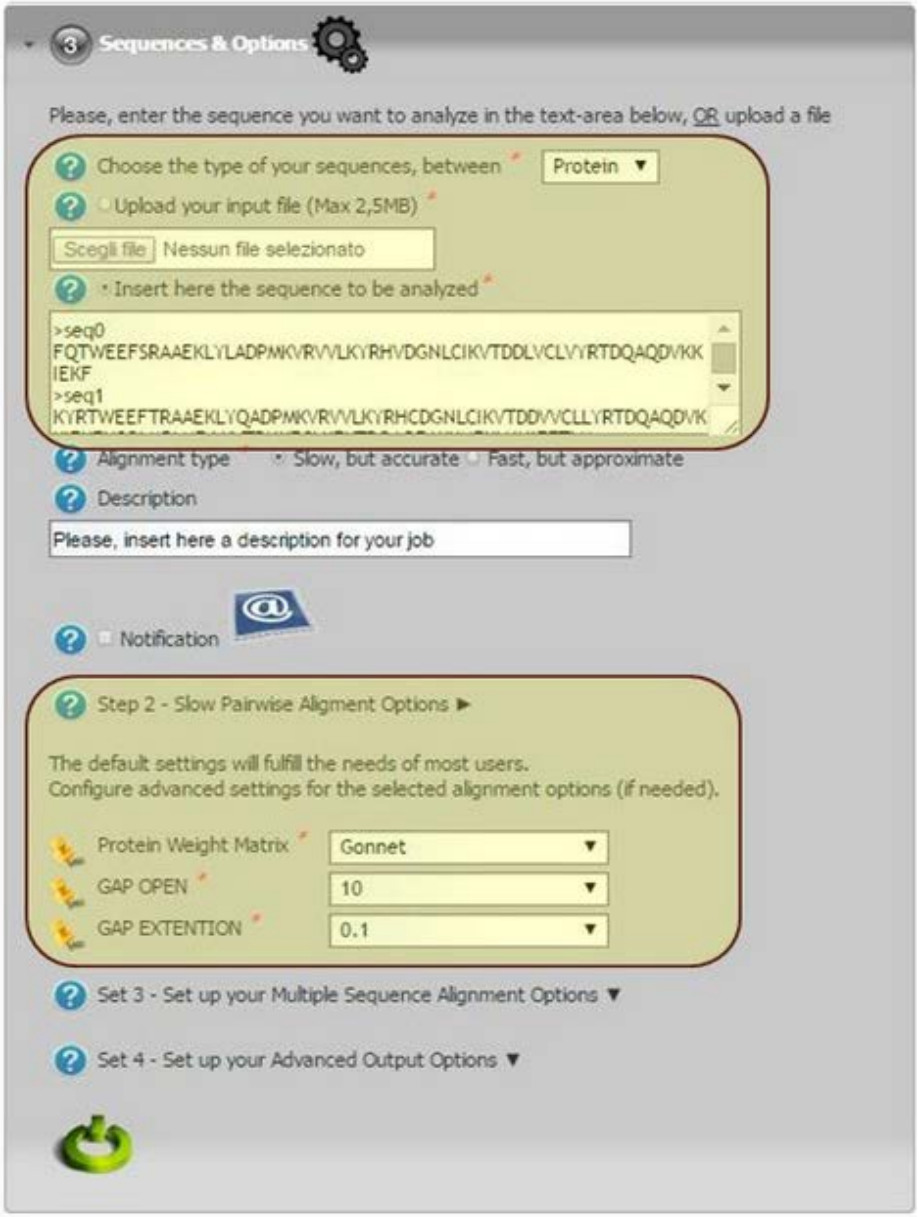
A dedicated web page to access the service and execute the Multi Sequence Alignments for DNA or protein is available at this page[20]. The service is organized in three different accordions. The first one shows some generic information about the service, the second shows the list of computing resources entitled to run the multi sequence alignments process.



[19] http://www.clustal.org/
[20] https://csgf.egi.eu/run-clustalw2

By default, the LToS CSG will send the request to one of these resources selected randomly, but the user can also select one of them if s(he) wants to force the execution of the sequence alignment on a specify one.

All the input settings needed to start the multi sequence alignment can be specified in the last accordion, as shown in the figure below.



Each simulation will produce the following files:

- `std.out`: the standard output file;
- `std.err`: the standard error file;
- `outputs.tar.gz`: the file containing the results of the Monte Carlo simulation.

2.4 Semantic Search Engine

The Semantic Search Engine (SSE) is a framework conceived to demonstrate the potential of information coupled with semantic web technologies to address the issues of data discovery and correlation. Two different version of the SSE are available on the CSG:

A simple Semantic Search Engine (SSE) customized for the DARIAH Competence Centre[21] project which allows users to search in the e-Infrastructure Knowledge Base, in more than 100 languages across more than 30 million resources contained in the thousands of semantically enriched Open Access Document Repositories[22] and Data Repositories[23]. Search results are ranked according to the Ranking Web of Repositories[24].

The search can be performed:

- By typing one or more keywords in the textbox search. After entering the search string, clicking the "Search" button the system submit a SPARQL[25] query that retrieve the results from the e-Infrastructure Knowledge Base.

  Example of query to retrieve the resources with the keyword in title.

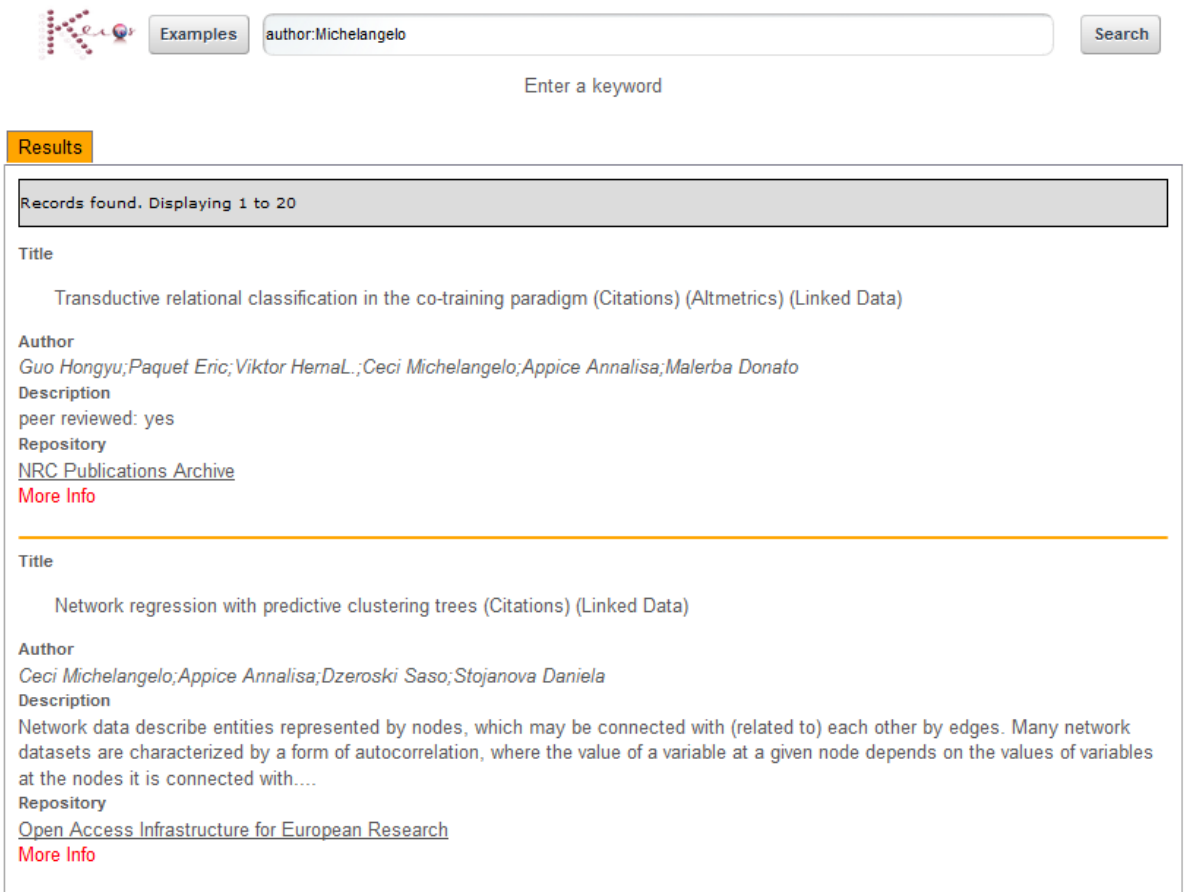
```
SELECT distinct ?s WHERE {

?s dc:title ?title. ?title bif:contains ″ + keyword + ″. ?s
<http://semanticweb.org/ontologies/2013/2/7/RepositoryOntology.owl#isResourc
eOf> ?rep. ?rep
<http://www.semanticweb.org/ontologies/2013/2/7/RepositoryOntology.owl#rank>
?rank. }ORDER BY ASC(?rank) limit 20 offset 0
```

- Using search filters. Inserting in textbox search string, for example, "author: G. Smith" and clicking "Search" will be launched a SPARQL query that will return only the resources of the e-Infrastructure Knowledge Base that have at least one author who matches "Smith G.". The possible filters (according to the Dublin Core Standard) are : dc: author, dc: subject dc: type and dc: publisher. Clicking the "Examples"

[21] https://wiki.egi.eu/wiki/Competence_centre_DARIAH
[22] http://www.sci-gaia.eu/e-infrastructures/knowledge-base/oadr-map/
[23] http://www.sci-gaia.eu/e-infrastructures/knowledge-base/data-repositories-map/
[24] http://repositories.webometrics.info/
[25] https://www.w3.org/TR/rdf-sparql-query/

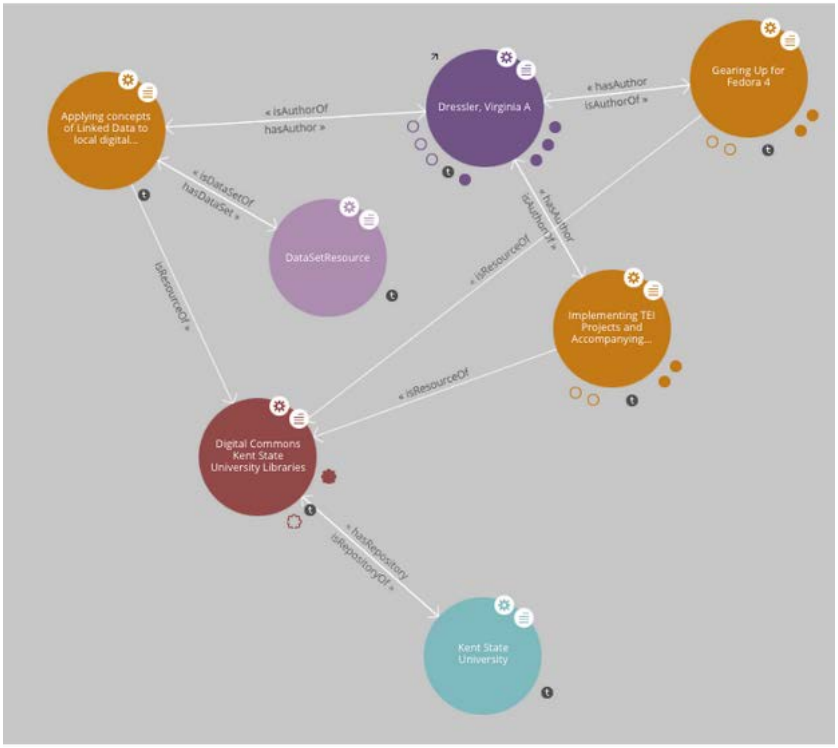
button are show examples of search filters in 4 of the 110 languages supported by the system.

In both cases, the results of the query will be processed further, to obtain the final results, as shown in the following figure.



Clicking on "More Info" link you can access all the details of the resource.
Clicking on "Citations" link, if available, are displayed more information from Google Scholar about the versions and the quotations of the resource.
Clicking on "Linked Data" link will be open a new tab that shows a LodLive[26]-based graph of the resource with all its metadata.

[26] http://en.lodlive.it/



- A parallelised SSE (PSSE) configured to simultaneously search across the above e-Infrastructure Knowledge Base, Europeana[27], Cultura Italia[28], Isidore[29], OpenAgris[30], PubMed[31] and DBpedia[32] platforms (others can be added upon request to sg-licence@ct.infn.it).

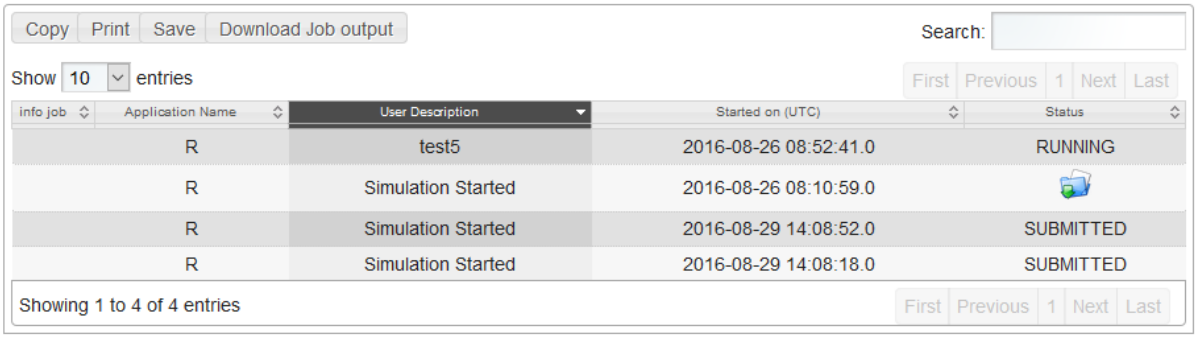
[27] http://www.europeana.eu/portal/
[28] http://dati.culturaitalia.it/?locale=it
[29] http://www.rechercheisidore.fr/
[30] http://agris.fao.org/openagris/index.do
[31] http://www.ncbi.nlm.nih.gov/pubmed
[32] http://wiki.dbpedia.org/

3. MyJobs

To monitor active jobs and download results at the end of the computation, users just need click on the MyWorkSpace link which appears in the top bar.

A complete list of Active and Done jobs will be showed in a table view as shown in the Figure below. Statuses are automatically updated every 15 minutes so there is no need to reload this page more frequently.

Once your jobs have finished, user has 96 hours to retrieve the output. Beyond that time, the output of jobs will automatically be deleted from the Science Gateway in order not to fill its storage with undesired stuff.



4. Support

If you have any question, please contact the Catania Science Gateway support team.