Geophysical Research Abstracts
Vol. 12, EGU2010-11935-1, 2010
EGU General Assembly 2010

Harvesting implementation for the GI-cat distributed catalog

Enrico Boldrini (1), Fabrizio Papeschi (1), Lorenzo Bigagli (1,2), Paolo Mazzetti (1,2)

(1) Italian National Research Council (CNR), Institute of Methodologies for Environmental Analysis (IMAA), Prato, Italy {boldrini, papeschi, bigagli, mazzetti}@imaa.cnr.it, (2) University of Florence, Prato, Italy {lorenzo.bigagli, paolo.mazzetti}@pin.unifi.it

GI-cat framework implements a distributed catalog service supporting different international standards and interoperability arrangements in use by the geoscientific community.

The distribution functionality in conjunction with the mediation functionality allows to seamlessly query remote heterogeneous data sources, including OGC Web Services - e.e. OGC CSW, WCS, WFS and WMS, community standards such as UNIDATA THREDDS/OPeNDAP, SeaDataNet CDI (Common Data Index), GBIF (Global Biodiversity Information Facility) services and OpenSearch engines.

In the GI-cat modular architecture a distributor component carry out the distribution functionality by query delegation to the mediator components (one for each different data source). Each of these mediator components is able to query a specific data source and convert back the results by mapping of the foreign data model to the GI-cat internal one, based on ISO 19139.

In order to cope with deployment scenarios in which local data is expected, an harvesting approach has been experimented.

The new strategy comes in addition to the consolidated distributed approach, allowing the user to switch between a remote and a local search at will for each federated resource; this extends GI-cat configuration possibilities.

The harvesting strategy is designed in GI-cat by the use at the core of a local cache component, implemented as a native XML database and based on eXist. The different heterogeneous sources are queried for the bulk of available data; this data is then injected into the cache component after being converted to the GI-cat data model.

The query and conversion steps are performed by the mediator components that were are part of the GI-cat framework. Afterward each new query can be exercised against local data that have been stored in the cache component.

Considering both advantages and shortcomings that affect harvesting and query distribution approaches, it comes out that a user driven tuning is required to take the best of them. This is often related to the specific user scenarios to be implemented.

GI-cat proved to be a flexible framework to address user need. The GI-cat configurator tool was updated to make such a tuning possible: each data source can be configured to enable either harvesting or query distribution approaches; in the former case an appropriate harvesting interval can be set.