Repository Infrastructure for Data and Provenance recording

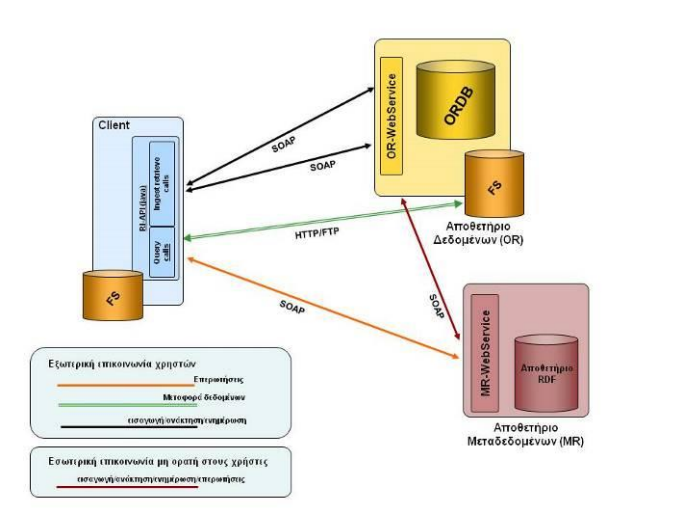
Overview

The Repository Infrastructure (*RI*) is an open source client-server system for **managing datasets produced in long and complicated workflows**, such as scientific research data and industry production datasets. Detailed digital documentation, sufficient to allow reproducibility, is of very high importance in this context. Client organizations can access at any time the repository services to upload new or update the existing material, to query and browse data, make annotations and download files.

The infrastructure provides a **flexible server installation** that can be either distributed or single-node, incorporating a central RI service communicating with several independent end-point services. A java-client API enables third-parties to build applications that ingest and retrieve data and metadata in a well-organized manner.

We have implemented three such applications (tools): (i) the ReposIt Tool for ingestion of complete workflows following predefined templates, (ii) the BrowseIt Tool for querying, browsing and viewing results and (iii) the CoRef Tool for identifying and resolving co-references.

An essential feature of the system, along with the data manipulation capabilities, is the ability to record semantic information on the data, enabling the users to query the data on their semantics and to make inferences based on these semantics.



System Architecture

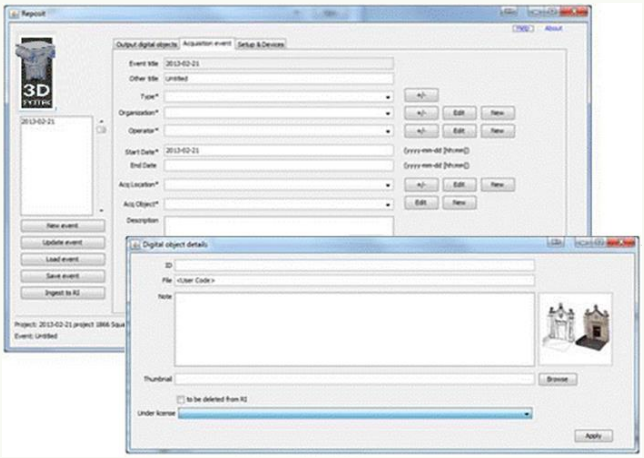
Target Applications

The RI is an ideal solution for recording data provenance for projects manipulating digital data involved in a data production-line of multiple steps. 2D-image and 3D-model production are typical use-cases of such applications that comprise multiple acquisitions and processing procedures. The system can be configured and parameterized in order to describe specific use-cases and data.

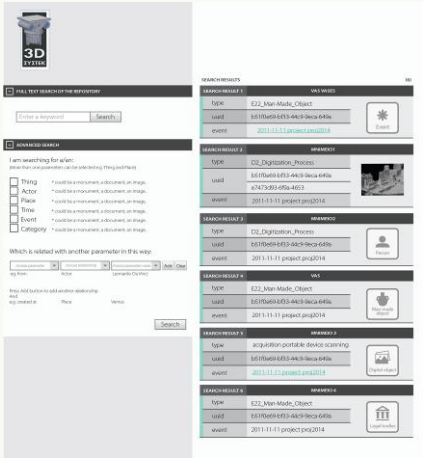
Description

Two essential components comprise the server-side part of the RI: (i) the Object Repository (*OR*) which is responsible for recording the ingested data creating the appropriate entries in the embedded relational database, (ii) the Metadata Repository (*MR*) which is responsible for handling the Semantic metadata and includes an integrated Query Manager (QM) module, used to enable complex querying of both the OR db and the MR triple-store. The querying mechanism of the RI is a powerful utility for the end-user, as it provides semantic inferred knowledge in conjunction to the ingested data.

A central web service (*RI web service*), which acts as the front-end of the RI server, is handling all client requests. It provides methods for ingesting data and metadata, updating metadata with newer versions, retrieving existing files and making queries. All 'write' operations are managed by the RI web service which is also responsible to implement the RI business logic before proceeding to changes on the low level database structures. All Tools communicate with the RI through this API interface.



ReposIt Tool for ingestion
(configured for 3D-SYSTEK project)



BrowseIt Tool for querying
(configured for 3D-SYSTEK project)

Contact details: **Martin Doerr**
martin@ics.forth.gr
www.ics.forth.gr/isl/cci.html