Tracking Dataset Transformations with HAPPI Toolkit

Wednesday, 11 November 2015 09:00 (1h 30m)

Results of the research community are based on three main pillars: models of phenomena, dataset gathered from missions and campaigns, validation and refinement of models based on dataset. Since its acquisition and during the whole life cycle of the research processes, dataset undergoes through many transformations (e.g. capture, migration, change of custody, aggregation, processing, extraction, ingestion) in order to be opportunely processed, analysed, exchanged with different researchers and (re-)used.

Consequently, trustworthiness of results, and of research community itself, rely on tracking dataset transformations within the whole life cycle of the research processes. Tracking dataset transformations becomes more important whenever dataset has to be treated from researchers communities of different domains and/or the research processes may span over a long interval of time.

Open Archival Information System (OAIS ¬- ISO:14721:2012) [1] has identified as "provenance information" the type of metadata where to store and track changes undergone to a generic digital object since its creation. Provenance is part of the so call OAIS Preservation Description Information, the metadata used to preserve digital object in a long-term digital archive, and it includes i) reference information (persistent identifier assigned to digital object); ii) provenance information; iii) context information (relationships to other digital objects), iv) fixity information (information used to ensure that digital object has not been altered in an uncontrolled manner) and v) rights information (permitted roles to access and transform digital object).

The HAPPI Toolkit [2], part of the Data Preservation e-Infrastructure produced by the SCIDIP-ES project [3], traces and documents dataset transformations by adopting the Open Provenance Model, a simple information model based on three basic entities (i.e. controller agent, transformation, digital object) that improves interoperability and capability to exchange information among different digital archives and/or research communities. Moreover, HAPPI Toolkit generates for each transformation a record (called Evidence Record) that includes reference information and integrity information. The collection of records represent the history of all the dataset transformations is called Evidence History, and this information is managed by HAPPI Toolkit and provides data managers with evidences that are used during the assessment of the integrity and authenticity of the dataset.

Since July 2014, HAPPI Toolkit is running on EGI FedCloud.

The tutorial aims to presents how HAPPI Toolkit works, and specifically: how HAPPI Toolkit is configured, how it creates the evidences of dataset transformations, how users can access evidences and dataset information.

Summary

interoperability, reusability, curation and preservation of data

Primary author: Mr BRIGUGLIO, Luigi (Engineering Ingegneria Informatica S.p.A.)

Presenter: Mr BRIGUGLIO, Luigi (Engineering Ingegneria Informatica S.p.A.)

Session Classification: Tutorial: HAPPI toolkit