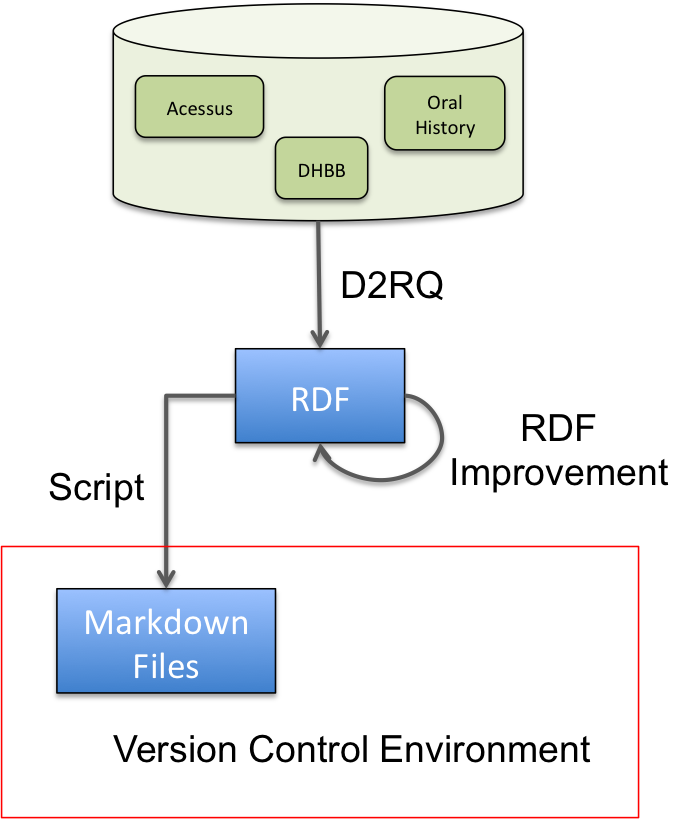
Many of the information systems proposed in the 80’s became obsolete with some of the technologies proposed lately for storing data. Relational databases are often hard to maintain, scale and share. Also the idea of having closed and very controlled systems is being increasingly replaced by the concept of having open and shared information systems. In such systems the responsibility of updating and creating new data is not sustained by a single institution but usually by a whole community that share knowledge and interests with associates. This way the data keep up-to-date, accessible and growing much faster due to the increased number of contributors.

In this context CPDOC is planning to improve the way their rich historical data is accessed, stored and shared. Currently the information is stored in relational databases that are made available restrictedly, but to face the challenge of maintaining the excellence in historical data storage and access, the institute decided to invest in technologies that privileges the open, shared way of dealing with data.

This new strategy is meant to be implemented in two different and subsequent moments: migration of the current relational technologies to new graph-based ones; and development of an environment for maintaining and creating data.



The figure above illustrates how the migration process is planned.

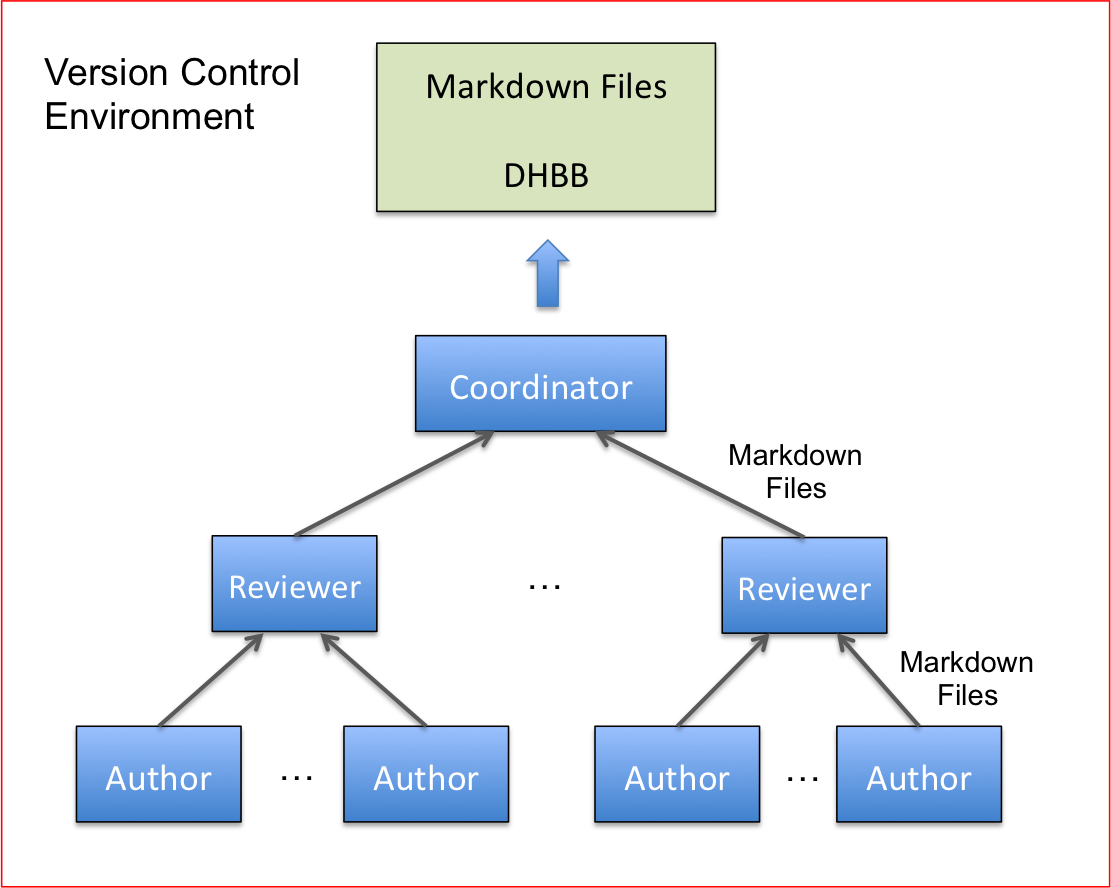
The idea is to migrate the databases in CPDOC from the current relational database systems to systems of data storage based in graphs, such as the RDF. It is important to mention that CPDOC has currently three different information systems, and the pilot project will involve only one of them: the DHBB. This system was chosen due to different administrative reasons, but also because it has the simplest data model.

Some tools are available for this kind of conversion and one of the most established is the D2RQ. Once the RDF is generated, a step for improving and enriching the RDF file, checking for data consistency and gathering new data connections using the available knowledge sources (ontologies and dictionaries?) is planned.

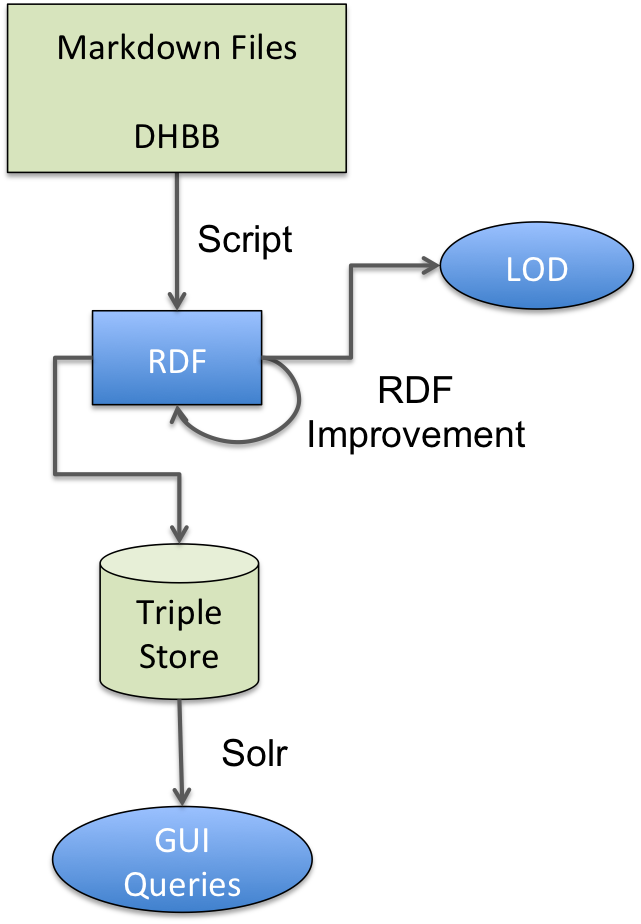
With the final RDF file ready, a script would be responsible for generating Markdown files in a version control environment. Markdown is a lightweight markup language that stores metadata in the header and comprises a raw text corresponding to the information to be stored. This environment of a group of files, a version control system and the RDF interface is meant to compose the new data storage for DHBB.

The following diagram shows how the CPDOC team would work to create new data for the DHBB database. The idea is to maintain the current hierarchy of cooperation in CPDOC. A version control system such as Git, allow for handling this straightforwardly. The authors, responsible for creating the raw versions of historical data, would create files following a Markdown file protocol. These files would then be revised by reviewers that submit their final versions to the coordinator, responsible for feeding the system with new data. It is interesting to notice that all these file exchanges are recorded and the whole process is controlled so as to ensure data consistency and security.

This new process would easy very much the current work for reviewers and coordinators, since presently there is no aid system for this task, basically performed using text files and emails



The new database system composed by Markdown files and periodic RDF releases can be easily used for accessing data in an efficient way. The figure below illustrates the process. CPDOC and their collaborators would be responsible for creating Markdown files that would be automatically analyzed by a script for generating the RDF file, which would be refined and improved. This refined RDF would be made available to the web as linked open data, which would allow the information to be accessed and improved by the community. On the other hand, the RDF would also be stored in a triple store that allows for queries using many different GUI configurations (for instance using Solr) to easy the access of data stored. This schema follows a modern, open and scalable way of sharing and improving the data stored in CPDOC.



.