Elements of a Curation Grid

Andreas ASCHENBRENNER

Goettingen State and University Library, DE

While earlier grid activities focused on the virtualisation of compute hardware, there is a growing awareness about the importance of data curation, as well as the utility of grid technologies in supporting curation. However, agreement about what "data curation" exactly means and what it entails still remains to be found, particularly when diverse communities, types of data, and research lifecycles are involved. The D-Grid project WissGrid [1] embarked on defining these elements of curation, and develops grid architecture for supporting curation activities. This paper presents the concepts and foundations of the curation grid developed thus far, identifying the responsibilities of e-Infrastructure as well as those of grid communities (i.e. the data creators and future users), and specifying technical requirements but also the organisational measures needed for data curation.

There is increasing recognition that adequate curation of digital data potentially improves - amongst other - the collaboration across fields (e.g. through interoperability), quality of research (e.g. through better validation of research results), and lowers overall costs (e.g. through re-usability). Projects like the Australian National Data Service (ANDS) [2], DataNet in the USA [3], and the nascent PARADE in Europe [4] aim to tap into these opportunities.

The Digital Curation Centre in the UK defines digital curation to involve "maintaining and adding value to a trusted body of digital information for current and future use" [5]. When looking closer at digital curation, we distinguish three curation levels. These follow the abstraction levels of digital objects suggested by Thibodeau [6] and they are well recognised in the preservation community. The levels build upon each other and may influence and support each other. The technical and organisational methods mentioned in the following are exemplary; they may change depending on the context and evolve over time.

1. Bitstream Preservation - the integrity of each bit Bitstream Preservation involves the technical and organisational infrastructure for monitoring the physical stability of data, and moving data to fresh and up-to-date carriers.

Components of Bitstream Preservation include durability of the hardware carriers and readers; their replication and recurrent integrity checks.
2. Content Preservation - processable data units Citability and accessibility of data items goes beyond the mere availability of the item's bits and touches e.g. upon the software with which it was created and its format.

Components of Content Preservation include persistent identification, technology watch, as well as preservation services such as format conversion, format validation, or emulation.

3. Data Curation - transporting meaning Data Curation views the object in its intellectual context and over its whole life cycle. It aims to preserve the meaning of the digital object to foster comprehension and re-usability in the future when the original context (e.g. the creators, organisational environment) has disappeared.

Components of Data Curation include data and metadata modelling, appraisal of the object's value, embedding in applications, object versioning, access control, and others. Examples include the life cycle of instrumental data, as it is captured, cleaned, filtered, and processed iteratively; or the enrichment of ancient texts with dictionaries and other contextual data from their time of creation.

The WissGrid project works together with partners in D-Grid and in the digital preservation community to offer Bit-Preservation on grid resources, and to support user communities in Content Preservation and Data Curation through tools and policy guidance. Its architecture is designed to be open, such that services can accommodate existing data management environments and evolve over time.

As part of its objective to also support Content Preservation and Data Curation, WissGrid analysed different integration patterns of e.g. grid technologies and preservation services such as JHOVE and CRiB, and developed integration approaches of repository systems into grids:

a. Repositories as data back ends for grid processing
b. Data grids as repository storage
c. Virtualisation of digital objects in federated repositories

By combining the experiences and technologies from both, the grid and the curation communities, WissGrid aims to achieve a comprehensive approach, which is supported by both communities and hence conducive to sustainability.