Data management in the humanities

Laurent Romary

Data management in the humanities

Laurent Romary

Inria

DARIAH, Transitional phase Director — DARIAH-EU

From the growing interest for digital methods in nearly all research domains in the humanities, reflected by the number of projects supported at national and European level, we need to understand the tenets of digitally based scholarship and how specific data management issues in the humanities are. To this end the ESFRI[1] roadmap on European infrastructures has been seminal in identifying the need for a coordinating e-infrastructure in the humanities, DARIAH[2], whose data policy is reflected in this paper.

Scholarly data in the humanities is a heterogeneous notion. Data creation, seen as the transcription of a primary document, the annotation of existing sources or the compilation of observations across collections of objects, is inherent to the scholarly activity and thus makes it strongly dependant upon the actual hypotheses or theoretical backgrounds of the researcher. There is indeed hardly any notion of data centre in the humanities since data production and enrichment is anchored on the individuals performing research.

The long-term vision[3] of DARIAH is to create a sound and solid infrastructure to ensure the long-term stability of digital assets, as well as the development of a wide range of even unanticipated services to carry out research on these assets. This comprises both technical aspects (identification, preservation), editorial (curation, standards) and sociological (openness, scholarly recognition).

Such a vision is articulated around the notion of *digital surrogates*, which are information structures intended to identify, document or represent a primary source used in a scholarly work. Surrogates can be metadata records, scanned image of a document, digital photographs, transcription of a textual source, or any kind of extract or transformation[4] on some existing data. Surrogates act as a stable reference for further scholarly work, in replacement – or in complement – to the original physical source it represents or describes. Moreover, a surrogate can act as a primary source for the creation of further surrogates, thus forming a network that reflects the various steps of the scholarly workflow where sources are combined and enriched before being further disseminated to a wider community.

Such a unified data landscape for humanities research implies that a clear policy in the domain of standards and good practices be defined. Scholars should both benefit from strong initiatives such as the TEI[5] (Text Encoding Initiative) and stabilize their experience by participating to the development of standards, in collaboration with other communities of practice (publishers, cultural heritage institutions, libraries).

---

[1] ec.europa.eu/research/esfri/
[2] www.dariah.eu/
[3] http://ec.europa.eu/information_society/newsroom/cf/itemlongdetail.cfm?item_id=6204
[4] E.g. the spectral analysis of a recorded speech signal
[5] www.tei-c.org

The vision also impacts on the technical priorities for DARIAH, namely:

- deploying a repository infrastructure where researchers can transparently and trustfully deposit their productions, comprising permanent identification and access, targeted dissemination (private, restricted and public) and rights management, possibly in a semi-centralized way allowing efficiency, reliability and evolution (cf. http://hal.archives-ouvertes.fr/hal-00399881);

- defining standardized interfaces for accessing data through such repositories, but also through third-party data sources, with facilities such as threading, searching, selecting, visualising and importing data;

- experimenting the agile development of virtual research spaces based on such services, integrating community based research workflows (see http://hal.inria.fr/inria-00593677).

Beyond the technical aspects, an adequate licensing policy must me defined to assert the legal conditions under which data assets can be disseminated. This should be a compromise between making all publicly financed scholarly productions available in open access and preventing the adoption of heterogeneous reuse constraints and/or licensing models. We contemplate encouraging the early dissemination of digital assets in the scholarly process and recommend, when applicable, the use of a Creative Commons CC-BY license, that supports systematic attribution (and thus citation) of the source.

From a political point of view, we need to see with potential data providers (cultural heritage entities, libraries or even private sector stakeholders such as Google) how to create a seamless data landscape where the following issues should be jointly tackled:

- General reuse agreements for scholars, comprising usage in publications, presentation on web sites, integration (or referencing) in digital editions, etc.;

- Definition of standardized formats and APIs that could make the access to one or the other data provider more transparent;

- Identification of scenarios by covering the archival of version of records as well as scholarly created enrichments. For example, TEI transcriptions made by scholars could be archived back in the library where the primary source is actually situated.

As a whole, DARIAH should contribute to excellence in research by being seminal in the establishment of a large coverage, coherent and accessible data space for the humanities. Whether acting at the level of standards, education or core IT services, we should keep this vision in mind when putting priorities as to what will impact the sustainability of the future digital ecology of scholars.