Stewardship of Cultural Heritage Data. In the shoes of a researcher.

Charles Riondet

Stewardship of Cultural Heritage Data

In the shoes of a researcher

Charles Riondet

9 avril 2018

Inria
charles.riondet@inria.fr

Outline

Stewardship ?

Extrapolation from a real life example

Challenges

How can the Charter help ?

Stewardship ?

Definition

Principle :
Long-time preservation, persistence, accessibility and legibility of cultural heritage data should be a priority.

Commitment :
Cultural Heritage Institutions, Researchers and Research Institutions will take the necessary steps and precautions to guarantee long-term stewardship of the original item or record and the resulting research. Whether primary digital surrogates or further representations, forms or enrichments, the various parties involved in the creation and curation of cultural heritage data, will ensure a proper hosting and preservation of all contents.

Definition

Principle :
Long-time preservation, persistence, accessibility and legibility of cultural heritage data should be a priority.

Commitment :
Cultural Heritage Institutions, Researchers and Research Institutions will take the necessary steps and precautions to guarantee long-term stewardship of the original item or record and the resulting research. Whether primary digital surrogates or further representations, forms or enrichments, the various parties involved in the creation and curation of cultural heritage data, will ensure a proper hosting and preservation of all contents.

Definition

Principle :
Long-time preservation, persistence, accessibility and legibility of cultural heritage data should be a priority.

Commitment :
Cultural Heritage Institutions, Researchers and Research Institutions will take the necessary steps and precautions to guarantee long-term stewardship of the original item or record and the resulting research. Whether primary digital surrogates or further representations, forms or enrichments, the various parties involved in the creation and curation of cultural heritage data, will ensure a proper hosting and preservation of all contents.

Definition

Principle :
Long-time preservation, persistence, accessibility and legibility of cultural heritage data should be a priority.

Commitment :
Cultural Heritage Institutions, Researchers and Research Institutions will take the necessary steps and precautions to guarantee long-term stewardship of the original item or record and the resulting research. Whether primary digital surrogates or further representations, forms or enrichments, the various parties involved in the creation and curation of cultural heritage data, will ensure a proper hosting and preservation of all contents.

What content ? Who is responsible ?

- Primary material (physical artefacts) → CHI *
- Digital surrogates → CHI & Researcher
- Metadata → CHI & Researcher
- Enrichments → Researcher only ?

∗.Cultural Heritage Institution

Hosting and preservation

For all these different contents, we have to determine :

- What to keep, under which format ?
- Who owns the rights ?
- Who manage the versioning ?
- Where is it stored ?
- For how long ?
- How is it identify ?
- What if it is, for some reasons, put offline ?

Extrapolation from a real life example

Digital edition of Léo Hamon's clandestine diary

Léo Hamon, born Lew Goldenberg (1908-1993) was a French lawyer of Russian origin, and one of the leaders of the Parisian Resistance.
His diary relates his underground daily life, reports on his comments on the course of the war, meetings he attended, the organization of the Resistance and on the preparation of the seizure of power in Paris.

Research project

Study the evolution of the discourse on the actors of the war (at every level) in the diary.

The tasks are :

- OCR the typed version (creation of a training model, several versions of the texts, output in PDF, txt,)
- HTR (Handwritten Text Recognition) the manuscript (same kind of data is produced)
- Structure of the text of the diary in chronological entries (using XML-TEI, merging the two sources)
- Annotate automatically the Named entities (Persons, organisations, places)
- Build a graph with the following elements :
  - All the Named entities mentioned (especially persons and organisations)
  - The different ways they are named in the text (nicknames, metonymies, ...)
  - How they are qualified (modifiers)

Source divided in two Institutions

Centre d'Histoire de Science Po Paris

- Part of the academic institution Science Po Paris.
- Holds archives from major modern French politicians, including Léo Hamon's.
- Manuscript version of the diary (years 1940 - 43)

Archives nationales Paris

- In the fonds of the WW2 Historical committee.
- Hundreds of testimonies and documents from 1945 to the early 80's
- Typed version of the diary (year 1944)

Outcome

Possibly a lot of data :

- Image
- Uncorrected OCR text (several formats)
- Corrected OCR text (several formats and versions)
- Metadata of the OCR process
- Structured text
- Aligned images with OCR (PDF)
- Annotations as they output by the annotation tools
- Annotations manually corrected
- Annotations aligned with the structured text
- Dictionaries, gazetteers with extracted information, possibly enriched with other sources
- Graph combining everything
- Visualization interfaces

Challenges

Challenges

Dialogue and concertation in any case, since the beginning

- Results hosted by the originating institution
- Results hosted by the research institution
- Results hosted elsewhere → agreement between three parties, new constraints.
  - Research Infrastructure like Dariah (HumaNum)
  - Repository like HAL
  - ...

Challenges

Ensure the connexion between the original data and the enrichment, and keep all the data clean

- Quality repository :
  - Metadata
  - Persistent Identifiers
  - Versioning
- Research good practices
  - Use of the appropriate standards according to the datatype (image, text, annotation, ...)
  - Documentation for all the data created/transformed

Challenges

Possible limits

A survey conducted in the summer 2017 on the Data Reuse Charter commitments outlined some potential users concerns about Stewardship :

- Time-consuming and possibly costly
- Not everything has to be stored on the long term
- CHI and researchers may put more efforts on more *valuable* data

How can the Charter help ?

How can the Charter help ?

No ready-made answers, but a frame where solutions can be designed.

- Initiate communication between Libraries and Researchers
- Distribute responsibilities
- Concrete stewardship agreement
- Good Data Management

How can the Charter help ?

Initiate communication between Libraries and Researchers : support the dialogue when several parties are involved

- How far should we go ?
- Release the data in the most appropriate format
- Allow for the reuse of the metadata
- Allow for remote and persistent access

How can the Charter help ?

Distribute stewardship responsibilities

- Allow researchers and CHI to make proposals and to make commitment
- What is the duty of the Library
- What can the researcher take care of
- Do they need to share everything ?

How can the Charter help ?

Concrete stewardship agreement

- Where do we store, and under which format ?
- How many copies ?
- For how long ?

How can the Charter help ?

Good Data Management

- Roadmap on the long run
- Creates confidence

How can the Charter help ?

No ready-made answers, but a frame where solutions can be designed.

- Communication between Libraries and Researchers
- Distribute responsibilities
- Concrete stewardship agreement
- Good Data Management