Workflow proposal

COST Action CA16204
Distant Reading for European Literary History
WG 1

February 13, 2018

Main steps

1. Selecting authors and novels.
2. Finding novels.
3. Cleaning-up and normalizing texts.
4. Annotations and encoding.
5. Publication.
6. Evaluation.

Selecting authors and novels

1. List of candidates.
   - Novels published during 1850-1920 (according to sampling criteria)
   - Web page (spreadsheet, for example) with metadata information (author, size, etc.).
   - Sources:
     - WorldCat
     - National libraries
     - Others: Appendix 1.

Selecting authors and novels

1. List of candidates.
   - Novels published during 1850-1920 (according to sampling criteria)
   - Web page (spreadsheet, for example) with metadata information (author, size, etc.).
   - Sources:
     - WorldCat
     - National libraries
     - Others: Appendix 1.
2. Selection of the appropriate novels.
   - Open to the advice of scholars and experts of each literary tradition.

2. Finding novels

To obtain the novels selected in a machine-readable format (plain text).
To look for them in digital repositories: txt, html, xml, ePub, etc.

- It must follow the sampling criteria.
- Licensed under Creative Commons or similar.

2. Finding novels

if not machine-readable format is available:

look for other digital formats: pdf, jpg, etc.

(else) if not digital version is available:

it must be digitized in this step.

2. Finding novels

if not machine-readable format is available:

look for other digital formats: pdf, jpg, etc.

(else) if not digital version is available:

it must be digitized in this step.

Problems / Questions

OCR.

Do we have access to the books (printed)?

2. Finding novels

Repository of raw texts:

- All novel in raw text could be stored in a repository.
- Back-up

3. Cleaning-up and normalizing texts

To check the texts in order to fix typos and normalize them.

- Semi-automatic process:
- manual revision will be necessary.

3. Cleaning-up and normalizing texts

To check the texts in order to fix typos and normalize them.

- Semi-automatic process:
- manual revision will be necessary.

Problem / Question

Manual revision will be a complex and time-consuming task. Do we have resources for this task? Who will do it? (Students?)

4. Annotation and encoding...

- Metadata annotation
- Annotation of the structure of the novel.

Metadata annotation

Semi-automatic annotation process

- XML structure with the metadata attributes could be created automatically from the plain texts.
  - Some metadata values could be extracted from the spreadsheet of step 1 (CSV file).
  - Annotators must review the automatic annotation and fill out remaining data.

Metadata annotation

Semi-automatic annotation process

- XML structure with the metadata attributes could be created automatically from the plain texts.
  - Some metadata values could be extracted from the spreadsheet of step 1 (CSV file).
  - Annotators must review the automatic annotation and fill out remaining data.
- Novel structure could be also annotated automatically.
  - NLP tools?
  - Ex. Prof. Jannidis "direct speech annotation tool" (?)

4. Annotating novel structure

At least a minimal manual revision will be necessary

The same problem (again)

Do we have resources for this task? Who will do it?

5. Publication

All XML files must be available in a repository (WG4).

- The repository should guarantee that novel could be downloaded as XML files or as plain text.

  Similar to `http://teipublisher.com` or `http://textometrie.org/`.

- The whole corpus could be available from the beginning, during the revision and annotation process.
- DOI from Zenodo

6. Evaluation

Two aspects must be evaluated:

- text quality (*sampling criteria*)
- annotation (*encoding guidelines*)

Double revision and annotation of a sample of novels in order to find recurrent mistakes.
Inter-annotator agreement?

Open questions

- Working groups: steps? literary tradition?
- Manual text revision and annotation? Developers team?
- Where to find printed novels. It is possible?