COST Action Distant Reading for European Literary History

WG1 Meeting

Distant [≡] Reading

Christof Schöch, Carolin Odebrecht, Lou Burnard, Borja Navarro-Colorado et al.

WG Meeting Budapest, 2019-09-24

Outline

1. Welcome and schedule

2. Introduction to the new schema

3. Dissemination of ELTeC: strategies for archiving, documentation and distribution together with WG4

4. Zenodo

Schedule

Tuesday	
9:00-10:45	Joint opening session
11:15-13:00	Introduction of the "dev" schema and associated automatic updates of TEI Headers
13:00-14:00	Lunch Break
14:00-15:15	Dissemination of ELTeC: strategies for archiving, documentation and distribution, Introduction to Zenodo (together with WG4)
15:45-17:00	Joint session WG1, WG2, WG3
17:15-18:45	Katherine Bode Lecture (optional, all WG members are invited!)
Wednesday	
9:00-10:30	Joint closing session
11:15-12:45	Joint final session of the Training School (optional, all WG members are invited!)

Current state

ELTeC Language Collections:
`https://distantreading.github.io/ELTeC/index.html`

Outline

1. Welcome and schedule

2. Introduction to the new schema

3. Dissemination of ELTeC: strategies for archiving, documentation and distribution together with WG4

4. Zenodo

New schema – Encoding Guides by Lou

- report https://distantreading.github.io/Training/Budapest/wg1-schema-report.html
- teiHeader https://distantreading.github.io/Training/Budapest/encodingGuide-1.html
- body https://distantreading.github.io/Training/Budapest/encodingGuide-2.html

Outline

1. Welcome and schedule

2. Introduction to the new schema

3. Dissemination of ELTeC: strategies for archiving, documentation and distribution together with WG4

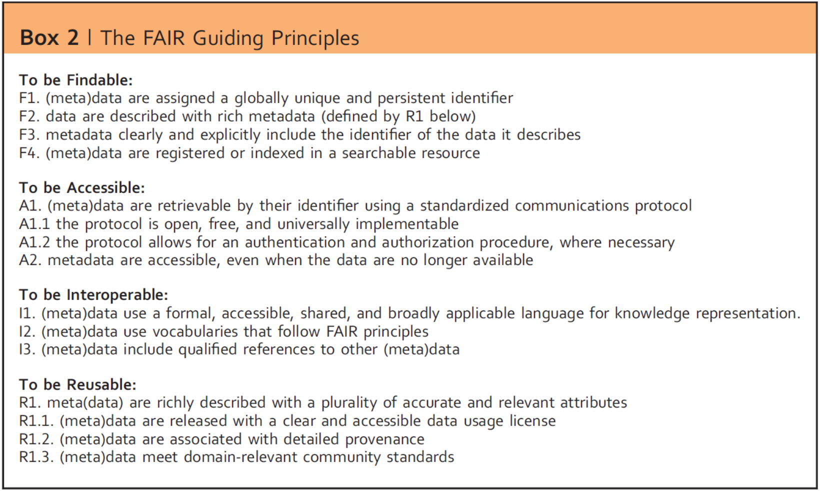
4. Zenodo

Purpose

- Evaluate dissemination strategies for ELTeC
- Zenodo for archiving and referencing purposes
- Applications to foster visibility and re-usability in research disciplines and communities, e.g.:
  - Repositories
  - Meta-search data bases /catalogue applications
  - Blogs
  - Analysis environments

→ FAIR Guiding Principles

FAIR Principles



Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

Figure: Wilkinson et al. (2016)

Current state – ELTeC

- Working environment on GitHub (upload, modify data and metadata), documentation of workflow
- Archiving and referencing with Zenodo (DOI, long-term storage, licence)
- Both generic tools (not community-specific)
    - Researchers might not search exactly for ELTeC but for historical corpora / novels etc.
    - Researchers might not know (yet) our Action *Distant Reading*
    - Tools do not provide search and visualization features
    - How can we foster the visibility and findability of ELTeC? (also fostering its reusability)

Approaches

- Use other domain-specific, discipline-specific repositories
  - Higher visibility in scientific communities
  - Foster re-usability within these communities
- Use analysis environments
  - Create or use pre-customized visualisations and analysis
  - Gain different representations for the same data set

Approaches

- Create entries in data catalogues / meta search services
    - Gain higher findability
    - Most of the services offer search and filter mechanisms
- Blogs, talks, other publications
- In general: Do not forget to inform WG4! (`https://www.distant-reading.net/news/`)

Some suggestions

- Catalogues and meta data bases for resources
    - varieng corpus finder http://www.helsinki.fi/varieng/CoRD/corpora/corpusfinder/
    - DARIAH Research collection registry https://colreg.de.dariah.eu/colreg/?lang=en
    - LRE http://lremap.elra.info/
- Discipline-specific Repositories
    - LAUDATIO https://www.laudatio-repository.org/
    - europeana https://www.europeana.eu/portal/de
    - LDC https://www.ldc.upenn.edu/
- Search and Visualisation
    - TEI publisher https://teipublisher.com/index.html
    - DramaCor: https://dracor.org/
    - TXM http://textometrie.ens-lyon.fr/?lang=en
    - ANNIS Search and Visualisation Tool http://corpus-tools.org/annis/
    - GutenTag: http://www.cs.toronto.edu/~jbrooke/gutentag/

Outline

1. Welcome and schedule

2. Introduction to the new schema

3. Dissemination of ELTeC: strategies for archiving, documentation and distribution together with WG4

4. Zenodo

Zenodo

"Zenodo is derived from Zenodotus, the first librarian of the Ancient Library of Alexandria and father of the first recorded use of metadata, a landmark in library history."[1]

- OpenAIRE Project
- "Built and developed by researchers, to ensure that everyone can join in Open Science."
- CERN, since 2013

[1] `https://about.zenodo.org/`

ELTeC on Zenodo

https://zenodo.org/communities/distant-reading/

Zenodo publication

- ELTeC Community (incl. DOI) with a repository for each language collection
  - Responsibilities: Christof, Carolin, Lou, Borja, and Maciej
- Each language collections gets a DOI
- Responsibilities of each language collection
  - Members of the action developing the language collection, e.g. Gábor and his colleagues for ELTeC-hun
- Language collections can be updated and released individually

Referencing

Different levels:

- Global: ELTeC on Zenodo community, DOI
- Language collection on Zenodo, DOI
- Any individual novel: collection's general umbrella DOI plus the ELTeC identifier (e.g. ENG18881)
- Any specific file from a specific release: collection's specific release DOI, plus the ELTeC identifier, plus the file's encoding level (e.g. level1, level2)
- include the collection's general umbrella DOI in the teiHeader

Versioning

For each language collection:

- Each version gets a version number: MAJOR.MINOR.PATCH
- PATCH: bug fixes on file level
- MINOR: add encoding level 1 or 2 to a text
- MAJOR: a collection is complete (100 novels meeting sampling and balancing criteria)

ELTeC

- The deadline for the first release of ELTeC is Oct 7!
- We will include language collections that have at least 20 or more texts for this first release.
- There will be regular releases on Zenodo.

References I

Wilkinson, Mark D., Michel Dumontier, Aalbersberg, I Jsbrand Jan, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, da Silva Santos, Luiz Bonino, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, et al. (2016). "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific data* 3, p. 160018.