DARIAH Winter School in Prague

Open Data Citation for Social Sciences and Humanities

24th to 28 of October 2016

Session 2: Open Critical Edition

2-Open Critical Edition. The Missing Link Between Digital Humanities and Open Science

Marjorie Burghart (CNRS) is a medievalist and Digital Humanist, and Emmanuelle Morlock (CNRS) is a Digital Humanist and a research officer specialised in information architecture, and Research Data Management.
This session will consist in a general introduction to digital editions followed by a practical presentation of TEI. We tried to separate the topics and to have at the same time a complementary approach.

What is a Digital edition: Some Interesting Examples

- We can start with this Google Book. It is a scanned volume from a famous series of 19th century books, the *Patrologia Latina*, a major collection of latin works. These volumes are now available and 200 of them are on Google Books.

But is this is a *digital edition*?
=> Some answers: It is digital and it is an edition. Or it can be used as digital edition by scholars.

Is it used as an *edition in digital form*?
=> Different types of users have different definitions, but there are more and more strict criteria. People speak about *digitised edition* for such type of material. This is an *edition*, *digitised* (it became digital) and it is a little bit *critical*. But this is not what most people today define as a digital edition. There is more and more reflection on this topic. See for instance RIDE: A review journal for digital editions and resources, from a German center based in Cologne. This journal produces a *review of digital edition projects* with great criteria for reviewing. It shows the state of the art criteria for the best practices in digital editions. One of them is the distinction between *digitised edition* that is just a scanned book put online and the *native critical edition* which is meant to take advantage of all the perks of the Internet connected data. There are a lot of criteria to achieve but it is more of an ideal to reach, to tend towards, that can be used as a benchmark for projects.

- Another example is this text from the Corpus Corporum: It is in fact the same text as the digitised edition we have just seen on Google Books, but presented in a searchable corpus of Latin text.

Is it a *digital* or a *digitised edition*?

=> It is actually debatable because it is only the text of the 19th century edition which has been OCRed and put online, there is no extra work. So it is a *print-born edition* which has been *digitised* into text mode, *structured* and put online, but there is no added value beyond that, except the fact that it is easier to search and that you can reference it.

- Edition of the poems of Anne Finch

It is the digital archive of her work, a very nice work: the layout is really pleasant to read, there are a lot of features, you can access all versions of the poems, you can access the sources material, images, etc.

Is it a *digital* or a *digitised edition*?

=> This is a *digital edition*: it is born-digital, it takes advantage of more than just full text because you have different media which are linked, etc. but I think that if you run the RIDE criteria on this archive, it would not get the highest score, because it is a limited in the way you can take advantage of all the data that have been gathered. For instance, I have performed a search query on the corpus and all I get is some kind of interface. When you do a Google citation search on a website, it is the same thing, it just searches for a string of characters, you have no combination. So it is very basic in the way you can take advantage of the material. You cannot download the source work if for instance you wanted to integrate these poems into a corpus of poetry from the same period. It is a very valid scholarly work but it is "*self-contained*", not connected with the outer world. It is digital but not open yet. It is open access: you can access it, everyone can access it, there are no barriers; but it is not yet open data: you cannot access the data underlying the scholarly work, reuse it and connect it to different things. This is a step further.

- Another kind of edition: Map of London

It represents a 17th century map of London which has been edited just as a text can be edited with interesting features: for example, you can highlight all the churches with an overlay layout on the map and you can zoom until reaching a single building, for example St Paul's Cathedral. You also have a text explaining what it is and you have a list of all documents in which St Paul cathedral is mentioned.
So you have, around the map, a library of edited documents with links between the map and the documents. Here you have the step further: linked data within the sub-corpora of the website, and you also have references of place from and to you can link from other projects. Here we enter the world of connected data: when you click on a link, you have the transcription of the document in which you find the highlighted term (that brought you here).

- Plaoul Commentary

It is an edition, the interface where you can read the scholarly digital edition. The author is Jeffrey C. Witt, a US medievalist. He has a vision: he sees critical editing as building a huge database with assertions on the editions and a lot of annotations. So he modeled precisely all these pieces of information and annotations and on top of that he builds printed editions, but also a workspace for an edition. It is a real complete environment and it is open because you can see the corpora and view the underlying data on Github. It is also a big database of precise data annotation with a service to query in this huge database: you can see relations and properties, in a formalised way. And that's not all, as you can also see the images of the manuscripts, it is important to note that he didn't digitise the manuscripts himself, and in his editions, he didn't have to keep a copy of the images on a server. With the properties of linked data, he just accessed the manuscripts images that are published as linked data by the institution that keeps the manuscripts. There is a protocol for images that is now widely used (IIIF). It is a new model that allows you to build your edition, your

transcription, your view of this text on top of some images that you don't curate at all. There is a visualiser and you can also built, as a researcher, a critical editor. If the images are in different places, you can build on top of it your workspace, your interface, just with links. As previously said, this is a great workspace with statistical tools, for example we can have the frequency of use of biblical quotations.

Group exercise with a poem

North of Everywhere, Helen Mort:
http://www.manifold.group.shef.ac.uk/issue7/HelenMort7.html
=> Goal: Think how to approach this document if you had to make an edition of it, from your background: what would you consider important to underline, to be able to share it with other people so they understand the document and take benefit from it, with of course a particular attention at what strategies to open the data.

Group presentation & paperboard
=> Synthesis: Groups had different approaches but a lot can be connected together. What the text is and how it does function in itself, with context and metadata? Some others had already in mind how the edition will operate in a broader system with API, as a kind of technical functioning. We heard also about intertextuality and expression of the relations with other words. What is interesting is that at the beginning of the analysis, you have to take into account the ecosystem in which you will publish and what you want to do with it (maps, etc.).
The context of your aim influences the decision about the representation. The question is can we represent all that in a practical way? Of course we can, but in the economy of a project you have limitations (money, resources, time). So you will have to list all the possible features and make choices. You can have an Interface, a displaying device, on top of digital data organised as a system.

Text Encoding Initiative

How can the Text Encoding Initiative help to prepare digital editions and encode text? Critical editions are an important part of Digital Humanities and the TEI allows to encode a text and take advantage of this encoding with an attention to open data.

Critical edition

Digital Humanities are everywhere: you can practice them by making bibliographic searches on databases, on Google, etc. You can search manuscripts, read books in digital libraries, you can generate reports or use corpora to identify the sources of a text. You can use computer assisted collation or stemmatics. Collation is the process of collecting all the witnesses, manuscripts or editions of the work you are editing and to compare them to see how the text differs, its variance. Once the collation is created, you have to try to determine what is the "genealogical tree" of the witnesses of the work, to try and see which one has been copied on which one. This is called a stemma.

TEI

You can also use digital tools to structure and analyse the edited text with TEI as it is a common frame to analyse and structure text, especially text from the Humanities, from historical and linguistic sources.
The first reason to use TEI is that you have the TEI guidelines, you can share something that a lot of scholars used for the past 40 years, it is “battle-tested”. It is not something you can think out yourself and decide it is ok for everyone else, you have to discuss encoding options with many different types of scholars from different fields to reach an agreement. The TEI helped to find common ground from different fields and scholars around the world, in order to share a same model of information for text. Besides, TEI makes it easy to differentiate between the aspect of a book and its analysis. This aspect is important, specifically if you are working with ancient documents, medieval or epigraphic, but also with contemporary digital-born documents. It allows to report that there are for example three lines in a particular place in a document. And it is also possible to add references to “Isabelle” for example. It is important to hit both sides of the document: the physical aspect but also the meaning. Finally, it is a good way to be completely free from proprietary formats (pdf, word). Otherwise you are completely tied to the format used for your file and you have no warranty that in the long term it will be preservable.

The Text Encoding Initiative is:

- Human-friendly rules for modeling the text: TEI Guidelines. In printed version, it is more than 1 000 pages because it covers a huge range of texts - you don’t have to know everything if you want to do a particular work.
- Computer-friendly way to implement the rules of the Guidelines through an XML schema. Guidelines are written for humans and the schema applies the same rules as developed in the guidelines, but for computer programs.
- Community of users that can provide support in different ways. It can be advices, discussions about your own issues and also software that has been prepared for other projects but in a generic enough way to be useful to others, sharing the same formalism, the same TEI modeling. It saves time and gives a better insurance for quality of reflection.

XML

In a way, it is really close to html. If you look at the code of a webpage, you can see tags, etc. XML is basically the same principle, you have tags, except that html is a closed vocabulary and xml is not, it is extensible. The rules are stricter than in html, it has to be a tree structure. TEI XML has the advantage of full text plus a database, so you don’t have to choose between transcribing on one side and creating a database on the other side. You can have both together if the data analysis links into the text. It permits to retrieve text mentioning data and vice versa, access data pertaining to the text.

For medieval writing

Diplomatic edition: it follows strictly the aspect of documents, for example you do not expand abbreviations, you respect the layout of the document, etc. Otherwise you have the Transcription for research purpose, where you can expand the abbreviations for a better readability.
Here you can have both, a versatile document, a record of all the data about the aspect of the document, the diplomatic view; and a record of all the analytic data. It is possible to have two views of a document, one is a diplomatic view where you can see which words were abbreviated and what is the expanded form; and another view where you can see what are the sections of the document, like chapters, that strictly define certain parts. It is interesting for researchers to be able to search and extract different types of parts from a corpus.
Here, the reader also has options! Classically, the editor makes all decisions once and for all. Now you can have a system allowing users to choose their options. They might be interested in mixing different kinds of visualisation with scripts that produce a webpage that readers can use.

Inside TEI

The key idea is that it is not just TEI or just XML, it is a family, a constellation of technologies that work together to work some magic in the end. It starts with XML that has the role of *representing* the text. With XML you describe your data with *tags* that can be qualified with *attributes* and you have to produce a *tree structure*: one *root* for your document and this root has several *children* which may also have children. This is the only golden rule of XML.
An example of XML source:
It begins with a declaration and then starts with the XML itself. The root of the tree structure is <text> and the children of this root are <p>.

```
<?xml version="1.0" encoding="UTF-8"?>
<text>
  <p n="1">I am reading a book by   <persName>Jack London</persName></p>
  <p n="2">I live in   <placeName>London</placeName></p>
</text>
```

- Controlling the text: TEI schema

The TEI provides the rules for structuring the document beyond the rules of XML. This schema is the implementation of the TEI guidelines, from a human-readable version. There are several TEI XML schemas and people can create sub-schemas based on the TEI, but only using a sub-part of the TEI. They can extend the TEI if they want, for example, to make an edition of a musical text, they need another format description of the music model. The model can be extended or reduced.

- Displaying the text: CSS and XSLT

You can display what you have encoded and transform it into something else. CSS is a web language that you apply to XML pages. XSLT is more developed and powerful, it can

transform data in a web page for example or in a different type of XML, or extract and transform data from your XML into RDF or JSON and share it.

- Querying the Text: XQuery and XPath

Defined by W3C recommendations: XPath is commonly used within XSLT and XQuery is more complex and more powerful as it allows to query data and structure together, so you can extract all the words in a particular part of the document, for example.

Open science

In fact, the main goal is to prepare data before publishing it in a way that machine can understand.
=> Opening the principles of open access, of openness to the whole cycle of research.
To explain this, we have to see some principles: *semantic web* and *Linked Open Data*.
Then we will come back to TEI to see how to interconnect TEI files with this web of data that are linked and exposed by machine.

- Giving access is not sufficient to research data and publication.
- *Open Access*: Free and persistent access to research data and publications.
- With *Open Access*, it is more about an access for the reader. So when you have a huge volume of information, how can you read it?
- *Open Data:* Files made publicly available by official organisms for re-use.
- *Open Process*: Right to openly observe the underlying data and workflows of research project.
- Openness also influences research as way of improvement as the underlying data are accessible. As we saw, if we just show you the result of an edition, you don't really understand what is at stake, what is the work of interpretation that has been done. In order to validate the scientific work on an edition, you also have to look in the underlying data. The workflows are also very important to be documented.
- *Open Science*: Free and persistent access to research data with the right to observe openly these data with digital tools.

=> Open Science = Open Access + Open Process

It means that it is not only the readers that can access the research and analysis but also machines. To do that you need data to be expressed in a particular way.
The difference with TEI and the schema is to know the meaning of the tags, a machine can parse it and build an interface, but the machine has to know the schema. And it is not always the case because the schema is inside the edition, even if we have standardisation, it is not enough to have a broader interpretation of the data. TEI allows to have data and with our interpretation.

Semantic Web

The *semantic web* is like a parallel web that differs from the original web by the kind of knowledge presented and accessed.
The knowledge found on the semantic web is *formal* knowledge with:

- a machine readable notation

- a formal syntax
- a formal semantics with inference mechanisms

The Semantic Web started as a vision by Tim Berners-Lee and became true via *Linked Data*.

=> Open Data + Linked Data = Linked Open Data (LOD)

The idea is to share machine-readable and interlinked data that are on the web with two aspects:

- A language aspect: how to interpret data that are in documents or in web pages
- Interoperability aspect: how to understand all this without referring to a schema

So it is a system to identify resources where everything is a resource.

Linked data

Design principles for sharing machine-readable interlinked data on the Web:

- Name resources with unique identifiers (URIs)
- Use the architecture of the web to get some information about theses resources (http)
- Use a standard model to give information about these resources (RDF)

RDF: Resource Description Framework

It expresses information about identified resources with very simple sentences and composed of three elements:

- a subject: identifying the resource
- a predicate: identifying a property of the subject
- an object: identifying the resource linked to the subject by the property

Ex. Helen Mort (subject) --- is the author of (predicate) --- the poem “North of everywhere” (object)

The result of the aggregation of triples is a graph and the specificity of this information model is that:

- relations are part of the data
- each triple is autonomous, complete, persistent
- a distributed model

TEI to LOD

TEI explicates the data but not exactly the relations. The relations expressed in the hierarchy.

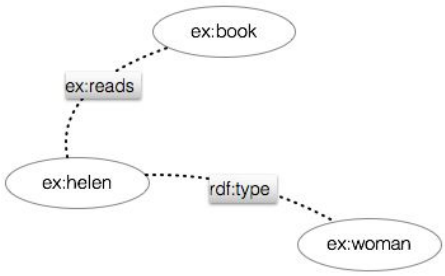
In the metadata, you have the title statement (titleStmt) and an author with a reference to the URI of the dbpedia page of Helen Mort. Dbpedia is the database made with wikipedia articles and facts extracted and transformed into a database which is accessible by humans and machines:

```
<titleStmt>
        <title>North of Everywhere</title>
        <author ref="http://dbpedia.org/resource/Helen_Mort">Helen Mort</author>
</titleStmt>
```

It is possible to extract the triples: the text represented in the TEI document has a title; the title of the text is “North of everywhere”, Helen Mort is the author of the text.. You have to select what could be interesting for others and express it in the RDF formal language to expose it and to make it available. You can also have “Hermaness” as the English (attribute) name of a place; this place is identied by the URI http://dbpedia.org/page/Hermaness, the longitude of this place is “60.837222”.
=> The sum of the triples produces a graph and the “magic is also done by the XSLT”

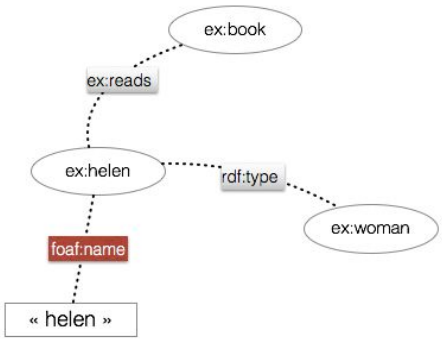
Step of conceptualisation: it is a point of view on the reality, ex: two resources: Helen (a woman) and a book; relation: she reads the book.
Step of language selection: where the URIs comes into play; pairs of resources are connected by the binary relation they belong in: ex:helen ex:reads ex:books; Unitary relations are connected to a class: ex:helen rdf:type ex:woman

A set of RDF triples is a graph



ex:helen ex:reads ex:book
ex:helen rdf:type ex:woman

Literals to associate a natural language fragment to a resource



Linked Open Vocabulary
Foaf (friend of a friend): It is a vocabulary that gives a property to the described relations between person. There is common vocabulary to prepare possible relations that some users will then activate, the catalogue of open vocabulary: https://lov.okfn.org/dataset/lov/.
On Helen Mort's page on dbpedia, there are information on triple you can find with for example: sameAs. You can find further information with the Linked Ancient World Data Institute.

DBpedia Demo: Basic exploration of a RDF graph with simple SPARQL queries

Two simple sentences or assertions:

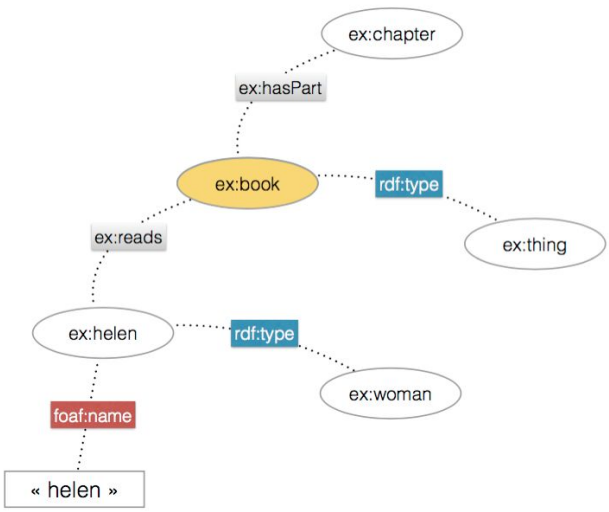
- "Helen" "reads" "a book"
- "Helen" "is" "a woman"

In RDF, with the prefix "ex:" we have our identifier:

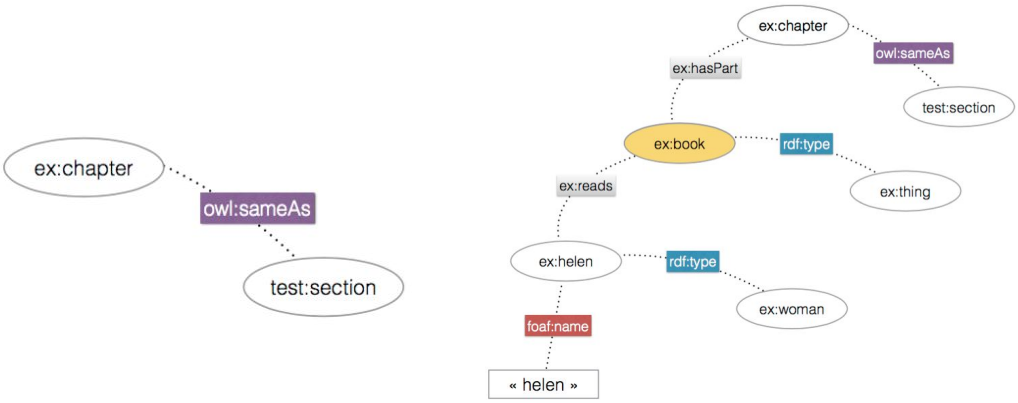
- ex:helen ex:reads ex:book
- ex:helen rdf:type ex:woman

=> subject, object and the relation

It is here a unitary relation, this means that it is the class of the resource. The three elements are resources.



Here you can express that it is the same thing and make the aggregation function. This is why in linked data publishing practices, it is highly recommended to be generous and to try to align with "sameas" as much as possible your data.



SPARQL with DBPedia

DBPedia is the RDF graph built extracting all the data that is curated in Wikipedia. When you are human you see an html page, when you are a machine you see a RDF file for the same information. There is, I guess, a duplication of the database that you can directly query with a dedicated interface with the SPARQL language.

Simple Protocol And Query Language A query has a structure:

- SELECT distinct * == select all resources
- WHERE { } == the query
- LIMIT, GROUP BY, ORDER BY...

Simple queries 1

Find resource with the English label « Prague »
Answer:
http://dbpedia.org/resource/Category:Prague ===>
http://dbpedia.org/page/Category:Prague
http://dbpedia.org/resource/Prague ===> http://dbpedia.org/page/Prague
With this you find the name of the resource and can use it for further queries.
Find all the properties of this resource
Find the types of this resource
Choose a type (ex. "?o"" for object)
Find the resources with the type

Simple queries 2

select distinct * where {?s rdfs:label "Prague"@en} LIMIT 100
select distinct * where {<http://dbpedia.org/resource/Prague> ?p ?o} LIMIT 100
select distinct * where {<http://dbpedia.org/resource/Prague>rdf:type ?o} LIMIT 100

select distinct * where {?s rdf:type <http://dbpedia.org/ontology/PopulatedPlace>} LIMIT 1000

This is a good to explore material when the relations are precisely defined. And you can start building a database without having in mind the whole schema, it can be flexible and adapted.

RDF is much more fact oriented and TEI is more precise to express subtle documents and gather a lot of precise annotations and distinctions. But they can work well in collaboration.

One of the key stakes of Linked Open Data is the quality of data and TEI is really good, it can be like a database where you keep all your scientific information and then extract some datasets in RDF or other language, in a continuous work of repackaging your data for different purposes.

CORESE

- Simple inference in action with Corese, a Semantic Web Factory (triple store & SPARQL endpoint) implementing RDF, RDFS, SPARQL 1.1 Query & Update, developed by INRIA: http://wimmics.inria.fr/corese
- Tutorial: http://wimmics.inria.fr/node/34
- Linked Data Navigator using Corese and SPARQL Template Transformation Language: https://corese.inria.fr/

What is the best way to share a good body of generated RDF?

EM: The best way is to find the appropriate data repository, one that is certified (Data Seal of Approval) and expose it here. If you want people to actually use it, I would do a **data paper** explaining concisely where the data come from and the context (technical but not only) that other researchers would need to use it for another research. All things that are obvious must be explicated. Like this, you delegate the stewardship of the dataset and you give all information and associated metadata.

Conclusion

Knowing that it will influence the way you prepare you text with TEI and at the same time, it opens to the notion that these **triples are not a technical thing, it is an editorial thing**. You have to decide which are the **interesting triples in a text and for a community**. **This is a new task of the publisher: to design. If you consider publisher or editor as a designer of information artefact, these RDF exposition of data must be editorialised and designed.**

Useful links

- Dbpedia example: http://dbpedia.org/resource/Helen_Mort
- DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia: http://svn.aksw.org/papers/2013/SWJ_DBpedia/public.pdf
- https://lov.okfn.org/dataset/lov/
- http://www.foaf-project.org/

- http://dublincore.org/documents/2012/06/14/dcmi-terms/?v=terms
- Using SPARQL to access Linked Open Data from SSH perspective: http://programminghistorian.org/lessons/graph-databases-and-SPARQL#searching-rdf-with-sparql
- To see more refined uses of sparql queries in combination with nice displays for the result, watch that youtube video about wikidata (16 min): https://www.youtube.com/watch?v=1jHoUkj_mKw

Contact

Marjorie Burghart & Emmanuëlle Morlock, CNRS

Marjorie Burghart is a research fellow at the CNRS (French National Center of Scientific Research) and she is working in the CIHAM UMR 5648 research center in Lyon and is specialised in medieval history and computer science. She is an elected member of the board of directors of the *Text Encoding Initiative* (TEI) consortium, the scientist in charge for the EHESS partner of the Erasmus SP+ *Digital Edition of Medieval Manuscripts*, and also the scientist in charge for the EHESS partner of the DIXIT (*Digital Scholarly Editions Initial Training Network*) Marie Curie european project. She has published several papers and softwares, and is involved in differents projects of electronic edition of medieval documents in TEI format.

Marjorie Burghart's website: http://marjorie.burghart.online.fr/?q=en

Email: marjorie.burghart@gmail.com

Emmanuelle Morlock is a digital humanities research officer at the French National Center for Scientific Research (CNRS) and currently works at HiSoMA, a research center dedicated to Archaeology and Philology of the Ancient Worlds. Her main mission is to assist researchers in their application of information technologies and solutions for scholarly editions of ancient texts and inscriptions. Her activities include project ownership assistance and technical implementation of online publications (metadata modeling, definition of encoding strategies, TEI framework implementation, information architecture and digital curation of research data). She was educated in France where she studied French literature and received a Master's Degree in Information Science and Documentation from SciencePo Paris.

Twitter: @emma_morlock

Email: emmanuelle.morlock@gmail.com