From DOCX via TEI to Literature Map

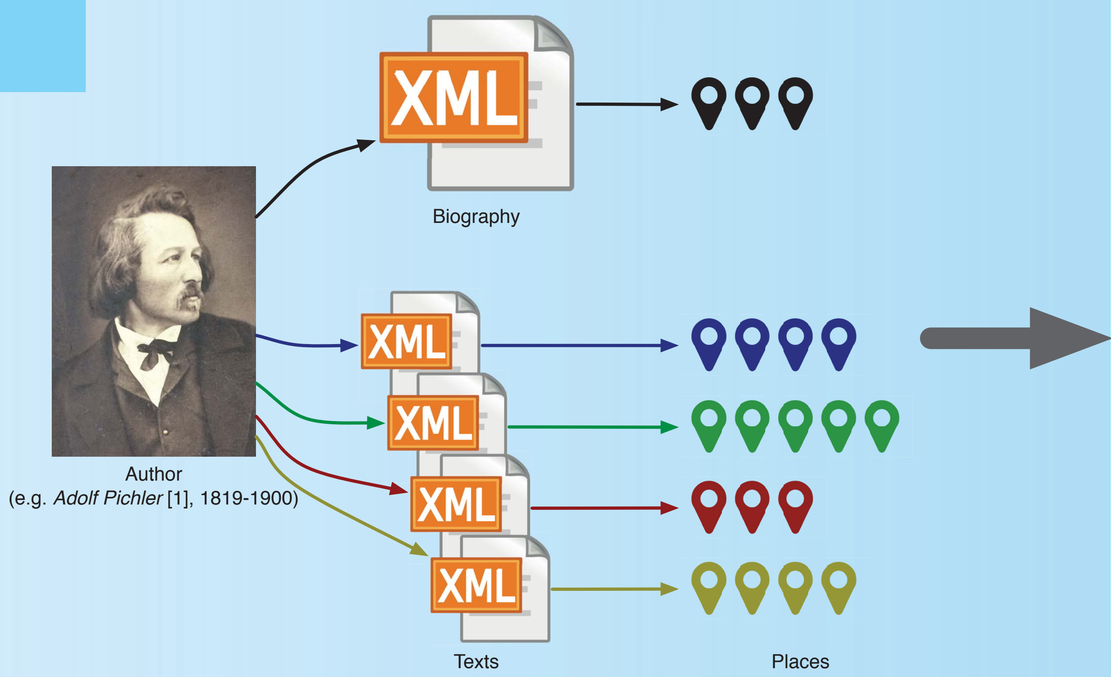
A presentation at the *TEI Conference and Members' Meeting 2016* in Vienna, September 2016, by Joseph Wang, University of Innsbruck

Aim of the Project

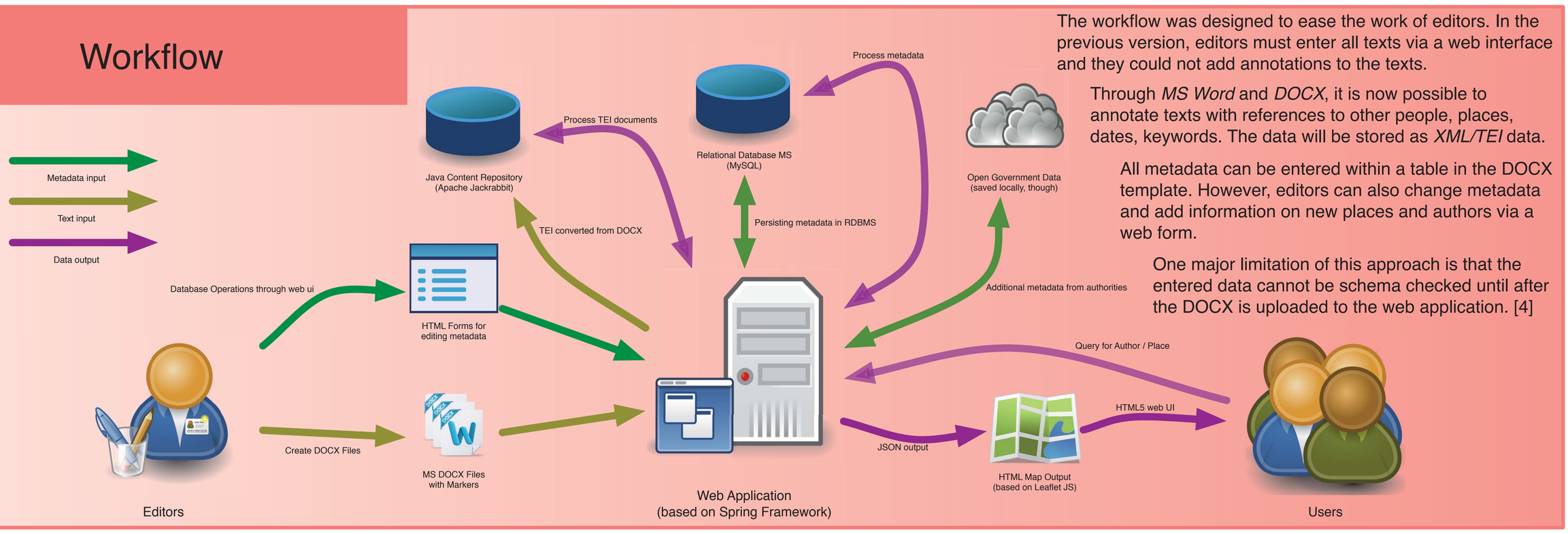
The aim of the project is to create an input system to feed the *Literature Georeferencing Database*, so the software is able to generate Literature Map out of the data.

A *Literature Map Software* must be able to:

1. create a geographical map given an author showing georeferences in his life and in his works.
2. create a geographical map given a work showing georeferences in this work
3. create a geographical map given a geo-location showing other georeferences linked to this location through literary works [3]



Workflow



The workflow was designed to ease the work of editors. In the previous version, editors must enter all texts via a web interface and they could not add annotations to the texts.

Through *MS Word* and *DOCX*, it is now possible to annotate texts with references to other people, places, dates, keywords. The data will be stored as *XML/TEI* data.

All metadata can be entered within a table in the DOCX template. However, editors can also change metadata and add information on new places and authors via a web form.

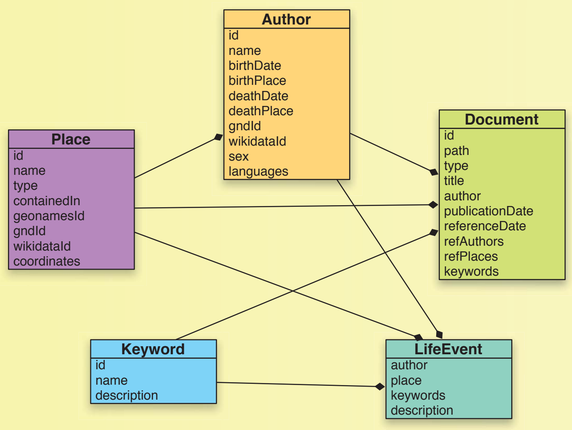
One major limitation of this approach is that the entered data cannot be schema checked until after the DOCX is uploaded to the web application. [4]

Data Modelling

Rational behind the data modelling:

1. The researchers should be able to enter large amount of data quickly.
2. We apply DRY-Principle (*Don't Repeat Yourself*) as good as possible.
3. Maps are generated from the database, and not drawn by the scholars using GIS tools.

Thus, we come up with this solution: For the metadata, a web frontend for database operations is created; for the TEI data, we found a solution based upon DOCX to TEI Conversion. Furthermore, queries on the authority givers (*GND*, *Geonames.org*, and *Wikidata*) is made during the data ingest, thus the database entities are linked to other authorities. However, the same cannot be done with open government data made available by the local government of Tyrol and South Tyrol; the reason: these data can only be downloaded but not queried through a web service. Therefore, the data is stored locally and queried during the data ingest.



Main object classes in the new data modelling

Map Generation:

Places have coordinates, they can be plotted in a HTML map. Using *LeafletJS*, one can embed interactive maps within a web page. Different colors and different icons mark different meaning of a place.

Benefits of this modelling:

1. Captures many aspects of use cases (e.g.: queries on objects are just queries on databases.)
2. Database operations on metadata of objects can be done quickly.
3. Inserting and Updating texts can be done easily.
4. Data models reflect their usage in humanities research.

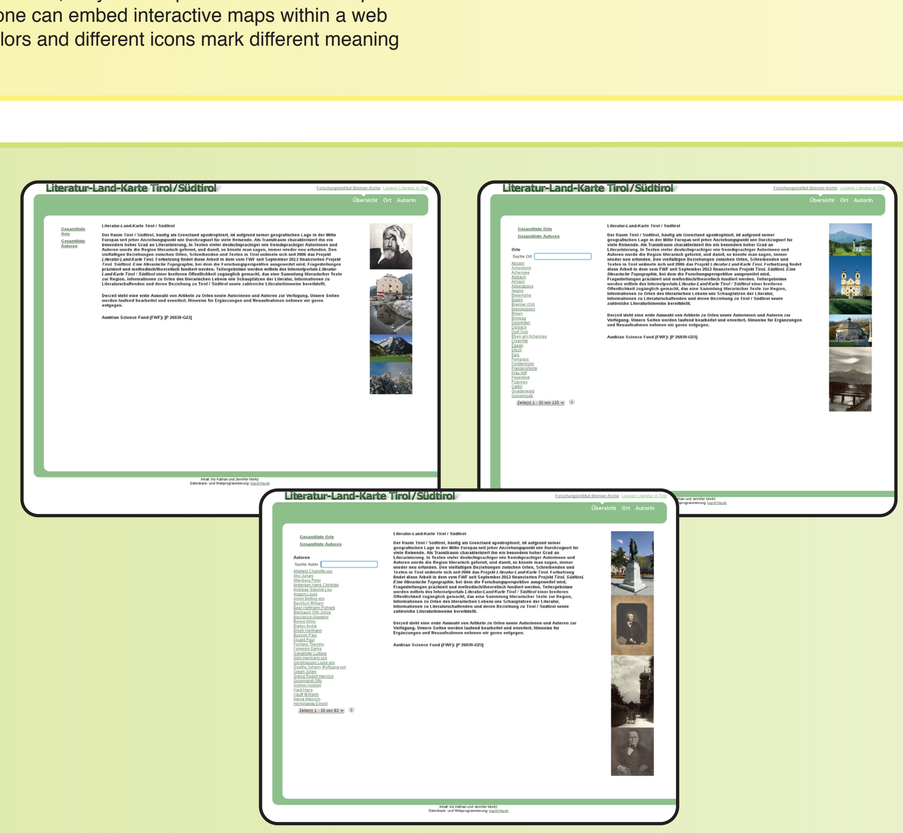
Drawbacks of this modelling:

1. Complicated database scheme is difficult to maintain.
2. Redundant data storage, no back-propagation of changes in RDBMS to TEI Document.
3. Dependencies on data of other resources.

Having the relationships between authors, works and geographical places in focus, the Project *The Tyrol / The South Tyrol - A literary Topography* (FWF P26039) created a database on these entities.

At the beginning, data modelling only captures basic metadata of the entities and their relationships. However, it was impossible to annotate texts, thus scholars could not tag the actual section of the texts with additional references.

After careful consideration, it was decided to redo the data modelling: While the metadata should still be kept and maintained by a RDBMS (in our case: *MySQL)*, the actual texts (primary sources, biographies and description of places) should be modelled as XML/TEI data and kept in a repository (in our case: *Apache Jackrabbit)*. Furthermore, we want to reuse authority controlled data and open government data.



Screenshot from the Database *Literatur-Land-Karte Tirol/Südtirol* (access on 2016/09/15)

Starting Point

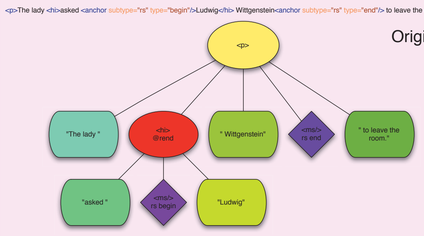
Convert comments marked using milestones to <rs> in 3 steps [5]

A simplified example withiout attributes

The lady <hi>asked <anchor subtype="rs" type="begin"/>Ludwig</hi> Wittgenstein<anchor subtype="rs" type="end"/> to leave the room.

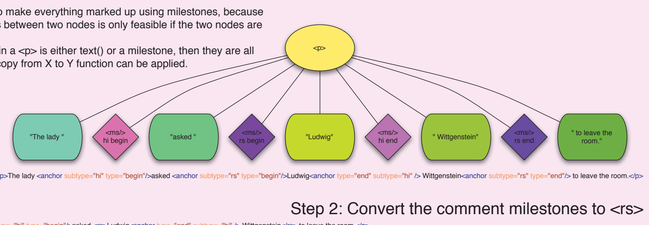
Origin: Using milestones to encode boundaries

After transforming the initial docx with the customized Docx2Tei-Stylesheet the boundaries of a comment is marked using milestones, while the highlighted passage is marked up within a <hi>.
As one can see, the two milestones marking the beginning and the end of a comment do not have the same parent, i.e. they are not siblings.



Step 1: convert all elements to milestones and texts.

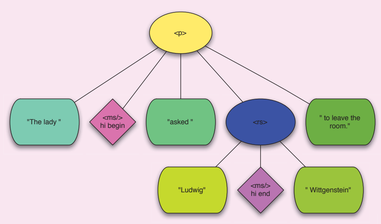
The first step is to make everything marked up using milestones, because copying all nodes between two nodes is only feasible if the two nodes are siblings.
If everything within a is either text() or a milestone, then they are all siblings and the copy from X to Y function can be applied.



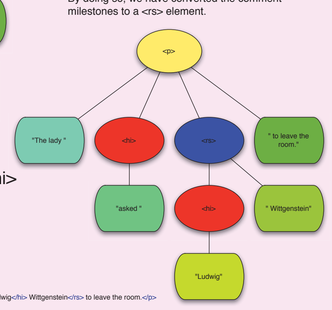
The lady <anchor subtype="hi" type="begin"/>asked <anchor subtype="rs" type="begin"/>Ludwig<anchor type="end" subtype="hi" /> Wittgenstein<anchor subtype="rs" type="end"/> to leave the room.

Step 2: Convert the comment milestones to <rs>

The lady <anchor subtype="hi" type="begin"/>asked <rs>Ludwig<anchor type="end" subtype="hi" /> Wittgenstein</rs> to leave the room.



Since all nodes are now siblings within the we can safely transform the nodes like this: copy a node if it is not between the beginning and the end comment milestone. When you see the beginning comment milestone, make a <rs> and copy the whole siblings between the beginning and the ending comment milestone as children of <rs>, otherwise just do nothing.
By doing so, we have converted the comment milestones to a <rs> element.



Step 3: convert the remaining milestones back to <hi>

Now copy the result of step 2, but test for each text() whether it is between two milestones marking up the boundaries of a <hi>. If it is, make a <hi> and copy the text into it, if it is not, just copy it.
By doing so we have doubled the <hi> to the child- and "grandchild-"axes of .

The lady <hi>asked </hi><rs><hi>Ludwig</hi> Wittgenstein</rs> to leave the room.

Parts of this project are supported by grants from the Austrian Science Fund FWF and the Government of Vorarlberg. The project benefits from results of other projects carried out by the Commission for Modern Austrian History.

This project is carried out in the Research Institute Brenner-Archives of the University Innsbruck.

References

[1] From the literary estate of Adolf Pichler in the Research Institute Brenner-Archives, sig. 4/3.1.1..
[2] Map generated from data taken from the OpenStreetMap Foundation. See: https://www.openstreetmap.org/ (2016-09-15).
[3] Icons taken from:
xml document icon: by AnonMoos, under https://commons.wikimedia.org/wiki/File:Text-xml.svg (2016-09-15)
map marker icon: by Kondicherry, under https://commons.wikimedia.org/wiki/File:Map_marker_font_awesome.svg (2016-09-15)
[4] Icons taken from:
map icon: by theway, under: https://commons.wikimedia.org/wiki/File:Map-icon-pure.svg (2016-09-12)
web ui icon: online resource under from: https://pixabay.com/de/men%C3%BC-gui-schnittstelle-vorlage-ui-303121/ (2016-09-12)
docx icon: by Renato Veras, under https://www.iconfinder.com/icons/65890/documento_docx_file_microsoft_word_icon (2016-09-12)
all other: by Open Security Architecture, under http://opensecurityarchitecture.com/cms/library/icon-library (2016-09-15)
[5] Taken from a poster by the author, presented during the workshop "Datenmodellierung in digitalen Briefeditionen und ihre interpretatorische Leistung: Ontologien, Textgenetik und Visualisierungsstrategien", Berlin May, 2014.

Joseph Wang
joseph.wang@uibk.ac.at
Research Institute Brenner-Archives /
Institute for Historical Research and European Ethnology
University of Innsbruck