Web Service for Easy Text-to-TEI Normalization and Metadata Creation

Bart Jongejan
University of Copenhagen, Denmark
bartj@hum.ku.dk

Lene Offersgaard
University of Copenhagen, Denmark
leneo@hum.ku.dk

Dorte Haltrup Hansen
University of Copenhagen, Denmark
dorteh@hum.ku.dk

Abstract

In CLARIN-DK we experience that it is too difficult for users to create data and metadata in the specific formats required in the repository. As most of the deposited resources in CLARIN-DK are text resources, we have decided to make a web service that assists researchers in the preparation phase of text resources. The aim is on the one hand to transform text to the TEI P5 format and on the other hand to create sufficient metadata for the resource – from the point of view of the repository as well as that of the researcher. To reach out to many scholars with a solution to these problems, the information available online has also been extended with tutorials. The goal is to make providing data and metadata of acceptable quality much easier.

1 Introduction

Since 2012 researchers have had the opportunity to deposit resources in the CLARIN-DK repository (Offersgaard et al. 2013). The deposit service only accepts resources and accompanying metadata in a number of specific formats based on widely used standards, aimed at maximal usability in a wide range of current and future research settings. The formats for data and the required metadata have been specified in collaboration with researchers from different cultural institutions and research groups. These partners have also delivered resources to be deposited.

About forty-thousand texts have been deposited, but to date, no individual researchers have made the conversions of their data and metadata that are required before depositing can be undertaken. Instead, the preparation of all these resources for depositing has been done by the CLARIN-DK team. During the last years dialogues with researchers in humanities have revealed that providing data in the required formats and defining metadata is too hard for many of them. It is much more difficult to supply data with metadata in a consistent way than expected (Hansen et.al. 2014), and defining metadata for resources and converting resources from one format to another is a challenging task for many researchers in humanities.

Much of the textual data that researchers might want to share with others is not in one of the formats that CLARIN-DK accepts as depositing format. Converting a resource to the required format involves a number of steps, requiring a mix of skills that few have. Some resources inevitably require scholarly diligence, for example if a text is manually annotated in a free style, or if a text contains elements that the researcher does not want or is not allowed to share. However, after this manual work, one or more steps follow that require technical prowess. To assist scholars with these technical steps, we have created a web service that makes providing data and metadata of acceptable quality much easier and we have extended the online information about this subject.

In continuation of the text cleaning functionality that was implemented during the DASISH infrastructure project[1] we have created a workflow for converting different text formats according to the TEI P5 schema that is required by CLARIN-DK.

[1] http://dasish.eu/publications/projectreports/DASISH-D5-3-II.PDF

2 Related Work

2.1 Text conversion

There are a number of document conversion tools that can produce TEI output, but currently not according to the TEI P5 schema that is required by CLARIN-DK. SaltNPepper (Zipser & Romary 2010) is a tool for converting digitised texts between a number of formats, especially from TEI to formats that are designed to handle a wide range of linguistic annotations, such as TCF 4, TigerXML, CoNLL and the Penn Treebank format. OxGarage[2] is a web service that can transform documents beween a number of formats. Most transformations use the TEI format at some stage. PAndoc[3] likewise converts between a large number of formats, among which formats for technical writing, such as markdown.

From the perspective of CLARIN-DK these systems also have the disadvantage that conversion from images containing text, from PDF and from presentation formats (PowerPoint, OpenDocument presentation) to TEI is impossible.

2.2 Metadata creation

Creating metadata in a specified format is not an easy task. Different tools have been created to support this task, e.g. COMEDI (Lyse et.al 2015) that enables users to create and edit metadata in various CMDI-formats. We see COMEDI as a very usable and flexible editor for users with some experience in metadata creation, but too advanced for beginners, and for researchers who only rarely create metadata. In META-SHARE an easy to use metadata editor allows users to create metadata for various resource types, but these metadata are expressed in an internal format only used in the META-SHARE repositories. The focus in our approach is to enable creation of metadata in the TEI P5 format used in the CLARIN-DK repository in an easy and user friendly manner. TEI P5 is chosen as it is a well-documented and widely used standard in the text research community.

3 Text preparation service for easy conversion of text to TEI P5 format and creation of metadata

The main goal of the web service is twofold: It transforms uploaded text to the TEI P5 format and it asks the user for an amount of metadata that is easy to provide and at the same time informative enough for other researchers when they search and browse through the repository. The work included among others the following activities:

1. Changing the TEI P5 schema for metadata by making many metadata elements optional and by giving more strict advice concerning the content of the metadata elements,
2. Allowing more diversity in the text format (also TEI P5) and therefore making tools more robust when handling the text by the available online tools,
3. A toolbox with conversion tools to convert text from a large number of formats into the CLARIN-DK TEI P5 format, and
4. An easy to use web-interface for adding the obligatory metadata and also, if the user wishes, a few important optional metadata for the converted text.

Activities 1 and 2 are not further described in this paper, but have been carried out.

3.1 Conversion of text to TEI P5 format

Often the textual data is not in one of the formats that CLARIN-DK accepts as depositing format. Some text resources will require human attention as the first step, but in other cases conversion to the accepted TEI P5 format can be completely mechanized. The CLARIN-DK text preparation service can handle uploaded text in a still growing variety of formats: plain text (txt), office document formats (DOC, docx, rtf, ODF), web documents (html), Adobe PDF, and even various OCR recognised image

[2] http://help.it.ox.ac.uk/oxgarage/index, https://github.com/TEIC/oxgarage
[3] http://pandoc.org

formats (gif, jpeg, pdf, png, tiff). The CLARIN-DK conversion service uses open source software (LibreOffice, html2text, pdfminer, Cuneiform, Tesseract) to convert the input to plain text or to RTF, from where tools developed by the CLARIN-DK team take over and complete the creation of TEI-output. If the need arises, the CLARIN-DK staff can implement specialized transformation tools for text data that require an idiosyncratic treatment and add those tools to the existing set of integrated tools. Once these tools are registered (URL, input and output specifications) they become part of the CLARIN-DK ecosystem and are automatically incorporated in work-flows when needed. (Jongejan 2013).

3.2 Applying metadata

It is also important to make it easier to apply metadata to the texts. From the experiences with research data providers for whom we have converted data and helped with applying metadata to the texts, we have concluded that initially it seems easy to fill out metadata, but that it is very difficult to create the TEI P5 metadata[4] in a standardized and consistent manner, especially for the technically inexperienced user (Hansen et. al. 2014, Offersgaard et. al. 2016).

After uploading a text resource for conversion, the user is confronted with a single page of metadata entry fields. Each field is accompanied by a help text in Danish or English. The web form does not require any XML and XSLT skills or knowledge of schema validation[5]. The web form asks for all obligatory metadata, which are fewer than before. In addition, there is a selection of optional metadata fields. Each field is validated and the user is asked to correct any errors that are discovered.

In this way the metadata creation is linked to the specific text resource type and the specific TEI P5 schema that CLARIN-DK is using. For use of the resource and the metadata outside CLARIN, the generated metadata can be extended compliant with TEI P5.

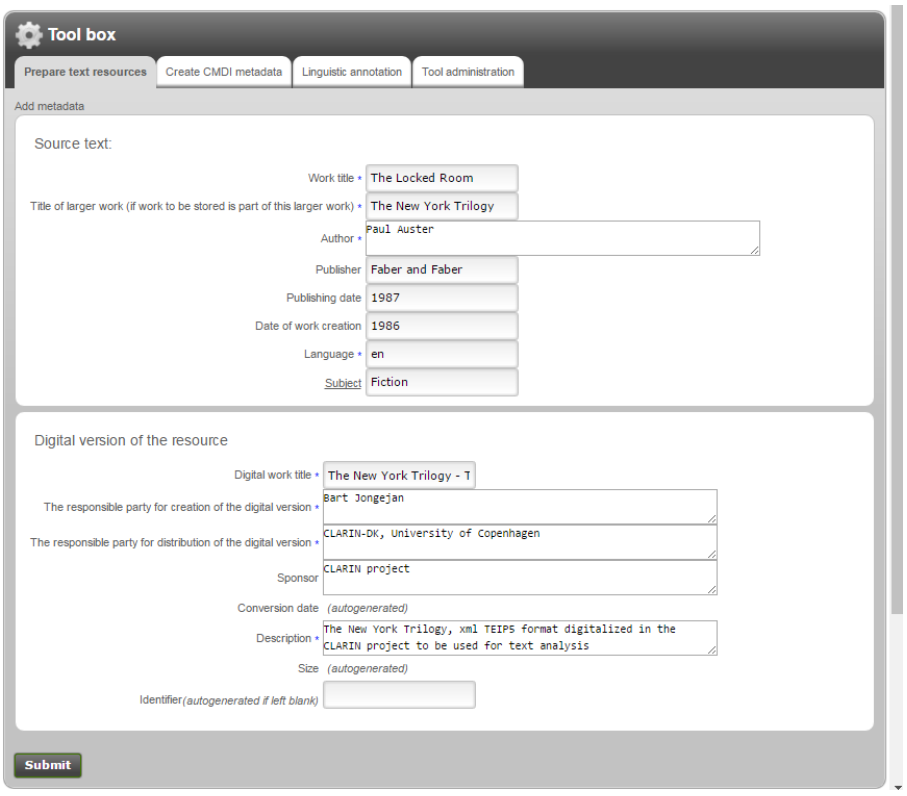

Figure 1: The metadata editor with metadata values for the novel "The Locked Room", part of "The New York Trilogy" by Paul Auster.

[4] The metadata used in CLARIN-DK conform to the TEI P5 scheme: https://clarin.dk/schemas/tei/TEIDKCLARIN.rng

[5] After downloading the TEI output experienced users can extend the metadata with more details using an XML editor. There are currently no plans to include this advanced functionality in the webservice.

3.3 The converted result

After completing the web form, the conversion of the uploaded text starts. The user downloads the results to her computer and inspects the end product, which is a TEI P5 file containing a metadata header and a body with text. Also the intermediary results for each conversion step are available to the user in easily accessible formats: plain text, RTF, or HTML. If needed, corrections can be made in the intermediary files, which then can be uploaded to the conversion web service. A readme file is also included, together with an index file with all metadata information the user provided in the web form, and details about the conversion process.

```
<TEI><teiHeader type="text"> <fileDesc><titleStmt>
        <title>The New York Trilogy – TEI P5 version</title>
        <respStmt>
          <resp>a_annotation</resp>
            <name><name>Bart Jongejan</name>
            <date when="2016-09-26"/></name>
        </respStmt>
        <sponsor>CLARIN project</sponsor>
      </titleStmt>
   <extent><num type="words">35083</num></extent>
     <publicationStmt>
       <distributor>CLARIN-DK, University of Copenha-
gen</distributor>
   </publicationStmt>
   <notesStmt>
     <note> The New York Trilogy, xml TEIP5 format digitalized
in the CLARIN project to be used for text analysis</note>
</notesStmt>
   <sourceDesc> <biblStruct>
        <analytic>
          <title xml:lang="en">The Locked Room</title>
          <author><name>Paul Auster</name></author>
        </analytic>
        <monogr>
          <title> The New York Trilogy </title>
          <imprint>
          <publisher n="n/a">Faber and Faber</publisher>
          <date when="1987"/>
         </imprint>
      </monogr> </biblStruct></sourceDesc>
etc.. …. </teiHeader>
```

```
<text>
    <body>
    <p n="1">
          <w xml:id="i1.1">And</w>
          <c xml:id="i1.2" type="s"/>
          <w xml:id="i1.3">death</w>
          <c xml:id="i1.4" type="s"/>
          <c xml:id="i1.5" type="p">….</c>
          <c xml:id="i1.6" type="s"/>
          <w xml:id="i1.7">happens</w>
          <c xml:id="i1.8" type="s"/>
          <w xml:id="i1.9">to</w>
          <c xml:id="i1.10" type="s"/>
          <w xml:id="i1.11">us</w>
          <c xml:id="i1.12" type="s"/>
          <w xml:id="i1.13">every</w>
          <c xml:id="i1.14" type="s"/>
          <w xml:id="i1.15">day</w>
          <c xml:id="i1.16" type="p">.</c>
    </p>
    etc.. ….
</text>
</TEI>
```

Figure 2: The text uploaded and the metadata entered now converted to TEI P5

If the result of the conversion is satisfying, the user can choose to upload the TEI P5 xml file to the CLARIN-DK repository. In this process CMDI metadata are created on the basis of the TEI P5 metadata and a teiHeader CMDI profile[6].

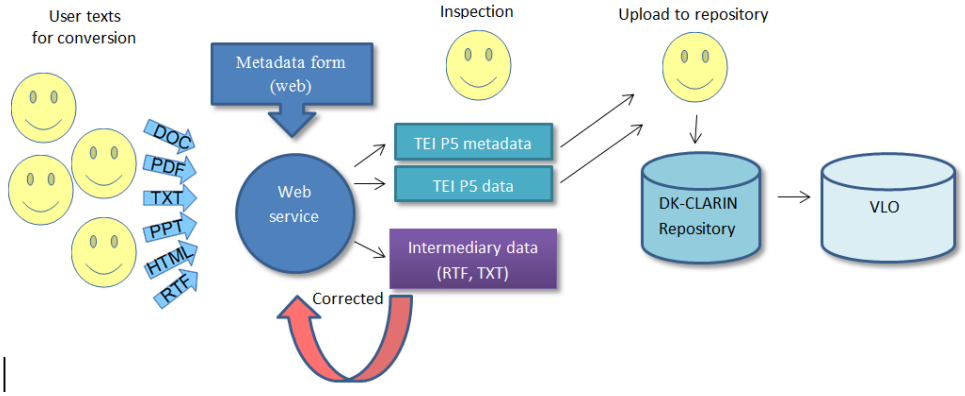

Figure 3: The architecture of the web service

[6] The CLARIN-DK CMDI teiHeader profile in the Component Registry (https://catalog.clarin.eu/ds/ComponentRegistry) ID: clarin.eu:cr1:p_1380106710826

It is important to notice that, although the whole process of text conversion and metadata creation shares functionality and user interface with the CLARIN-DK repository, it is not integrated in the repository itself. That means that also users who have no plans to deposit resources in CLARIN-DK can use the conversion utility to generate TEI P5 output. The shared functionality is embodied by the workflow management system that can operate on input that originates from the repository, but that alternatively can be uploaded by the user or fetched from a URL the user provides.

4 Knowledge sharing

In addition to the web service we find the availability of user support important. CLARIN-DK therefore contains information in form of tutorials[7], a helpdesk[8] and links to the schemas[9] used. We will extend the support facilities and prepare more tutorials for both new users and advanced users. The advanced tutorials, templates and schemas will allow experts to extend the metadata created in text preparation service with more details.

5 Conclusion

With a web-based text preparation service we reach out to researchers who hitherto have been unable to share their textual data due to limited resources, while ensuring that textual data and their metadata are made available in a uniform fashion. The service is in no way limited to use in connection to the CLARIN-DK repository, but can be used by all researchers.

Initially we only addressed the preparation of text resources, but after gathering some experiences we are currently expanding the facility to handle more kinds of resources.

The web service enables on the one hand the transformation of text to the TEI P5 format and on the other hand the creation of sufficient metadata for the resource – from the point of view of the repository as well as that of the researcher.

Reference

Hansen, D. H., Offersgaard, L., & Olsen, S. (2014). Using TEI, CMDI and ISOcat in CLARIN-DK. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation: LREC 2014.* (pp. 613 - 618). Reykjavik.

Jongejan, B. (2013). Workflow Management in CLARIN-DK. In *Proceedings of the workshop on Nordic language research infrastructure at NODALIDA 2013.* (Vol. 089, pp. 11-20). Linköpings Universitet: Linköping University Electronic Press. (NEALT (Northern European Association of Language Technology) Proceedings Series, Vol. 20).

Offersgaard, L., Hansen, D. H (2016). Facilitating Metadata Interoperability in CLARIN-DK. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation: LREC 2016.* (pp. 2510 - 2515). Portoroz, Slovenia.

Offersgaard, L., Jongejan, B., & Hansen, D. H. (2013). CLARIN-DK – status and challenges. In *Proceedings of the workshop on Nordic language research infrastructure at NODALIDA 2013.* (Vol. 089, pp. 21-32). Linköping University Electronic Press. (NEALT (Northern European Association of Language Technology) Proceedings Series; Vol. 20).

F. Zipser & L. Romary (2010). A model oriented approach to the mapping of annotation formats using standards. In Proceedings of the Workshop on Language Resource and Language Technology Standards, LREC 2010. Malta. URL: http://hal.archives-ouvertes.fr/inria-00527799/en/

Lyse, G. I, Meurer, P. & Smedt, K. COMEDI: A component metadata editor. In *CLARIN 2014 Selected Papers;* Linköping Electronic Conference Proceedings. (NEALT Series: NEALT Proceedings Series; 28)

[7] http://info.clarin.dk/clarin-dk-platformen/tutorials (not yet translated to English)
[8] http://info.clarin.dk/en/clarinhelpdesk
[9] https://clarin.dk/schemas/tei