Production and Presentation of Historical Text Data

The developments of information science produced perspectives in the area of the humanities only a very few scholars dared to think about twenty years ago. Storing and retrieving enormous quantities of text-related data has become much faster and easier than compiling a haystack and searching a needle in it. The national corpora present valuable text repositories containing tagged and untagged texts, the database systems are stable and fast enough to create useful research environments.

In this paper I will focus on the obstacles which hinder the use of these developments for a better international data interchange and and for an improvement of scholarly research.

1. The Peculiarities of Scholarly Production of Text Data

The preparation of texts for scholarly research requires in many cases more than typing and verifying the text itself. Often information about structure and presentation is added, and - concerning historical texts - different versions are used: a facsimile-like graphical representation with all abbreviatures, a diplomatic version with expanded abbreviatures, and a normalized version. Further levels of tagging are morphological features on word level and sequential and hierarchical data for syntactical and stylistic description.

Because of the geographical and historical diversity of languages many projects have produced their own tags and attributes to mark features important for their research. In earlier days, database tables were used to store this information, since the upcoming of XML many texts have been converted into this format (often using the column names as tags or attributes). Both types of data are sometimes accompanied by specialized search facilities.

A remaining problem in this area of text data production is that after the end of a project there is no money to preserve the data in an up-to-date status. Thus sometimes the data are not transformed to UniCode, sometimes the database system used for storage is outdated, sometimes SGML data needs a complicated conversion into XML, sometimes texts were produced with word processors marking features by color or font size or style. In many cases, it would be a valuable work to update these text data, and the expenses for converting them are often acceptable. Another remaining problem I will focus on later is the diversity of tagsets.

We are now in a situation that XML is a widely accepted standard language for text and data storage and exchange. The tools working with XML texts are now at a degree of sophistication; convenient editor programs help to type and to annotate texts, powerful transformation and retrieval processors guarantee a wide range of usability from web presentation and prepress reproduction to the preparation of search engine data. This situation opens an opportunity we should seize to prepare the infrastucture for the future.

2. The Possibilities of Presentation of Text Data

The world wide web had and has still an important impact on the development described above. On the one hand, it was the most important factor which pushed standardization and cooperation between competing companies. On the other hand, it opened a new world of participation and cooperation between research institutes which is still to be explored.

One of the upcoming opportunities I would like to focus on is the flexible combination of

text and image presentation based on XML and XSL, more precisely on the transformation of TEI-annotated texts to (X)HTML-conformant web interfaces.

The example I would like to present is taken from my work at the Arnamagnaen Institute at Copenhagen University and provides a novel called Fridthjofs saga written about 1550 by Ari Jónsson and either Jón or Tómas Arason. Matthew Driscoll prepared the electronic text in 2004 using TEI P4 and providing three versions: facsimile-like, diplomatic, and normalized. The original, named AM 510 4to, was shipped back to Iceland in 2008.

My task was to prepare the TEI P5 version and to create a new web interface providing all information available in the electronic text in an XHTML version. The processing of the TEI text is done automatically, i.e. the browser (we used Firefox and Opera) produces the presented version while loading the XML file. For longer texts, XHTML versions already prepared save some loading time.

Two main ideas determined the development. One of them was to combine image data and text data closely, realized by a function which allows to open exactly that part of an image which contains the line of the transcribed text the researcher is interested in. The other one was to present a search facility providing the (meta) information of the text data, in case of Fridthjofs Saga a word list based on the (facsimile, diplomatic, normalized) version.

Both ideas resulted in an XSL transformation with some parameters regarding the version presented and the display of images, lists, names, and editorial meta data. [These features will be demostrated during the presentation of the paper.]

3. The Necessity of an Improved Text Management

The remaining problem in the area of text-oriented scholarly research is the great variety of encoding and annotating schemes. On the one hand, the attempts to solve the encoding problem seems to be on a good way since UniCode has been established as a basis of encoding and additional character repositories like MUFI can be used for manuscript studies. On the other hand, the variety of annotation schemes is still a serious barrier for text exchange and use.

The reasons for this situation is obvious: literary scholars and linguists have different viewpoints regarding information to be added, the tagsets focus on content and style or on morphology and syntax. Whereas different levels of annotation could be combined with only few serious problems, the grammatical tagsets differ in such a range that most of them are nearly incompatible. As a result, to run a search on texts enriched by different tagsets will produce hitlists which are rarely comparable.

The attempts to standardize these tagsets - especially TEI and partly DocBook - aimed at incorporating the different interests. The result was a flexible tagset useful for many annotation goals but also with a flexible usage of tags and attributes which sometimes produces again incomparable information. For example, try to combine different bibliographies even if all of them are based on TEI.

The question is what can be done to solve or at least to defuse this problem causing incompatibilities and obstructing cooperation. Steps could be taken into two directions. Firstly, it seems to be essential to define an international standard and invite (or to urge) scholars to use this standard. Secondly, it seems necessary to create an interface which allows

a mutual unambiguous mapping. TEI was intended to be the basis of such an "international standard for text interchange" but its tag structure is too flexible to ensure unambiguity as already said.

The solution I would like to suggest could be a more restrictive subset of TEI which reduces the manifold tagging possiblities to one clear definition of tags and attributes. This tagset (perhaps called TEIcore) could be used for annotating texts but the main idea is that text-encoding projects could und should create transformations from and into this TEIcore tagset (using for example XSL).

As an example, I tried to create bilateral transformations between TEI P5 and the MENOTA tagset. The latter one has been created in order to fill the gaps in TEI P4 in the area of grammatical data and manuscript representaion. It is hard to proclaim one of them as the better solution; both provide the same information, MENOTA is friendlier to the annotator and the reader, TEI is less redundant and easier to process. Employing this XSL-transformation, you can produce text data using MENOTA, convert them to TEI P5, and vice versa. Updating and extending the MENOTA data is much easier because the conversion could be done on the fly, and furthermore, the converter will help to find inconsistent annotations.

These bilateral transformations are only the first step, because a star-shaped system is more effective, at least if more than three tagsets should interact. This is the most important impetus to establish a new effort to create such a TEIcore (or a new) tagset which could be used as a basis for this type of transformations. This effort should cover

(a) a comparison of the existing tagsets used in the main repositories of tagged historical (and contemporary?) text corpora,

(b) the definition of a core set of tags and attributes regarding text organization and presentation as well as stylistic and linguistic metadata on different levels,

(c) recommendations (and examples) for transformation programs.

There will be a lot of objections that the definition of such a core set is impossible because of the variety of linguistic theories regarding (for example) sentence structures. I agree insofar that the "x-bar theory" will result in other attribute values than the "head-driven phrase structure grammar" or the classical "part-of-sentence assignment". As a first answer, I would like to point out that the tags and the attributes itself could be the same, and it might be a very interesting task to compare the different values. To create a program for such a comparison will then last a few days and no longer several weeks or months. But the primary advantage would be the stability of the structure and the flexibility of the attribute values.

To conclude, the main problem to create transformation or presentation programs on the basis of TEI is its ambiguity as a result of its flexibility. Therefore a TEIcore (or another unambiguous) tagset will be a great advantage for producing and using texts and their meta data in literary and linguistic research. The possibilities of participating in the results of projects at other institutes will simplify and improve international cooperation. Furthermore it will produce more reliable results (because of a broader text basis) and it will - not at least - save a lot of money (because of a reduction of twice invented wheels).

Wolf-Dieter Syring (July 2009)