Published in: Calzolari,Nicoletta/Choukri, Khalid/ Maegaard, Bente/Mariani, Joseph/Odijk, Jan/Piperidis, Stelios/Tapias, Daniel (eds.): Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08. May 28-30, Marrakesch, Marokko). - European Language Resources Association (ELRA), 2008. pp. 371-378.

The Metadata-Database of a Next Generation Sustainability Web-Platform for Language Resources

Georg Rehm[1], Oliver Schonefeld[1], Andreas Witt[1], Timm Lehmberg[2], Christian Chiarcos[3], Hanan Bechara[1], Florian Eishold[1], Kilian Evang[1], Magdalena Leshtanska[1], Aleksandar Savkov[1], Matthias Stark[1]

University of Tübingen, Germany[1] | University of Hamburg, Germany[2] | University of Potsdam, Germany[3]
SFB 441: Linguistic Data Structures | SFB 538: Multilingualism | SFB 632: Information Structure
Nauklerstr. 35, Tübingen | Max-Brauer-Allee 60, Hamburg | Karl-Liebknecht-Str. 24-25, Potsdam

Corresponding author: georg.rehm@uni-tuebingen.de

Abstract

Our goal is to provide a web-based platform for the long-term preservation and distribution of a heterogeneous collection of linguistic resources. We discuss the corpus preprocessing and normalisation phase that results in sets of multi-rooted trees. At the same time we transform the original metadata records, just like the corpora annotated using different annotation approaches and exhibiting different levels of granularity, into the all-encompassing and highly flexible format eTEI for which we present editing and parsing tools. We also discuss the architecture of the sustainability platform. Its primary components are an XML database that contains corpus and metadata files and an SQL database that contains user accounts and access control lists. A staging area, whose structure, contents, and consistency can be checked using tools, is used to make sure that new resources about to be imported into the platform have the correct structure.

1. Introduction

This article describes a comprehensive database of metadata records that can be explored and searched in order to find language resources that are appropriate for one's specific research needs. It is one of the most crucial architectural components of a next generation sustainability platform for language resources that is currently under development in the project "Sustainability of Linguistic Data" (funded by the German Research Foundation, DFG).

Our project aims at sustainably archiving (Trilsbeek and Wittenburg, 2006) the language resources that have been developed or are still work in progress in three large-scale collaborative research centres. The groups in Tübingen (SFB 441: "Linguistic Data Structures"), Hamburg (SFB 538: "Multilingualism"), and Potsdam/Berlin (SFB 632: "Information Structure") built a total of 56 resources (mostly corpora and treebanks, but also lexicons, collections of sentences and associated grammaticality judgements etc.).[1] According to estimates it took more than one hundred person years to collect and to annotate these resources. The project has two primary goals:

1. To process and to sustainably archive the three SFBs' language resources so that they are still available to the research community and other interested parties in five, ten, or even 20 years time (Schmidt et al., 2006).
2. To enable researchers to query the resources both on the level of their metadata (for example, if a linguist who wants to work on a specific research question, tries to see whether there is an appropriate corpus he or she could use) as well as on the level of linguistic annotations (e. g., query one or more corpora for certain keywords, part-of-speech tags or syntactic patterns).

[1] We process 27 resources (16 corpora, five lexicons, and six sentence collections) from SFB 441, 18 corpora from SFB 538, and 11 corpora from SFB 632.

In more general terms, our main goal is to enable solutions that leverage the interoperability, reusability, and sustainability of a large collection of heterogeneous language resources. A web-based platform is our tool of choice to make sure that as many researchers as possible can access the language resources – even in the very long term.

2. Corpus Normalisation and Preprocessing

Language resources are almost exclusively built using XML-based markup languages nowadays (Wörner et al., 2006). Most current resources contain several annotation layers that correspond to multiple levels of linguistic description (for example, part-of-speech, syntax, coreference or other information related to semantics, etc.). As we have to process a heterogeneous set of corpora based on a number of different corpus markup languages, our approach includes the normalisation of XML-annotated resources, for example, for cases in which XML-annotated corpora use PCDATA content to capture both primary data (i. e., the original text or transcription) as well as annotation information (for example, part-of-speech tags). We use a set of tools to ensure that only primary data is encoded in PCDATA content and that all annotations proper are encoded using XML elements and attributes. Different annotation layers are separated into multiple files (see figure 1). For each layer, a tree consisting of XML elements and attributes is created. All trees share the same primary data. Thus the normalisation of each XML-annotated corpus results in a multi-rooted tree (Witt et al., 2007). Depending on the resource, the process of normalising the corpora and separating the annotation layers to a multi-rooted tree can be achieved using fully automatic or manual techniques. Custom tools are used to check that all files generated in the latter processing step are in fact identical with regard to their primary data.

Another reason for the normalisation procedure is that both hierarchical and timeline-based corpora need to be trans-

formed into a shared annotation approach, because we want our users to be able to query both types of resources at the same time and in a uniform way. In fact, the original annotation format will be irrelevant to the user, as the graphical interface and the underlying technology abstracts from any idiosyncrasies and peculiarities of the original data formats. Our approach can be compared to the NITE Object Model (Carletta et al., 2003): we developed tools that semiautomatically split hierarchically annotated corpora that typically consist of a single XML document instance into individual XML files, so that each file represents all the information related to a single annotation layer; this approach guarantees that overlapping structures can be represented straightforwardly. Timeline-based corpora are processed using other tools in order to separate graph annotations. This approach enables us to represent arbitrary types of XML-annotated corpora as individual files, i. e., individual XML element trees. These multi-rooted trees are represented as regular XML document instances. Details can be found in (Rehm et al., 2007a) and (Rehm et al., 2008a).

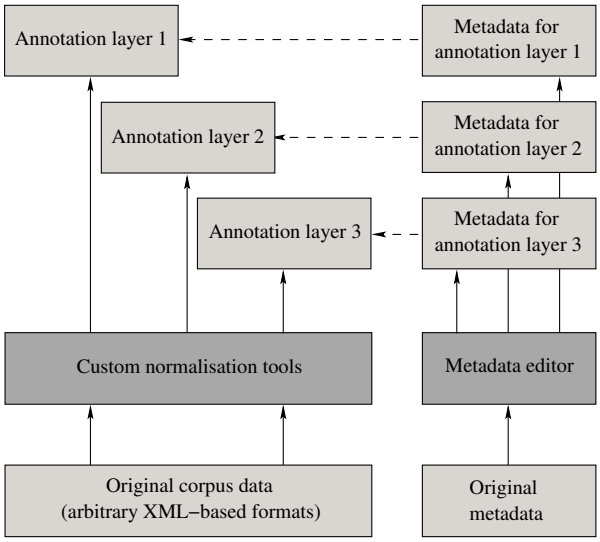

Figure 1: Metadata records for each annotation layer file

3. Legal Issues and the Need for Fine-Grained Access Control

Modern corpora contain multiple levels of annotation that refer to multiple levels of linguistic description. Before we are able to import corpora into our sustainability platform they are normalised and their individual annotation layers are separated (see section 2). This processing step has serious consequences with regard to legal issues that we have to take into account (Zimmermann and Lehmberg, 2007; Lehmberg et al., 2007a; Lehmberg et al., 2007b; Lehmberg et al., 2008; Rehm et al., 2007b): due to copyright and personal rights specifics that usually apply to a corpus's primary data (for example, copyrighted newspaper articles, transcribed doctor-patient-dialogues etc.) we provide a fine-grained access control layer to regulate access by means of user accounts and corpus-specific roles. In other words: it is not enough to specify that a certain user is allowed to access a certain corpus. We have to be able to explicitly represent the fact that a certain user only has access to the set of, say, six annotation layers (in this example they might be available free of charge for research purposes) but not to the primary data, because the primary data might be copyright-protected – for unrestricted access the user is required to provide proof-of-purpose of the primary data (such as, for example, a CD ROM). It is exactly this scenario that applies to the German treebank TüBa-D/Z (Rehm et al., 2007b). A closely related scenario refers to widely used corpora that are created and distributed by a specific research group and that are extended with additional or alternative annotation layers (that usually refer to new linguistic description layers) by other research groups. As the licence restrictions that apply to the original corpus and the annotation layers might be different, we need to be able to control access with regard to individual annotation layers. For this reason, every single annotation layer (as well as the corpus, the raw data, the primary data, and the setting) has an associated metadata record that, among others, contains information about potential access restrictions for the corresponding corpus files (see figure 1).

4. The Metadata Schema eTEI

Due to our primary goals, we use open, community- as well as industry-accepted and, therefore, sustainable standards wherever possible. We store the metadata records and the corpora themselves in a native XML database. We currently use eXist but we are still in the process of evaluating other databases for the back-end. The underlying assumption is that XML-annotated datasets are more sustainable than, for example, data stored in a proprietary relational database management system (the risk being that it might prove difficult or even impossible to run proprietary software on a modern operating system in, say, 15 years time).

Our generic metadata schema is based on the TEI P4 header (Sperberg-McQueen and Burnard, 2002) and extended by informational units that are missing in P4, but that are available in Dublin Core (`http://dublincore.org`), the ISLE Meta Data Initiative (IMDI, `http://www.mpi.nl/IMDI/`), the Open Language Archives Community (OLAC, `http://www.language-archives.org`), and that are among a set of requirements we collected.

Section 2 shows that multiple sets of annotation can refer to the same set of primary data. We do not follow the monolithic paradigm and treat the corresponding resource as *one* file (or as a set of small files) that has *one* metadata record. Rather, our position is a modular one so that we can apply metadata records to, for example, every single annotation layer, and to the set of primary data. This approach was born out of necessity: we process several corpora in which the set of primary data is available under a different licence than the set of annotations, so that we need to be able to distinguish between them (see below). Therefore, the main difference between eTEI and other approaches is that the generic eTEI metadata schema, currently formalised as a single document type definition (DTD), can be applied to five different levels of description (Trippel, 2004; Himmelmann, 2006). One set of metadata contains information on one of the following levels:

Metadata-RawData-File_1_of_1.xml:

```
<eTEIheader levelOfDescription="rawData" [...]>
  <fileDesc>
    <titleStmt>
      <title level="series">Asteriks</title>
      <author type="text"><person>Goscinny</person></author>
      <author type="drawing"><person>Uderzo</person></author>
      <editor><institution>Politikin [...]</institution></editor>
    </titleStmt>
    <publicationStmt>
      <publisher><institution>Politikin zabavnik -
        Politika AD</institution></publisher>
      <pubPlace><address>Beograd, Yugoslavia</address></pubPlace>
      <date>1995-1997</date>
    </publicationStmt>
    [...]
    <sourceDesc>
      <bibl>
        <biblPart>Comic series 'Asteriks' 1995-1997, ed. by "Politikin
          zabavnik - Politika AD", Beograd, Yugoslavia</biblPart>
      </bibl>
    </sourceDesc>
  </fileDesc>
  [...]
</eTEIheader>
```

Metadata-Corpus-File_1_of_1.xml:

```
<eTEIheader levelOfDescription="corpus" [...]>
  <fileDesc>
    <titleStmt>
      <title level="corpus">Comic Corpus</title>
      <respStmt>
        <resp type="project"><name>B8</name></resp>
      </respStmt>
      <respStmt>
        <resp type="annotation"><name><person>Gabi Fulir</person></name>
          <name><person>Slavica Stevanovic</person></name></resp>
      </respStmt>
    </titleStmt>
    <editionStmt><edition version="1.0"/></editionStmt>
    <extent><wordCount>57104</wordCount>[...]</extent>
    <publicationStmt>
      <distributor><institution>SFB 441</institution></distributor>
      <pubPlace><address>Nauklerstr. 35 [...]</address></pubPlace>
        [...]
    </publicationStmt>
  </fileDesc>
  <encodingDesc>
    <projectDesc>The Comic Corpus consists of [...]</projectDesc>
  </encodingDesc>
  <profileDesc>[...]</profileDesc>
</eTEIheader>
```

Metadata-Annotation-Deictics-File_1_of_1.xml:

```
<eTEIheader levelOfDescription="annotation" [...]>
  <fileDesc>
    <titleStmt>
      <title level="annotation">Deictics</title>
      <respStmt>
        <resp type="annotation">
          <name><person>Gabi Fulir</person></name>
          <name><person>Slavica Stevanovic</person></name></resp>
      </respStmt>
    </titleStmt>
    <editionStmt><edition version="1.0"/></editionStmt>
    <publicationStmt>
      <distributor><institution>SFB 441</institution></distributor>
      <pubPlace><address>Nauklerstr. 35 [...]</address></pubPlace>
    </publicationStmt>
  </fileDesc>
  <encodingDesc>
    <tagsDecl>[...]</tagsDecl>
  </encodingDesc>
  <profileDesc>
    [...]
  </profileDesc>
</eTEIheader>
```

Metadata-Annotation-Conversation-File_1_of_1.xml:

```
<eTEIheader levelOfDescription="annotation" [...]>
  <fileDesc>
    <titleStmt>
      <title level="annotation">Conversation</title>
      <respStmt>
        <resp type="annotation"><name><person>Gabi Fulir</person></name>
          <name><person>Slavica Stevanovic</person></name></resp>
      </respStmt>
    </titleStmt>
    <editionStmt><edition version="1.0"/></editionStmt>
    <publicationStmt>
      <distributor><institution>SFB 441</institution></distributor>
      <pubPlace><address>Nauklerstr. 35 [...]</address></pubPlace>
    </publicationStmt>
    <notesStmt><note [...]>Denotes speakers and utterances (contained
      in round boxes in the comic).</note></notesStmt>
  </fileDesc>
  <encodingDesc>
    <tagsDecl>
      <tagUsage gi="figure">Every panel of a comic [...]</tagUsage>
    </tagsDecl>
  </encodingDesc>
  <profileDesc>[...]</profileDesc>
</eTEIheader>
```

Figure 2: Four of the seven eTEI files (abridged) containing the metadata of the "Asterix" corpus

1. *setting* (applies to recordings or transcripts of spoken language primarily and describes the situation in which the speech or dialogue took place);
2. *raw data* (e. g., a book, a piece of paper, an audio or video recording of a conversation etc.);
3. *primary data* (transcribed speech, digital texts etc.);
4. *annotations* (that add information to primary data);
5. *a corpus* (consists of primary data with one or more annotation levels).

There are several additional reasons why we need to be able to represent metadata on these five different levels explicitly. Often used informational units in metadata records comprise terms such as "author", "creator", "date", "place" etc. Terms such as these are potentially ambiguous, e. g., does "author" refer to the author of the raw data, to the person who transcribed the raw data, to the author of the corpus, or maybe to the author of the annotation (e. g., a software tool, or a linguist who added a specific annotation to the corpus)? Moreover, corpora consist of at least two parts: a set of primary data and one or more layers of annotation. Different access restrictions can apply to these $2 + n$ layers of data: primary data is usually copyrighted (e. g., by a publishing house) and the different layers of annotation can have access restrictions of their own, (Rehm et al., 2007b; Rehm et al., 2007c) provide examples. As these metadata are of utmost importance for regulating web-based access to the corpora and their individual parts (see section 3), we need to be able to represent these subtle, but highly important properties that every single corpus has.

Figure 2 shows an eTEI example. The figure includes four heavily abridged files of the seven that contain the metadata describing the "Asterix" corpus, developed by the project B8 of SFB 441. One of the extensions we added to the TEI header is the obligatory attribute `levelOfDescription` that has five preset values (`setting`, `rawData`, `primaryData`, `annotation`, `corpus`). A project-internal technical document specifies several naming and structuring schemas that control the naming of corpus files, their metadata files, and the structure of the staging area (see section 5.2).

4.1. Editing, and Parsing eTEI Records

We developed an integrated workflow that helps users to edit, and parse eTEI records (see figure 3). The workflow's two primary components are the eTEI DTD and the highly flexible Oxygen XML editor. We use Oxygen's "project" facility to pre-configure the editor with several files. The eTEI DTD itself contains several structured annotations that are embedded into XML comments (`<!-- ... -->`) that apply to almost every single element and attribute declared in the DTD. The structured comments are anchored to their respective element or attribute using a unique naming scheme that repeats the element or element/attribute names and contain three informational units: (a) a short natural language description of the respective element or attribute, (b) if the element or attribute belongs to TEI P4 or if it

was taken from one of the standards mentioned in the introduction, and (c) annotations that specify if an element or attribute is valid with regard to the five levels of metadata description.
We process these structured XML comments using scripts (Perl, Python) in order to produce an empty XML document instance with embedded documentation and a Schematron schema. While the eTEI DTD can be used to validate the overall structure of an eTEI instance, the Schematron specification can be used to check whether all elements and attributes used in an eTEI instance conform to the current value of the `levelOfDescription` attribute; as it is impossible to specify corresponding rules in a DTD, we decided to implement this step using Schematron.
The empty eTEI document contains embedded documentation. We generate an empty eTEI document instance by converting the DTD into an XML Schema description using Oxygen and instantiating the Schema. Afterwards a Python script converts the comments that contain the short documentation remarks into `<_doc>` elements that precede the element they explain. These elements help users to assess the semantics of all elements and attributes. After editing an eTEI XML document instance, the user can activate an XSLT stylesheet that removes all `<_doc>` elements.

4.2. From Existing Metadata Records to eTEI

We process corpora and their metadata from three different research centres that have their own approaches for data handling. In the following three subsections we briefly discuss these heterogeneous approaches and the state of the corresponding metadata collections.

4.2.1. SFB 441 (Tübingen University)

The metadata records of SFB 441 are encoded using the Tusnelda annotation standard (Wagner, 2005) that consists of inline annotation, and nested hierarchies without overlaps. It is based on TEI and CES, but was adapted to meet the specific research purposes of the SFB projects. Since eTEI is also based on the TEI header, the transition from Tusnelda to eTEI is straightforward, and not too much further processing is necessary. What is problematic, however, are numerous idiosyncrasies found in the existing metadata records. Most idiosyncrasies can be traced back to the relative freedom given by the PCDATA content in most header elements. Plus, the metadata records were created by different researchers in different projects. Element content was entered inconsistently, yielding small, yet numerous differences that make the option of automatic processing rather hard. Idiosyncrasies include:

- *Character variations* – capitalisation is a common problem, differences exist even in the metadata from one and the same project, such as in, for example,
  - `<language>brazilian-portuguese</language>`
  - `<language>Brazilian Portuguese</language>`.

  Another type of variation occurs with special characters, such as the German umlauts. The name "Tübingen", for example, is written in two different ways, one with the original character ("ü"), and the other with its alternative international representation ("ue").
- *Delimiter usage* – a common problem is the use of different delimiters. In addition, whitespace is inconsistently used, especially with regard to the SFB itself, its projects and their notation. In some cases, the SFB and its number are noted (sometimes in different languages) along with the project number. In some cases, the SFB is omitted, in other cases, the principal investigator is given in brackets.
  - `<creator>SFB 441, B3</creator>`
  - `<creator>SFB441/project B8</creator>`
  - `<creator>SFB 441 / Projekt B1</creator>`
  - `<creator>B9</creator>`
  - `<creator>B9 (Schlieben-Lange) </creator>`.
- *Abbreviations* – often the full form of, e. g., a named entity is reduced to its more common abbreviation:
  - `<dist>Seminar für Sprachwissenschaft</dist>`
  - `<dist>SfS</dist>`.

  Similarly, there are some cases, in which the full name of a language is used in contrast to the standardised abbreviation:
  - `<language id="Russian"/>`
  - `<language id="ru"/>`
- *Dates and numbers* – there are several idiosyncrasies with regard to dates and numbers. Sometimes a date expression is given in a complete form, in other cases only the final two digits of the year are noted, in multiple cases the day and month are missing (sometimes they are swapped). In contrast, several date expressions contain dashes instead of periods.
  - `<pubDate>14.09.2001</pubDate>`
  - `<pubDate>06.09.01</pubDate>`
  - `<pubDate>1994</pubDate>`
  - `<pubDate>09-13-2001</pubDate>`

  Numbers are also displayed in different ways. Version numbers, for example, are written both with and without minor numbers (`version='1'` vs. `version='1.0'`).
- *Addresses* – they contain several information units and, therefore, multiple variations exist with regard to the notation and the level of detail specified in an address. In addition there are some cases of tag abuse when phone and fax numbers are included.
  - `<pubAddress>Nauklerstr. 35, 72074 Tübingen, Germany</pubAddress>`
  - `<pubAddress>Nauklerstr. 35, D-72074 Tuebingen</pubAddress>`
  - `<pubAddress>Nauklerstr. 35, 72074 Tübingen; tel: 07071-2977157, fax: 07071-295830</pubAddress>`
- *Named entities* – there are multiple variations concerning proper names. With regard to the names of persons, sometimes only the family name is specified (in contrast to both the first and the last name). Additional inconsistencies can be found in the names of sources (such as the names of newspapers).

As the data set is relatively small and as there are far too many idiosyncracies with regard to multiple informational units, we decided to transform the existing TEI headers into eTEI using a fully manual approach.

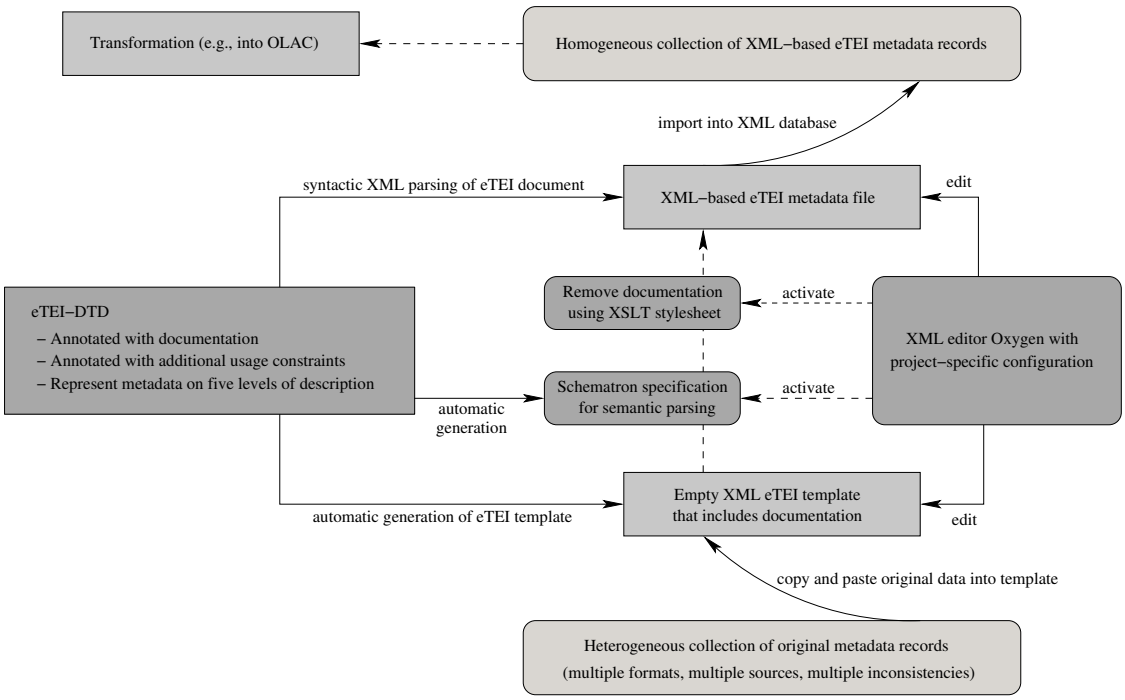

Figure 3: The integrated workflow for editing and parsing eTEI metadata records

4.2.2. SFB 538 (Hamburg University)

In SFB 538, large amounts of spoken as well as written corpora have been collected and processed. Whereas the majority of the written corpora have been encoded using the TEI-, or TEI-compliant formats such as MENOTA (`http://www.menota.org`), the situation for the spoken language corpora is more complex. In many cases they are associated with extensive metadata sets that contain information not only on speakers, and transcriptions, but also on situational contexts, and communication settings. For these purposes researchers normally define customized sets of metadata for their specific research questions.

The majority of the spoken language corpora have been processed using the Exmaralda system (Schmidt and Wörner, 2005). The Corpus-Manager (CoMa) was designed as an integral part of EXMARAaLDA to meet the special demands of metadata analysis mentioned above. CoMa enables researchers to bundle Exmaralda transcriptions into corpora and to structure them according to their individual metadata (contained in the header) into communications and speakers (Schmidt and Wörner, 2008). Thus it becomes possible to manage the complex relationship between speakers, transcriptions, and situational contexts. CoMa also allows the carrying out of metadata queries and application of filters to create subcorpora that only contain transcriptions with selected metadata attributes.

The individual and non-hierarchical metadata elements and structures cannot be integrated into a generic metadata scheme without an immense loss of information, of course. However, the transcriptions that constitute CoMa-based corpora still contain metadata that refer to speakers and annotations as well as settings and raw data. After running through the preprocessing steps described in section 2, these records can be easily transferred into eTEI.

4.2.3. SFB 632 (Potsdam University)

In SFB 632, metadata are collected according to different community-specific standards. Three types of metadata records are to be distinguished, for different types of collections, i. e., corpora of well-documented modern languages, historical documents, and typological data collections.

As for collections of modern language, these include corpora of written language, in particular, newspaper articles, but also spoken language (e. g., radio news and parliamentary debates). The metadata of these collections can be compared to those at SFB 441 and SFB 538 with few problems for the transformation into eTEI.

With regard to the historical corpora, the metadata of their primary data is specific, insofar as there is only a limited set of documents available for these languages (e. g., Old High German, Old Saxon), and most of these documents have a long editorial history. Therefore, metadata mostly concerns editorial information. These metadata, however, are often implicitly represented as many documents can be generally identified using their names alone. For example, *Heliand* and *Muspilli* are medieval manuscripts whose denotations act like proper names. Thus, explicit metadata for primary data of corpus languages is generally sparse, because this kind of knowledge is taken for granted within the respective community. In the transformation to eTEI, this information is preserved, but not extended.

The extreme opposite are the typological data collections. In SFB 632, an extended version of IMDI was established as the metadata standard for typological projects. The original IMDI elements can also be integrated with eTEI. Technically, metadata are an integral part of PAULA, the generic data format used in SFB 632 (Dipper, 2005). PAULA is a generic standoff-format comparable to the Linguistic Annotation Framework (Ide et al., 2005). Conceptually similar

to LAF, PAULA operates on the basis of a graph-theoretic data model, and, therefore, different types of annotations can be transformed into PAULA losslessly – this also includes different representations of metadata.
A PAULA project consists of a set of XML files with XPointers connecting files directly or indirectly with the primary data. There are basically four types of files: *text files* contain the primary data, *markables* define text spans that can be annotated, *structure files* define elements that are linked in a hierarchical structure, and *feature files* specify annotations assigned to markables and structure elements. The files are organised by means of an AnnoSet, i. e., a specialised structure file that specifies the hierarchical organisation imposed on the documents in the current PAULA project. This hierarchical organisation represents the grouping of markables or structure files together with the feature files pointing to these. This grouping corresponds to one single annotation layer. However, as the AnnoSet is a structure file itself, also its elements can be subject to feature specifications. These features may refer to a single XML document (structure, markables, feature, or text files), one annotation layer, or the project itself. Features that refer to the elements of an AnnoSet are *defined* as metadata, while features that point to other XML files are annotations. Thus, in PAULA metadata is not *structurally* distinguished from annotations, but *functionally*. Therefore, metadata can be processed and queried in the same way as other annotations. The following levels of metadata can be distinguished:

- Metadata of primary data (corresponding to *primary data* metadata in eTEI), i. e., feature specifications that refer to the text file.
- Metadata of segmentations, annotations and layers (corresponding to *annotation* metadata in eTEI), i. e., feature specifications that refer to groups of markable, structure, or feature files.
- Metadata of documents (corresponding to *setting* metadata in eTEI), i. e., feature specifications that refer to the AnnoSet file for a single document.
- Metadata of subcorpora and corpora (corresponding to *corpus* metadata in eTEI), i. e., feature specifications that refer to the AnnoSet file for several documents.

This classification corresponds to the five levels of metadata description presented in section 4 with the only exception that raw data and primary data are currently not distinguished in PAULA. Later versions will incorporate audio, and video files. Features that refer to these media files in an AnnoSet correpond to eTEI *raw data* metadata.
As both content and format of PAULA metadata and eTEI resemble each other, converters between both formats can be easily implemented. However, eTEI relies on consistent naming conventions for features that express metadata which is not guaranteed by the PAULA format and, thus, PAULA-to-eTEI conversion requires manual preprocessing. Moreover, it should be noted that normalisation issues as pointed out in section 4.2.1 have been assessed in the PAULA metadata only to a limited degree. For every exported metadata entry, manual correction cycles with the eTEI editor need to be performed.

5. An XML-Database of Metadata Records

The web-based sustainability platform has two main functional areas: (a) browsing of, and search within the database of corpus metadata, and (b) browsing of, and search within one or more corpora. Both functional areas rely on a database of metadata about the corpora contained in the platform. Following, we briefly discuss the system architecture (section 5.1), the staging area (section 5.2), and the basic functionality of the front-end (section 5.3).

5.1. The Architecture

The sustainability platform consists of two main components: the front-end and the back-end. The front-end is the user visible part and is realised using JSP (Java Server Pages) and Ajax technology. It runs in the user's browser and provides functions to search and to explore the metadata records. Based on the metadata, the user can choose one or more resources for further processing, such as querying or downloading. Query results can be displayed in formats such as KWIC or a tree view (Rehm et al., 2008a).
The back-end component of the platform hosts the Java Server Pages and related files. It accesses two different databases, the *corpus database* and the *system database*. The corpus database is an XML database in which all resources and metadata are stored, allowing users to query the data using XQuery. The system database is a relational database that contains all data about user accounts, resources (i. e., annotation layers), resource groups (i. e., corpora) and access rights to these resources. A specific user can only access a specific resource if the permissions for this user/resource tuple allow access. The system data is kept separate from the corpus data to allow for a cleaner separation of these repositories and for enhanced performance as well as security.

5.2. The Staging Area

A new resource is imported into the sustainability platform by copying all corresponding files into the staging area. The directory structure of the staging area is defined on six levels: the name of the *organisation*, the *organisational unit*, the name of the *project*, the name of the *corpus*, its *version* and the actual *corpus data*, i. e., a set of files processed for the platform (see section 2), the original corpus data and a set of metadata files. Strict naming rules apply for the processed corpus files, for the metadata files, and for the directories, but it is not necessary to alter the names of the original file and directories as they are stored in separate directories below the processed corpus files. Furthermore, each corpus contains a manifest file. Manifest files, represented in a simple XML format (see figure 5) act as corpus inventories and supply additional data about the files included in a corpus. They are automatically generated by the corpus normalisation tools described in section 2 and their contents are used by the import and export tools, and by the GUI.
The importer tool traverses the staging area, checks the data for consistency and imports the corpus data and metadata records into the XML corpus database. At the same time, new resource and resource group records as well as permissions are set up in the system database. The default permis-

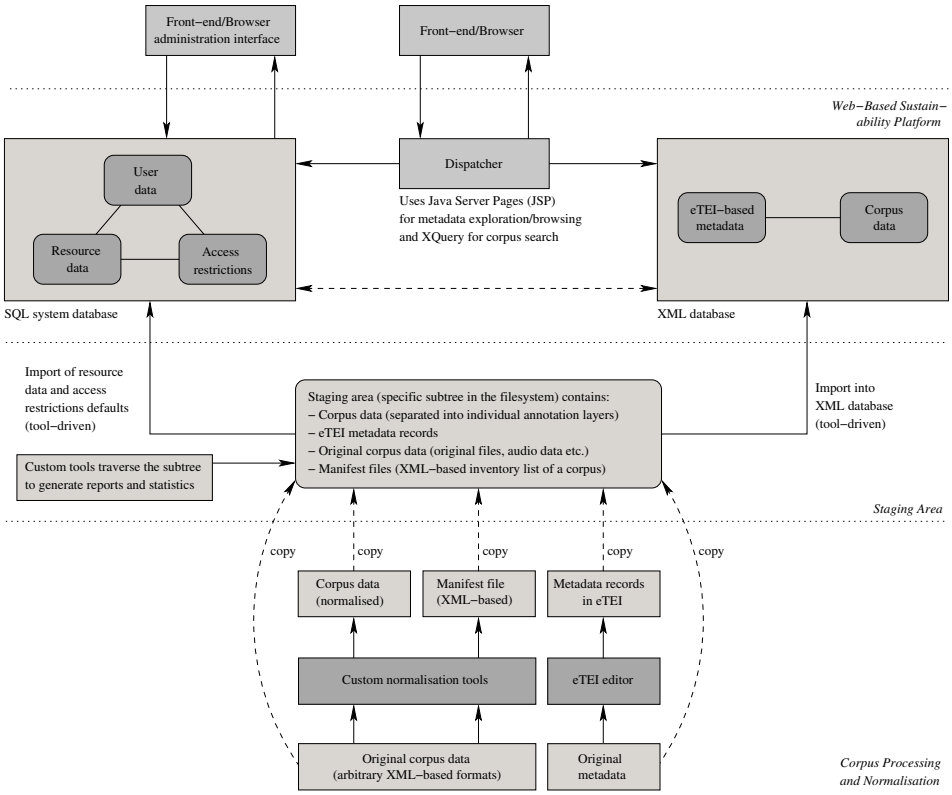

Figure 4: Normalisation of original corpus data and metadata and the staging area of the web-based sustainability platform

```
<!ELEMENT manifest (preamble,
                    (corpus|(subcorpus,subcorpus+)))>
<!ATTLIST manifest version CDATA #REQUIRED>

<!ELEMENT preamble (title,description+)>

<!ELEMENT corpus (processed,original,transformation)>

<!ELEMENT processed (dataset+)>
<!ELEMENT original (file+)>
<!ELEMENT transformation (file+)>

<!ELEMENT dataset (file+)>
<!ATTLIST dataset source IDREF #REQUIRED>

<!ELEMENT file EMPTY>
<!ATTLIST file id ID #REQUIRED
    filetype CDATA #IMPLIED
    contents (corpusdata|metadata|both|other) #REQUIRED
    linkanchor CDATA #IMPLIED
    location CDATA #REQUIRED
    checksum CDATA #IMPLIED>

<!ELEMENT subcorpus EMPTY>
<!ATTLIST subcorpus location CDATA #REQUIRED>

<!ELEMENT title (#PCDATA)>
<!ELEMENT description (#PCDATA)>
<!ATTLIST description lang CDATA #REQUIRED>
```

Figure 5: The DTD for manifest files

sions are chosen based on the restrictions defined in metadata records. Since the resources come from three different research centres it is vital to have a common and consistent naming scheme for directories and individual files.

5.3. Accessing and Exploring Metadata and Corpora

To work with the platform, a user first needs to log in (the credentials are validated against the system database). Then, the user can choose one or more corpora to work with. This is done based on the metadata records that can be searched and explored in several different ways. The permissions stored in the system database govern access to the data. After selecting a set of corpora to work with, the user can query or download them.

An administration interface enables administrator users to create, modify, or delete user accounts or specific access rights on resources. Figure 4 gives an overview of the workflow and the individual components. The lower part of the figure shows the workflow for creating a processed corpus with associated metadata which are copied to the staging area (middle). The importer tool imports the respective data sets into the databases in the back-end (top). The front-end accesses the data using a dispatcher that checks the access control lists and that submits requests to the databases.

6. Summary and Conclusions

Our goal is to provide a web-based platform for the long-term preservation and distribution of a set of linguistic resources. We briefly discussed the preprocessing phase in which we normalise the heterogeneous corpora into a common annotation format and into a set of multi-rooted trees – at the same time we transform the metadata records associated with the corpora into an all-encompassing format. This highly flexible eTEI format is able to represent practically all informational units contained in the original metadata records in a uniform and homogeneous way. In order to edit and to process eTEI metadata records we developed an integrated workflow that is based on an XML DTD and several tools. This workflow primarily aims at supporting the user for the conversion of existing metadata records

into the eTEI format. We also presented the architecture of the web-platform. Its primary components are an XML database that contains the corpus and metadata files and a relational database that contains all user accounts, associated security data, and access control lists. A staging area, whose structure and contents can be checked using tools, is used to make sure that new resources that are about to be imported into the platform have the correct structure.

7. Future Work

The research presented in this article is still work in progress. We want to highlight some of the aspects that we plan to realise by the end of 2008. While the corpus normalisation and preprocessing phase is, with only minor exceptions, finished, the process of transforming the existing metadata records into the eTEI format will be completed by the end of May. Work on the metadata exploration and on the graphical visualisation and querying front-end (Rehm et al., 2008a) as well as on the back-end is also ongoing; we plan to finish work on the platform by September.

We plan several extensions and modifications for the eTEI schema. Most notably, we plan to replace the current DTD, based on TEI P4, with an XML Schema description that is based on the current version of the TEI guidelines (P5). XML Schema has better and more appropriate facilities for including embedded documentation than the simple and unstructured comments available in DTDs. We use a web-based graphical customisation environment for TEI P5 tagsets, ROMA (`http://www.tei-c.org/Roma/`), to create and to edit an ODD, "one document does it all" specification (ODD documents are TEI instances that use the "tagdocs" module; the ODD format was completely revised for TEI P5) to maintain our modifications and extensions using a standardised and sustainable approach. The ODD file also contains the documentation and all related data.

An extension of eTEI that is very important for enhanced sustainability concerns several XML-based repositories in which data that is referenced in the eTEI metadata records multiple times will be stored in a centralised way. This approach will primarily make sure that entities (names of researchers, associates, annotators, projects, institutions, languages, countries etc.) and standardised sets of, for example, language codes, are stored only once in order to reduce redundancy and to enhance data consistency. We plan to use XLink/XPointer to reference these pieces of XML-represented information flexibly. To give another example, we are currently in the process of constructing a taxonomy of text types with the goal of annotating every text in each corpus with its genre, or text type, see also (Rehm et al., 2008b). Such references can easily be realised by pointing to the corresponding XML elements that, in this case, encapsulate texts (e. g., `<article>`, `<text>` etc.). This simple yet powerful mechanism allows us to add full sets of metadata to arbitrary XML elements. Especially to provide remote project teams with an editing and browsing environment for the central databases and eTEI records, we plan to implement a web-based eTEI editor.

As soon as all corpora and metadata records are finished, we plan to submit our metadata records to the aggregators `http://www.driver-repository.eu` and `http://www.language-archives.org` to make sure that interested parties are able to find them.

Acknowledgments The research presented in this paper was supported by a grant from *Deutsche Forschungsgemeinschaft* within the project *Nachhaltigkeit linguistischer Daten*.

8. References

J. Carletta, J. Kilgour, T. J. O'Donnell, S. Evert, and H. Voormann. 2003. The NITE Object Model Library for Handling Structured Linguistic Annotation on Multimodal Data Sets. In *Proc. of the EACL Workshop on Language Technology and the Semantic Web (3rd Workshop on NLP and XML)*.

S. Dipper. 2005. XML-Based Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation. In *Proc. of Berliner XML Tage 2005*, pages 39–50, Berlin, Germany.

N. P. Himmelmann. 2006. Daten und Datenhuberei. Keynote speech, 28th annual meeting of the DGfS, Bielefeld, February.

N. Ide, L. Romary, and E. de la Clergeri. 2005. International Standard for a Linguistic Annotation Framework. In *Proc. of the HLT-NAACL'03 Workshop Software Engineering and Architecture of Language Technology*.

T. Lehmberg, C. Chiarcos, E. Hinrichs, G. Rehm, and A. Witt. 2007a. Collecting Legally Relevant Metadata by Means of a Decision-Tree-Based Questionnaire System. In *Digital Humanities 2007*, pages 164–166, Urbana-Champaign, IL, USA, June.

T. Lehmberg, C. Chiarcos, G. Rehm, and A. Witt. 2007b. Rechtsfragen bei der Nutzung und Weitergabe linguistischer Daten. In G. Rehm, A. Witt, and L. Lemnitzer, editors, *Data Structures for Linguistic Resources and Applications: Proc. of the Biennial GLDV Conf. 2007*, pages 93–102. Gunter Narr, Tübingen.

T. Lehmberg, G. Rehm, A. Witt, and F. Zimmermann. 2008. Preserving Linguistic Resources: Licensing – Privacy Issues – Mashups. *Library Trends.* In print.

G. Rehm, R. Eckart, and C. Chiarcos. 2007a. An OWL- and XQuery-Based Mechanism for the Retrieval of Linguistic Patterns from XML-Corpora. In G. Angelova, K. Bontcheva, R. Mitkov, N. Nicolov, and N. Nikolov, editors, *Int. Conf. Recent Advances in NLP (RANLP 2007)*, pages 510–514, Borovets, Bulgaria, September.

G. Rehm, A. Witt, H. Zinsmeister, and J. Dellert. 2007b. Corpus Masking: Legally Bypassing Licensing Restrictions for the Free Distribution of Text Collections. In *Digital Humanities 2007*, pages 166–170, Urbana-Champaign, IL, USA, June.

G. Rehm, A. Witt, H. Zinsmeister, and J. Dellert. 2007c. Masking Treebanks for the Free Distribution of Linguistic Resources and Other Applications. In *Proc. of the Sixth Int. Workshop on Treebanks and Linguistic Theories (TLT 2007)*, pages 127–138, Bergen, Norway, December.

G. Rehm, R. Eckart, C. Chiarcos, and J. Dellert. 2008a. Ontology-Based XQuery'ing of XML-Encoded Language Resources on Multiple Annotation Layers. In *Proc. of the 6th Language Resources and Evaluation Conf. (LREC 2008)*, Marrakech, Morocco, May.

G. Rehm, M. Santini, A. Mehler, P. Braslavski, R. Gleim, A. Stubbe, S. Symonenko, M. Tavosanis, and V. Vidulin. 2008b. Towards a Reference Corpus of Web Genres for the Evaluation of Genre Identification Systems. In *Proc. of the 6th Language Resources and Evaluation Conf. (LREC 2008)*, Marrakech, Morocco, May.

T. Schmidt and K. Wörner. 2005. Erstellen und Analysieren von Gesprächskorpora mit EXMARaLDA. *Gesprächsforschung*, 6:171–195.

T. Schmidt and K. Wörner. 2008. EXMARaLDA – Creating, Analysing and Sharing Spoken Language Corpora for Pragmatic Research. In J. Allwood, editor, *Corpus Based Pragmatics – Proc. of the 10th Int. Pragmatics Conf. (Göteborg, 8–13 July 2007)*. To appear.

T. Schmidt, C. Chiarcos, T. Lehmberg, G. Rehm, A. Witt, and E. Hinrichs. 2006. Avoiding Data Graveyards: From Heterogeneous Data Collected in Multiple Research Projects to Sustainable Linguistic Resources. In *Proc. of the E-MELD 2006 Workshop on Digital Language Documentation: Tools and Standards – The State of the Art*, East Lansing, Michigan, June.

C.M. Sperberg-McQueen and L. Burnard, editors. 2002. *TEI P4: Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Initiative Consortium. XML Version: Oxford, Providence, Charlottesville, Bergen.

P. Trilsbeek and P. Wittenburg. 2006. Archiving Challenges. In J. Gippert, N. P. Himmelmann, and U. Mosel, editors, *Essentials of Language Documentation*, pages 311–335. Mouton de Gruyter, Berlin, New York.

T. Trippel. 2004. Metadata for Time Aligned Corpora. In *Proc. of the LREC Workshop: A Registry of Linguistic Data Categories within an Integrated Language Repository Area*, Lisbon.

A. Wagner. 2005. Unity in diversity: Integrating differing linguistic data in TUSNELDA. In S. Dipper, M. Götze, and M. Stede, editors, *Heterogeneity in Focus: Creating and Using Linguistic Databases*, volume 2 of *Working Papers of the SFB 632*, pages 1–20. Potsdam.

A. Witt, O. Schonefeld, G. Rehm, J. Khoo, and K. Evang. 2007. On the Lossless Transformation of Single-File, Multi-Layer Annotations into Multi-Rooted Trees. In B. T. Usdin, editor, *Proc. of Extreme Markup Languages 2007*, Montréal, Canada.

K. Wörner, A. Witt, G. Rehm, and S. Dipper. 2006. Modelling Linguistic Data Structures. In B. T. Usdin, editor, *Proc. of Extreme Markup Languages 2006*, Montréal, Canada.

F. Zimmermann and T. Lehmberg. 2007. Language Corpora – Copyright – Data Protection: The Legal Point of View. In *Digital Humanities 2007*, pages 162–164, Urbana-Champaign, IL, USA, June.