Sustainability of annotated resources in linguistics: A web-platform for exploring, querying, and distributing linguistic corpora and other resources

Georg Rehm
vionto GmbH, Berlin, Germany

Oliver Schonefeld
German National Library of Medicine (ZB MED), Cologne, Germany

Andreas Witt
Institute for the German Language (IDS), Mannheim, Germany

Erhard Hinrichs
General and Computational Linguistics, Tübingen University, Tübingen, Germany

Marga Reis
Deutsches Seminar, Tübingen University, Tübingen, Germany

Correspondence:
Georg Rehm,
vionto GmbH, Karl-Marx-Allee 90a, D-10243 Berlin, Germany
E-mail:
georg.rehm@vionto.com

Abstract

We report on finished work in a project that is concerned with providing methods, tools, best practice guidelines, and solutions for sustainable linguistic resources. The article discusses several general aspects of sustainability and introduces an approach to normalizing corpus data and metadata records. Moreover, the architecture of the sustainability platform implemented by the authors is described.

1 Introduction

In practically all scientific fields the task of ensuring the sustainability of resources, data collections, personal research journals, and databases is an increasingly important topic—linguistics is no exception (Dipper *et al.*, 2006; Trilsbeek and Wittenburg, 2006). We report on finished work in a project that is concerned with providing methods, tools, best-practice guidelines, and solutions for *sustainable* linguistic resources.[1] Our overall goal is to make sure that a large and heterogeneous set of

approximately sixty linguistic resources will be accessible, readable, and processible by interested parties such as, for example, other researchers than the ones who originally created said resources, in 5, 10, or even 20 years time—our method-of-choice for making sure that these corpora are available in the coming years is a web-based sustainability platform. The language resources we work with have been constructed or are still work in progress in three collaborative research centres (Sonderforschungsbereiche). 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, corpora, and treebanks mostly. According to our estimates it took more than one hundred person years to collect and to annotate these datasets. To summarize 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.

The agency that funded both our project as well as the projects who created the linguistic resources—the German Research Foundation (DFG)—would like to avoid a situation in which they have to fund yet another project to (re)create a dataset for whose creation they already provided funding in the past, but the existing version is no longer available, accessible or readable due to, for example, a proprietary file format, because it has been locked away in an academic's hidden vault, or the person who developed the annotation scheme can no longer be asked questions concerning specific details of the custom-built format.

The remainder of this article is structured as follows: first, Section 2 discusses several general aspects of sustainability, while Section 3 introduces our approach to normalizing corpus data and metadata records. The architecture of the sustainability platform we implemented is described in Section 4, although we are only able to highlight selected parts of the system due to space restrictions: the staging area is briefly discussed in Section 4.1, Section 4.2 describes the back-end of the system, the web-based graphical interface that includes a corpus query and visualization front-end is explained in Section 4.3. The article ends with concluding remarks (Section 5).

2 Linguistic Resources: Aspects of Sustainability

There are several text and data types that linguists and other scholars who regularly create or use language data work and interact with on a frequent basis, but the most common, by far, are linguistic corpora (Garside *et al.*, 1997; Zinsmeister *et al.*, 2008). In addition to rather simple word lists and sentence collections, empirical sets of grammaticality judgements, and lexical databases, the linguistic resources our sustainability project is primarily confronted with are linguistic corpora that contain either texts or transcribed speech in dozens of languages. These resources are annotated using multiple, mostly incompatible annotation schemes. We developed XML-based tools to normalize the existing resources into a common approach of representing linguistic data (Wörner *et al.*, 2006; Witt *et al.*, 2007) and represent knowledge about the individual annotation schemes used in the original resources with the help of OWL ontologies (Rehm *et al.*, 2007a).

The goal of providing durable, sustainable language corpora is faced with a multitude of challenges (see the extensive discussion by Bird and Simons, 2003). In the following, we briefly touch upon the most important ones:

- Markup languages: language resources which have been created using proprietary software tools or annotated based on custom-made, non-standard annotation schemes need to be transformed into a data format that can be considered sustainable and standardized (Wörner *et al.*, 2006; Witt *et al.*, 2007, 2009; Lehmberg and Wörner, 2009).
- Metadata encoding: heterogeneous metadata records across a large set of heterogeneous language resources that were created by several research groups at multiple sites need to be normalized and brought in line with one another so that these metadata records can be used in a

sustainability platform (Burnard and Bauman, 2007; Rehm *et al.*, 2008d).

- Legal aspects: while some of the corpora that we process belong to the public domain and can thus be made available without any need for access regulations, there are other corpora whose primary data consists of copyrighted material or highly sensitive data such as doctor–patient conversations; therefore, we need to make sure that only those users who are authorized to view or to download a specific resource are able to do so (Zimmermann and Lehmberg, 2007; Lehmberg *et al.*, 2007a,b; Rehm *et al.*, 2007b,c).

- Corpus querying and metadata search: in order to enable interested researchers to find and to evaluate the resources that we prepared, homogeneous metadata records as well as a common data format ensures that metadata and corpus data can be queried (Rehm *et al.*, 2007a, 2008a,d; Soehn *et al.*, 2008).

As already hinted at in the introduction, our method of choice for addressing the above-mentioned aspects and providing access to the corpora we processed in the course of our project is a web-based sustainability platform. First and foremost, this platform, called SPLICR (Sustainability Platform for Linguistic Corpora and Resources), is aimed at researchers who work in linguistics, computational linguistics, and related fields (Rehm et al., 2008b,c). SPLICR provides a comprehensive database of metadata records that can be explored and searched online in order to locate among the set of processed language resources those that could be appropriate for one's specific research needs. In addition, the system provides a graphical interface that enables users to query and to visualize corpora. Resources or specific parts thereof can also be downloaded.

The main advantage of SPLICR and its underlying architecture is that we designed and specified an integrated workflow that starts with the processing of individual corpora at multiple sites using custom-made tools. Afterwards, the processed corpora along with metadata files, their original data sets, HTML- or PDF-based manuals, and transformation logfiles are copied onto a server in a directory tree whose structure is specified by rigid protocols. In the next step, this directory tree is traversed using a lightweight importer client that checks the directory tree for consistency and copies the corpus files onto the SPLICR server.

3 Data Normalization and Representation

One of the obstacles we are confronted with is providing homogeneous means of accessing a large collection of diverse and complex linguistic resources. For this purpose we developed several custom tools in order to normalize the corpora (Section 3.1) and their metadata records (Section 3.2).

3.1 Normalization of linguistic resources

Language resources are usually built using XML-based markup languages nowadays (Ide *et al.*, 2000; Sperberg-McQueen and Burnard, 2002; Wörner *et al.*, 2006; Lehmberg and Wörner, 2009) and contain several concurrent annotation layers that correspond to multiple levels of linguistic description (e.g. part-of-speech, syntax, coreference). Our approach includes the normalization of XML-annotated resources, e.g. for cases in which corpora use PCDATA content to capture both primary data (i.e. the original text or transcription) as well as annotation information (e.g. POS 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. The transformation from PCDATA content (i.e. XML elements) to CDATA values (i.e. XML attributes) is performed semi-automatically (see Wörner *et al.*, 2006, for details).

Figure 1 illustrates this process by means of an excerpt from the TüBa-D/Z treebank (Telljohann *et al.*, 2004) in one of its four representation formats (Tusnelda, see Wagner, 2005). Beside the actual primary data content 'die AWO' (the PCDATA content of the XML element <orth>) other XML elements such as <pos> use PCDATA content to encode grammatical information. Since this information serves annotation purposes, the contents of
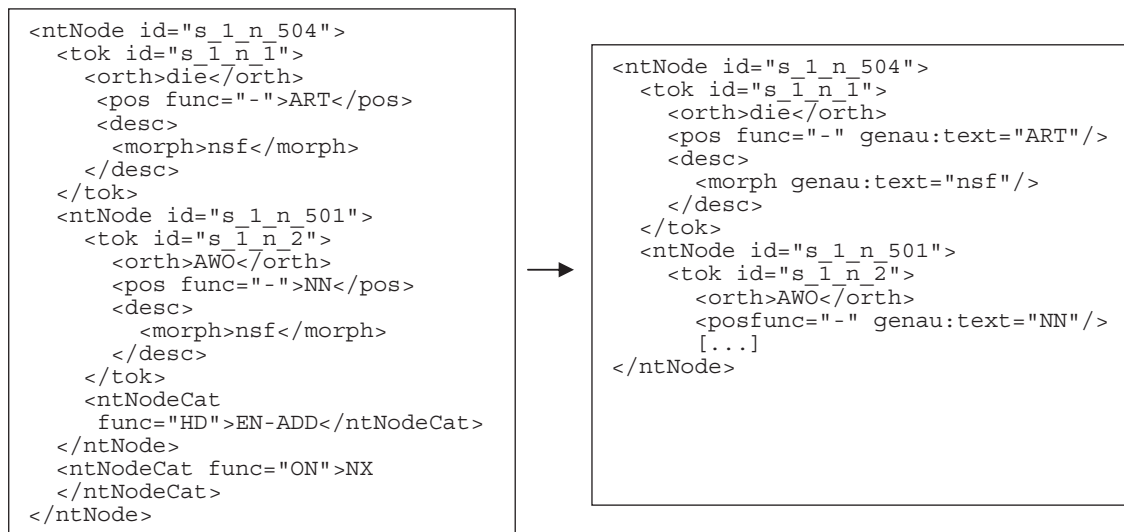
```
<ntNode id="s_1_n_504">
  <tok id="s_1_n_1">
    <orth>die</orth>
    <pos func="-">ART</pos>
    <desc>
      <morph>nsf</morph>
    </desc>
  </tok>
  <ntNode id="s_1_n_501">
    <tok id="s_1_n_2">
      <orth>AWO</orth>
      <pos func="-">NN</pos>
      <desc>
        <morph>nsf</morph>
      </desc>
    </tok>
    <ntNodeCat
    func="HD">EN-ADD</ntNodeCat>
  </ntNode>
  <ntNodeCat func="ON">NX
  </ntNodeCat>
</ntNode>
```

```
<ntNode id="s_1_n_504">
  <tok id="s_1_n_1">
    <orth>die</orth>
    <pos func="-" genau:text="ART"/>
    <desc>
      <morph genau:text="nsf"/>
    </desc>
  </tok>
  <ntNode id="s_1_n_501">
    <tok id="s_1_n_2">
      <orth>AWO</orth>
      <posfunc="-" genau:text="NN"/>
      [...]
</ntNode>
```

Fig. 1 An example from the TüBa-D/Z treebank (represented in the Tusnelda format) before (left) and after processing the resource with our normalization tools (right)

elements that do not contain primary data within their PCDATA content are transformed to the value of the attribute genau:text that is introduced by our tools. As Tusnelda documents comprise several levels of annotation in a single, monolithic XML element tree, the overall annotation is still extremely complex even though we perform a normalization procedure that includes the step described above. Therefore, we use additional processing methods to split the different conceptual levels, e.g. syntax, morphology, and named entities into multiple documents, that is, into a multi-rooted tree (Witt et al., 2007).

Another reason for the normalization procedure is that both hierarchical and timeline-based corpora (Bird and Liberman, 2001; Schmidt, 2005) need to be transformed into a common 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 (see Fig. 2). Our approach (Dipper et al., 2006; Schmidt et al., 2006; Wörner et al., 2006) 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 files, so that each file represents the information related to

a single annotation layer (Witt et al., 2007; Rehm et al., 2008d); this approach also guarantees that overlapping structures can be represented straightforwardly. Timeline-based corpora are also processed 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 are encoded as regular XML document instances, but, as a single corpus comprises *multiple* files, there is a need to go beyond the functionality offered by typical XML tools to enable us to process multiple files, as regular tools work with single files only (our approach for querying multi-rooted trees is described by Rehm et al., 2007a, 2008a).

The corpora that we process are marked up using several different markup languages and linguistic tag sets. As we want to enable users to query *multiple* corpora at the same time, we need to provide a unified view of the markup languages used in the original resources. For this sustainable operationalization of existing annotation schemes we employ the ontologies of linguistic annotation (OLiA, Chiarcos, 2008) approach (see Fig. 2): we built an OWL DL ontology that serves as a terminological reference. This reference model is based on the EAGLES recommendations for morphosyntax, the
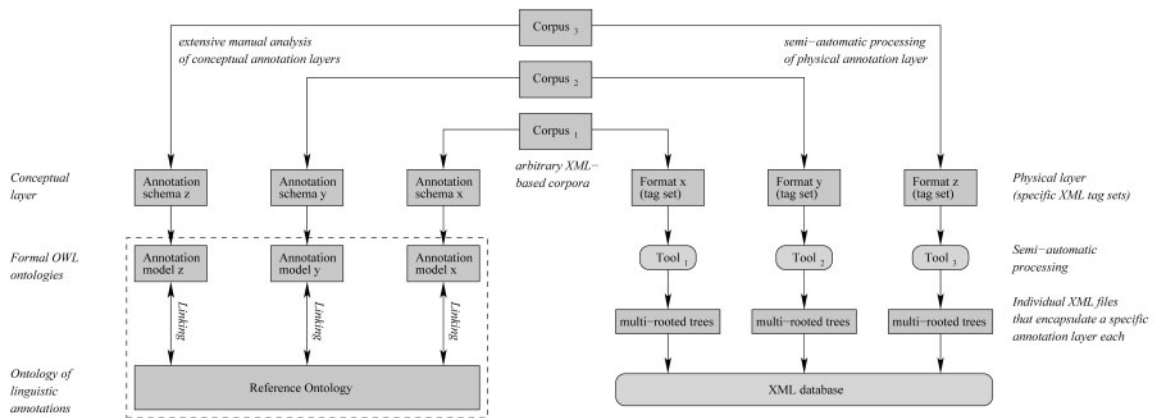
Fig. 2 The two main stages of processing a corpus: constructing an ontology-based formalization of the annotation model (left) and normalization as well as transformation of the physical corpus annotations into multi-rooted trees (right)

general ontology for linguistic description (Farrar and Langendoen, 2003), and the LISA annotation standard (Dipper et al., 2007). It covers reference specifications for word classes, and morpho-syntax, and is currently extended to syntax and information structure. The OLiA reference model represents a terminological backbone that different annotations are linked to and consists of three components: a taxonomy of linguistic categories (classes such as Noun, CommonNoun), a taxonomy of grammatical features (classes, e.g. Accusative), and relations (properties, e.g. hasCase). An OLiA annotation model is an ontology that represents one specific annotation scheme. Rehm et al. (2008a) describe the integration of the ontology into the overall querying environment.

Almost all resources that we process are linguistic corpora and treebanks. In addition, there are a few resources that belong to different data types. Four SFB 441 projects construct sentence collections that consist of, for example, suboptimal syntactic constructions taken from the linguistic literature and annotated with grammaticality judgements, or sentences that have a specific kind of verb phrases such as the stative passive. Furthermore, multiple projects create lexicons, some of which are augmented with empirical judgements gathered in online experiments. In a secondary line of research, we develop generic XML-based representation formats for these

types of linguistic resources for which we also implement query and visualization methods to be used within SPLICR. Our representation format for lexicons is based on the TEI P5 (Burnard and Bauman, 2007) guidelines and constructed using the Roma tool (http://www.tei-c.org/Roma/).

3.2 Normalization of metadata records

The separation of the individual annotation layers contained in a corpus has serious consequences with regard to legal issues (Zimmermann and Lehmberg, 2007; Lehmberg et al., 2007a,b, 2008; Rehm et al., 2007b): due to copyright and personal rights specifics that usually apply to a corpus' primary data we provide a fine-grained access control layer to regulate access by means of user accounts and access roles. We have to be able to explicitly specify 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 (Rehm et al., 2007b,c).

Our generic metadata schema, eTEI, is based on the TEI P4 header (Sperberg-McQueen and Burnard, 2002) and extended by a set of additional requirements. Both eTEI records and the corpora are stored in an XML database. The underlying assumption is that XML-annotated datasets are

more sustainable than, for example, data stored in a proprietary relational database management system. The main difference between eTEI and other approaches to representing metadata is that our generic eTEI metadata schema, currently formalized as a single document type definition (DTD), can be applied to five different levels of metadata description (Trippel, 2004; Himmelmann, 2006). A single eTEI file contains information on one of the following levels:

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

The pressing need for these five levels of metadata description can be illustrated using the ambiguity of the 'author' concept: while *setting* refers to a specific communication situation, the author of *raw data* can be the author of a certain book or the speaker whose monologue has been recorded. The author of *primary data* is the person who transcribed the raw data into one or more digital files. The authors of individual *annotation* files are those who analyse, interpret, and annotate the primary data (usually linguists, student assistants, or PhD students) with the help of specialized tools and the author of the *corpus* is the person who is responsible for constructing or collecting the corpus data (for example, the principal investigator of a research project). Important and relevant metadata exist on all five levels and can be captured using the eTEI approach.

We devised a workflow that helps users edit eTEI records (Rehm *et al.*, 2008d). Its primary components are the eTEI DTD and the oXygen XML editor. Based on annotations contained in the DTD we can generate automatically an empty XML document with embedded documentation and a Schematron schema. The Schematron specification can be used to check whether all elements and attributes instantiated in an eTEI document conform to the current level of metadata description.

4 The Sustainability Platform SPLICR

The sustainability platform consists of a front-end and a back-end. The front-end is the user visible part and is realized using JSP (Java Server Pages), JavaScript, and Ajax technologies. It runs in the user's browser and provides functions for searching and exploring metadata records and corpus data. The back-end is a web application that runs on top of the Tomcat application server.

In the following, we describe SPLICR's staging area and general architecture (Section 4.1) as well as its back-end (Section 4.2) and front-end (Section 4.3).

4.1 The staging area

SPLICR is the result of a joint project between the universities of Hamburg, Potsdam/Berlin, and Tübingen. While the system is primarily developed in Tübingen, it is meant to be used to import and make available corpora and additional resources from all different sites. For this reason we put special emphasis on making sure that all corpora from all sites can be imported into the platform and that the distributed research staff can carry out these procedures on their own.

First, all corpora and metadata are manipulated according to the processing steps described in Section 3 (see the lower part in Fig. 3). This also includes the creation of a manifest file for each corpus. They are represented in a simple XML format and act as a corpus inventory list that also contains a name and short description of a resource. Manifest files are generated semi-automatically by the normalization tools, their contents are used by the front-end and by the import and export tools. Each corpus consists of five parts: the manifest file, multiple files that contain the processed corpus data, multiple files that contain the eTEI metadata records, the original and unchanged corpus files (including the original metadata files,
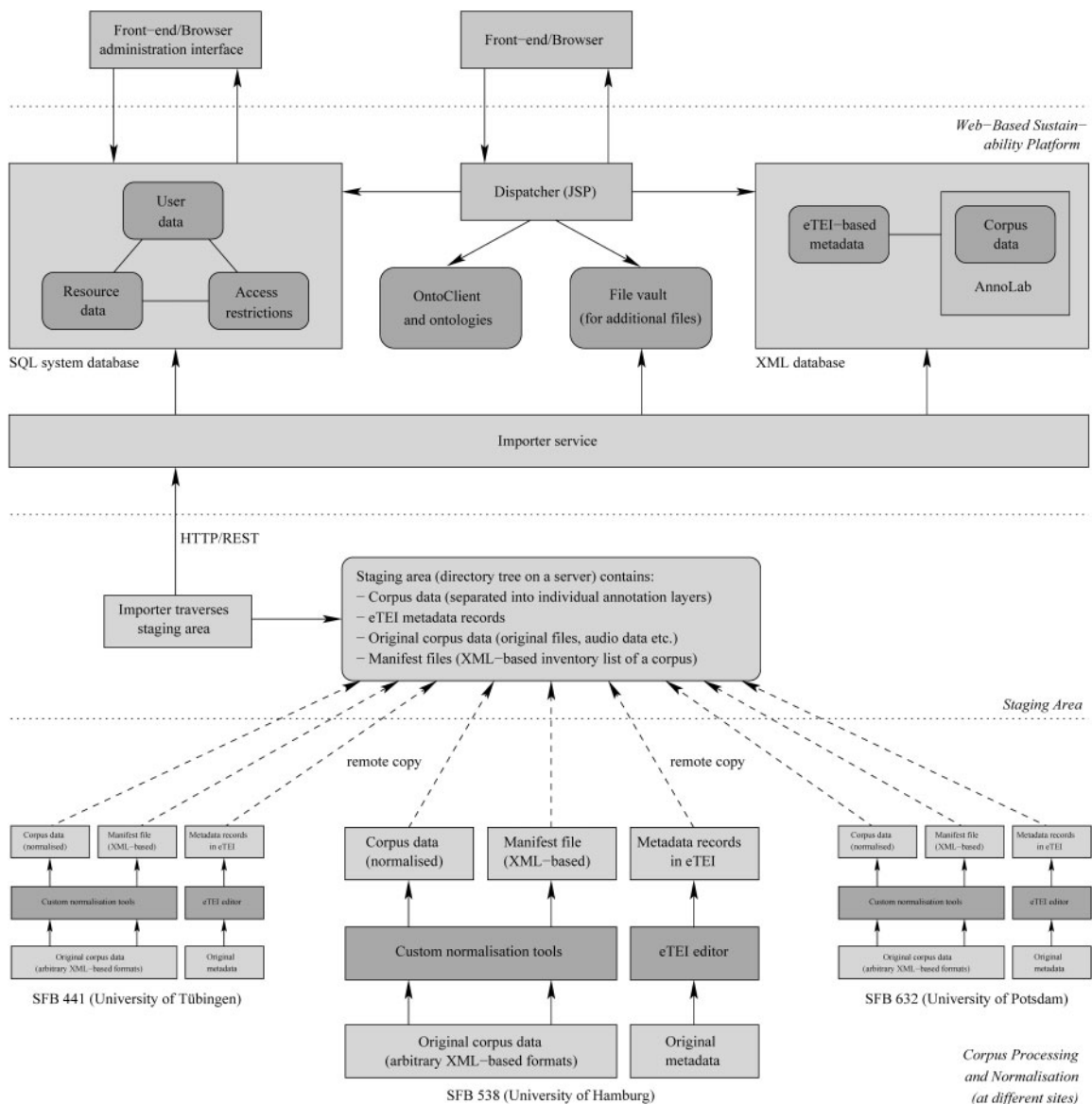
Fig. 3 Resource normalization, the staging area, and the primary SPLICR components

documentation, etc.), and stylesheets as well as log-files that were used for or generated by the corpus transformation phase. Then, each site prepares a local directory tree that houses all corpus directories and files. The structure of this directory tree (see Fig. 4) as well as strict naming conventions for files and folders are defined by a shared technical specification so that the tree can be traversed and processed

automatically. Afterwards, the three local directory trees are copied into the staging area, stored on our central project server which is located in Tübingen (see the middle part in Fig. 3). From here we use an importer tool to traverse the staging area fully automatically. The importer tool checks, among others, the data for consistency and imports the corpus data and metadata records into the back-end.
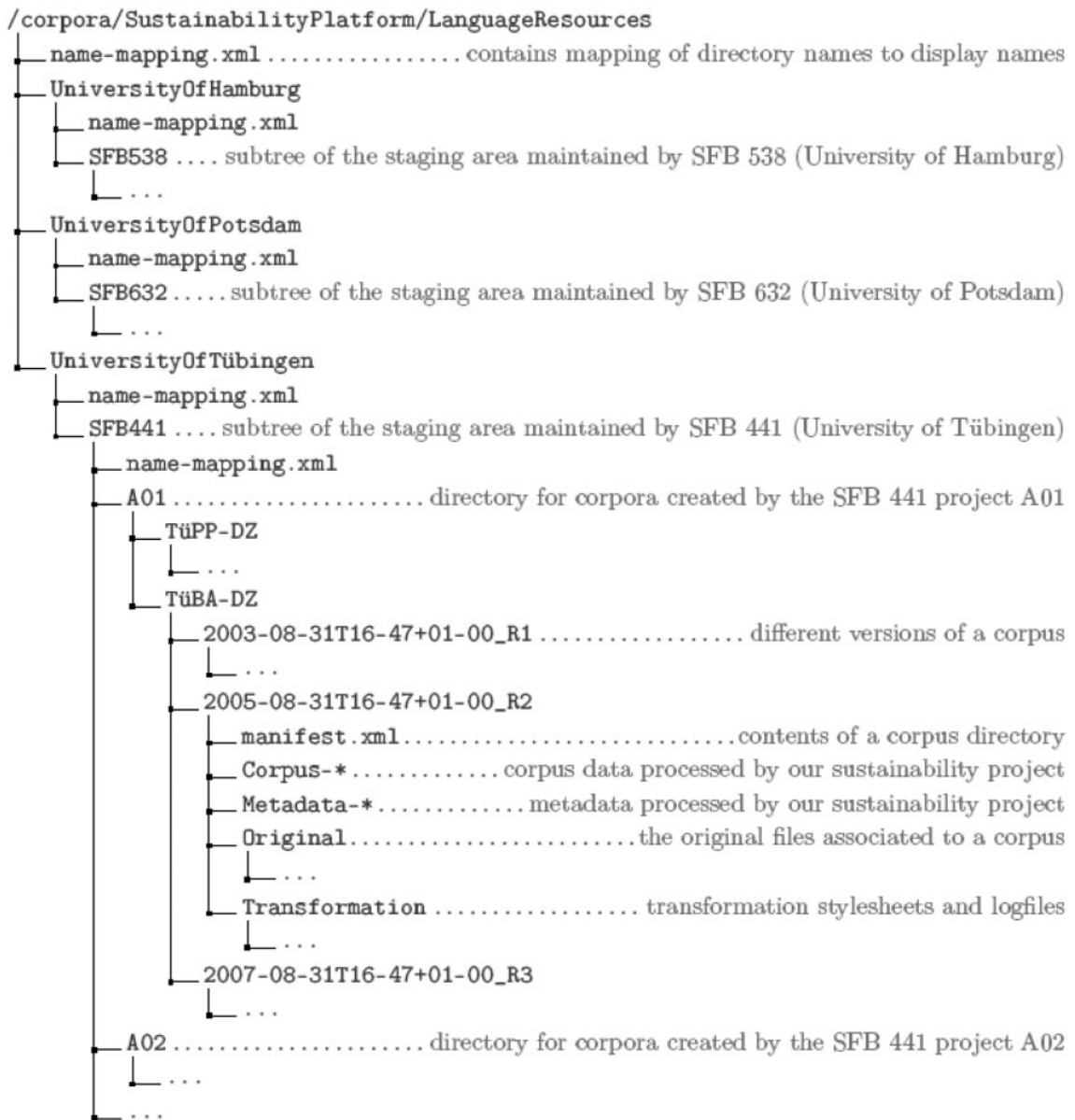
Fig. 4 The strictly specified directory structure of the staging area (excerpt)

4.2 The back-end

The back-end hosts the JSP files and related data. It accesses two different databases, the corpus database and the system database. Furthermore, we implemented a file vault that stores additional files that belong to a specific corpus, such as the original corpus data files, PDF files that act as documentation, and transformation scripts, among others. Several servlets provide means for exchanging information between the front-end and the back-end. The back-end is implemented as a web application that runs on top of Apache's Tomcat servlet container.

The corpus database is an eXist XML database, extended by the AnnoLab system (Eckart and Teich, 2007), in which all XML-encoded resources and metadata are stored. The back-end communicates with the corpus database using the XML:DB interface. The XML files that we processed (see Section 3.1) and that belong to a single annotation layer of a specific corpus are stored in a single collection. For the system database we use a relational database (MySQL) that contains data about user accounts and acts as a catalogue for corpus data. It stores information about single files in a corpus, resource groups (i.e. corpora), and access rights. A specific user can only access a specific resource if the permissions for this user/resource pair allow this operation. The file vault stores all additional resources (i.e. original corpus files, etc.) in regular files in a directory used by SPLICR exclusively. The corpus database and the file vault are regarded as a single component in the back-end and are accessed using a unified vault interface. Depending on their type, resources are transparently stored either in the corpus database or in the file vault. All operations that access these resources are carried out through this interface, so other back-end components do not need to communicate with either the corpus database or the file vault explicitly.

The import service servlet is the remote endpoint for the importer client (see Section 4.1). The import tool processes the staging area and sends each resource to the import service in a file-by-file fashion. Depending on the URI used and the HTTP headers set, the type of each file is determined and the appropriate records in the system database are created. Furthermore, files are stored in the appropriate location using the vault interface. The query service component is responsible for the asynchronous processing of corpus queries. A query is carried out in the following way: the front-end sends a JSON representation of the query and a list of the corpora currently selected by the user to the query dispatcher servlet. The servlet transforms the query into XQuery by generating, for every single file of all selected corpora, a dedicated XQuery expression. If the user does not have the permissions to access a certain file, it will be skipped. This set of XQuery expressions is linked to a query job which is executed using a worker-thread of the query service component. At the same time, a unique query ID is returned to the front-end, which will start polling results. The XQuery expressions are run sequentially against the corpus database. Results are buffered within the back-end until the front-end fetches them. This approach enables us to provide a much better user experience, since the user can already start exploring the first result even though the system is still running queries on the remaining files (see the progress bar at the bottom of the browser in Figs 9–11). A query monitor exists in the administration area of the front-end. It allows the administrator to display all currently running query jobs with additional details such as average query runtime per file and estimated remaining total runtime. Furthermore, the administrator can cancel query jobs. Corpus data is delivered either using the get service or transformation service servlets. If the user has access to a file, the get service servlet fetches the resource from the vault and sends it to the browser (we need this function for providing corpus download links in the front-end). The transformation service servlet applies an XSLT transformation to a specific file before it sends the result to the browser. Currently, this servlet is used for the HTML visualization of the eTEI metadata records.

4.3 The front-end

The target user group of our system consists of researchers who work in linguistics, computational linguistics, and related fields. SPLICR needs to be able to support its users answering the following questions (see Fig. 5): Which linguistic resources are stored in the platform? Can one or more of these corpora be used as empirical data bases for a specific research question one is working on? What is the extent of the annotations of these resources and do they cover what is needed for one's research endeavour? We meet this intended usage scenario with the following areas of functionality: metadata exploration (Section 4.3.1), multiple methods of querying corpora (Section 4.3.2), as well as corpus download (Section 4.3.3).
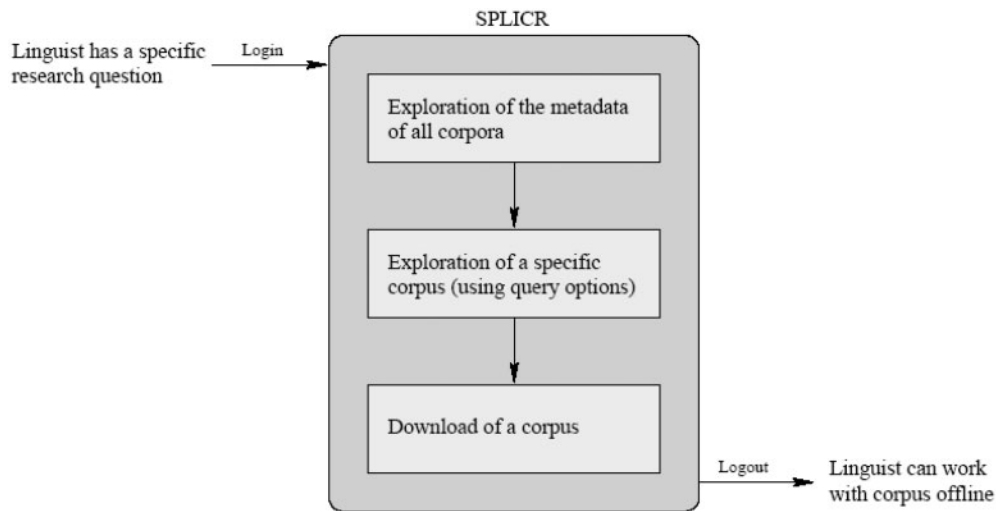
Fig. 5 One typical usage scenario of the sustainability platform SPLICR
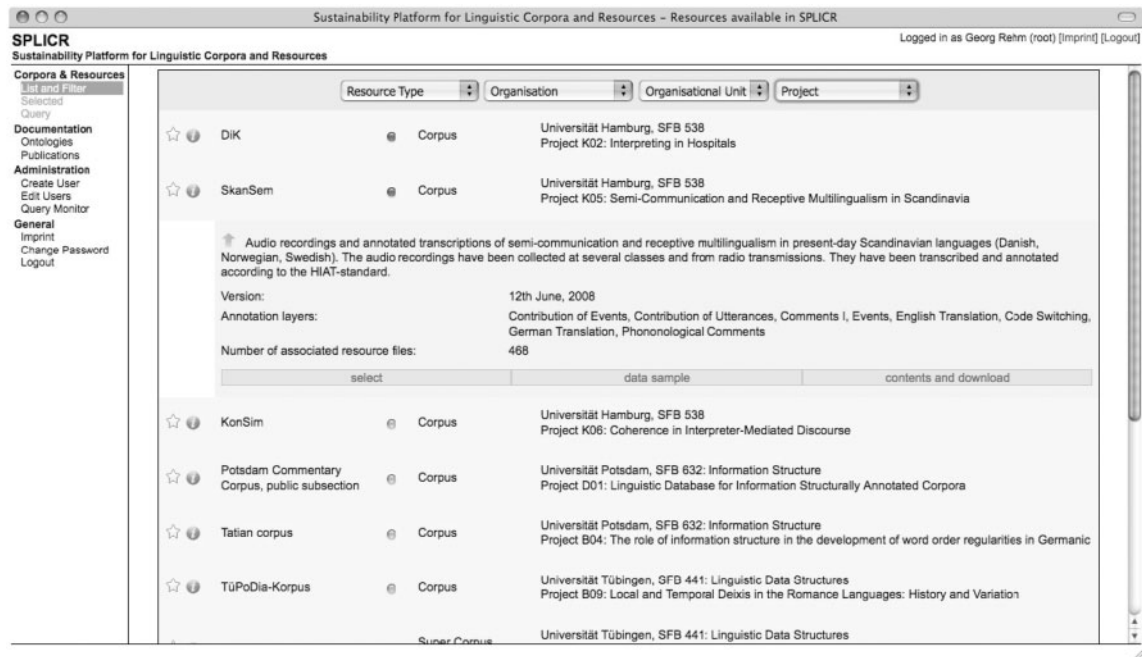

Fig. 6 The SPLICR front-end: a list of the available corpora and resources

4.3.1 *Metadata exploration*

As soon as a user logs onto the system he or she is presented with the complete list of resources currently stored in SPLICR (see Fig. 6).[2] Drop-down menus can be used to filter the list, for example, to restrict the view to the corpora from a specific organization or to those that contain a specific level of linguistic description. A click on the 'information'

icon expands the row that contains the name of the resource and its affiliation. This expanded view shows a brief description of the corpus, its version, the annotation layers, and the number of files associated with this resource. If the user wants to know more, the hyperlink 'contents and download' switches to a view that lists all files that belong to a corpus. From here, the eTEI-encoded XML metadata files can be displayed by automatically transforming them to XHTML (see Section 4.3.3).

4.3.2 Multiple methods of querying corpora

As we cannot expect our target users (i.e. linguists) to be proficient in XML query languages such as XPath and XQuery, we provide an intuitive query interface that generalizes from the underlying data structures and querying methods actually used. Before we started implementing SPLICR we collected a set of requirements and functions that the front-end should have by conducting in-depth interviews with the staff members of SFB 441 and by asking them to fill out a questionnaire (Soehn et al., 2008). The feedback we received to this questionnaire and during the interviews was structured and compiled into a set of features that we documented in a requirements document.

As soon as one or more corpora are selected in the list of resources by clicking the star icon or using the 'select' hyperlink (see Fig. 6), the user can access the query interface which is based on two main concepts. First, we provide three different kinds of search widgets (in increasing order of complexity: full-text search, concept search, tree fragment search). Second, the query interface supports multiple tabs that can be added and deleted at will. The idea behind this approach is that a complex search query can be constructed using arbitrary search widgets in multiple tabs. In reality, however, SPLICR currently only supports multiple tabs for the concept search query widget; to enable complex queries, these multiple tabs can be combined using '∧' and '∨' icons in a flexible way (not shown in any of the screenshots).

Query Widget: Full-Text Search. The full-text search query widget can be used to find certain words or simple patterns in corpora. Matches are highlighted in the result browser.

Query Widget: Concept Search. The concept search query widget presents a list of linguistic concepts that are contained in the individual annotation layers that make up a corpus (see Fig. 7). For example, if the user selects a certain corpus, the annotation layer to be queried needs to be selected from the floating 'Query Type & Layer' window ('Grammatical Function' in Fig. 7). As soon as one such layer is picked, the front-end presents the list of linguistic concepts that are encoded in this layer.[3] When the user selects one of these concepts, a second drop-down menu with a list of distinct values for this concept is dynamically filled. Finally, the user can select the context in which the matches are to be displayed. Depending on the selected corpus and annotation layer, the context can be 'words' only, 'phrases', 'sentences', 'paragraphs' or 'articles'.

Concept search is a very simple and user-friendly method of getting to know the actual annotations contained in a specific corpus. This querying mode is based on XML-encoded configuration files that represent the list of concepts, their values as well as a small set of corpus-specific return contexts. The respective set of information is represented as XPath fragments. At query time these fragments are combined into an XQuery expression that is evaluated against the XML database in which all corpora and resources are stored (see Fig. 3).

Query Widget: Tree-Fragment Search. The most sophisticated query widget is a fully-fledged interactive editor for constructing linguistic tree fragments that can be queried against the currently selected corpus (Rehm et al., 2008a). The structures defined by these graphs mirror the structures to be found; in Fig. 8 a query is constructed in which nodes of the phrase type NX (noun phrase) dominate nodes of the phrase type ADJX (adjective phrase).[4] The user can choose among several functions in order to build a tree in a step-by-step fashion. First, one or more node icons can be picked from the tool bar and placed on the assembly pane using drag and drop. Then, the user can connect these nodes with edges that represent dominance or precedence. In addition, one or more sets of attribute-value-pairs can be placed inside nodes in order to specify even more detailed
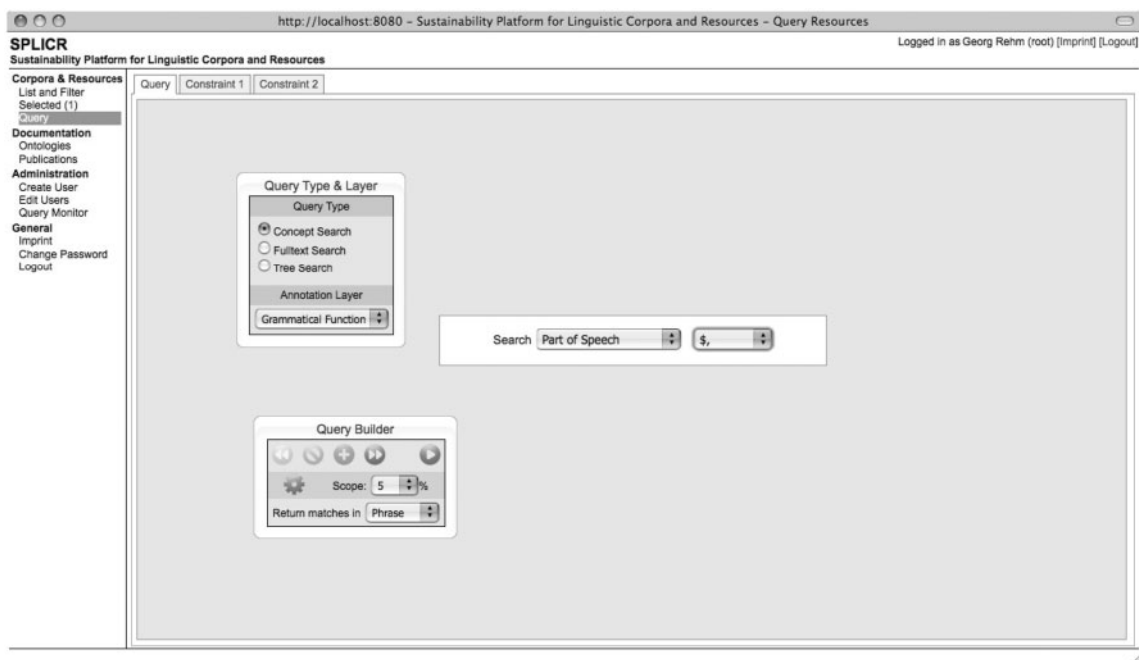
Fig. 7 The SPLICR front-end: the query mode 'concept search'
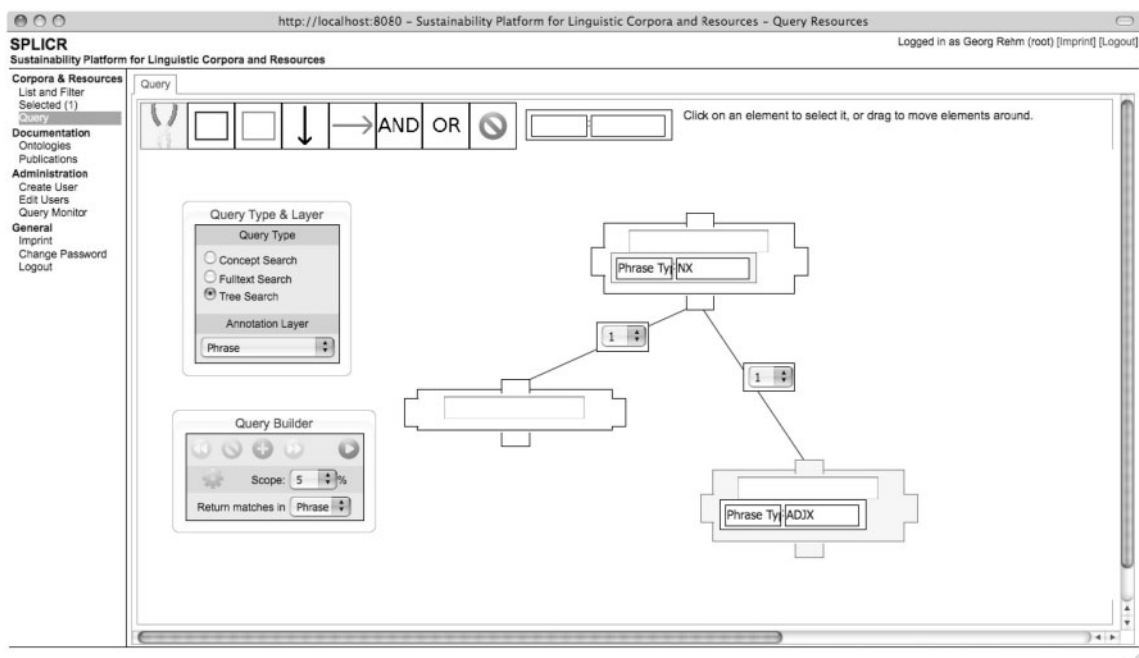

Fig. 8 The SPLICR front-end: the 'tree editor' query mode

constraints a node has to satisfy. For these constraints, the concept search configuration files (see the previous paragraph) are re-used so that the user is not expected to have intimate knowledge of the annotation scheme employed in the original corpus. The current state of the tree editor can be roughly compared to TIGERSearch's feature set (Lezius, 2002) enhanced by our specific requirements.

The tree editor is fully implemented in JavaScript extended by the frameworks Prototype and script.aculo.us. The editor communicates with the back-end via Ajax by posting XQuery requests to a servlet that runs on the back-end. The servlet responds with the XML-encoded matches, which are then interpreted by a variety of display modules (see below).

The Result Browser. We provide three different query widgets that can be used to search and query corpora. The results of these queries are visualized by the result browser that offers four different display modes: plain text view, XML view (see Fig. 9), box view (see Fig.10), and tree view (see Fig. 11).

While plain text view and XML view are self-explanatory, the box viewer is especially useful for visualizing transcribed speech data; additional information is presented in pop-up windows. The tree viewer was fully implemented in JavaScript and visualizes the tree structure of the XML document that was returned as a match by the XML database. The user can zoom in and out, attribute information can be displayed or hidden, and edges can be displayed in two different styles.

4.3.3 Corpus download

The corpus download facility is included in the view that displays the contents of a corpus (see Fig. 12). The contents are grouped into the sections 'processed data', 'metadata', 'original data', and 'transformation data' (see Section 4.1). These different sections can be further expanded in order to display the actual files that belong to a section. For the processed data area, an additional section is introduced, i.e. the annotation layers that are listed beneath the group of processed files (again, see Fig. 12). Should the user decide that he or she

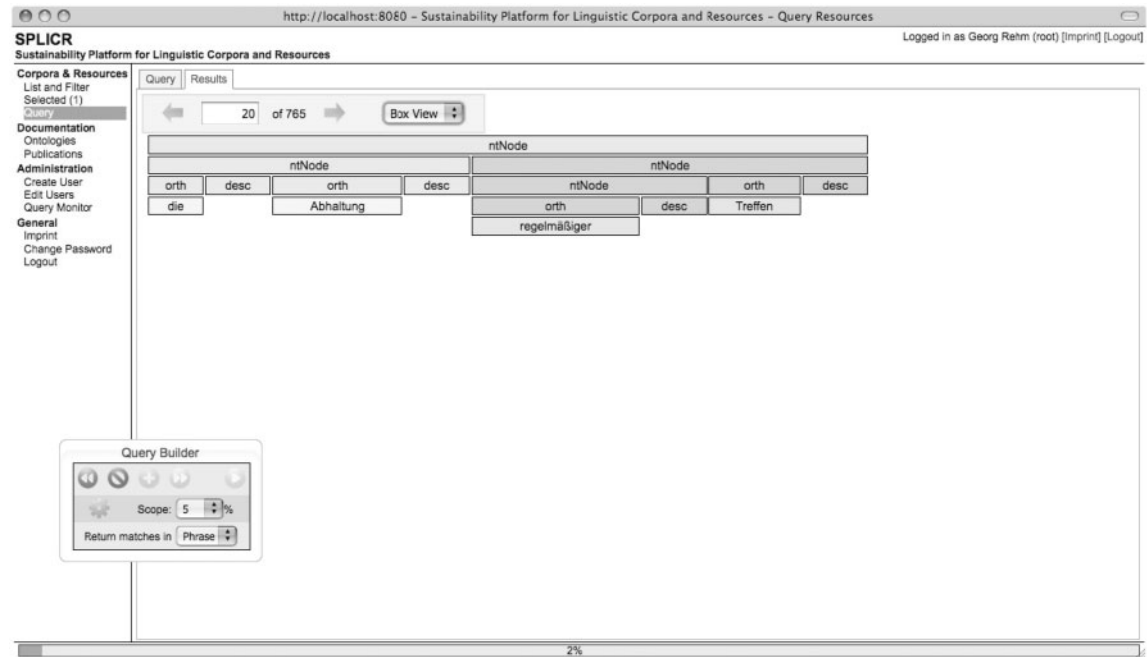**Fig. 9** The SPLICR front-end: result browser in 'XML viewer' mode

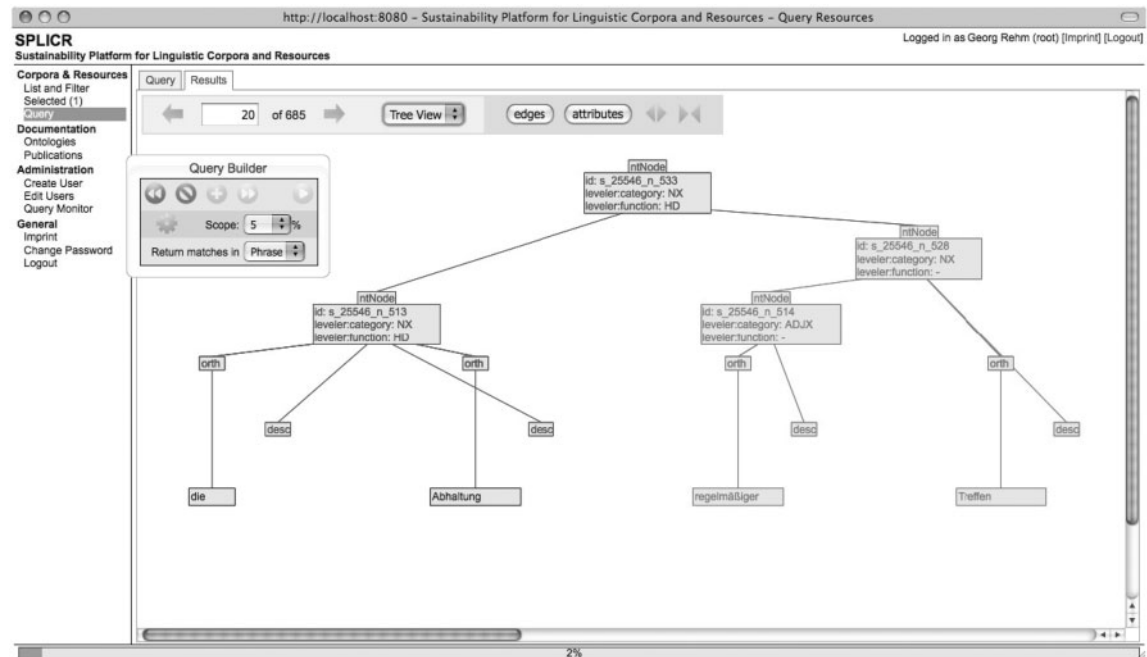**Fig. 10** The SPLICR front-end: result browser in 'box viewer' mode

Fig. 11 The SPLICR front-end: result browser in 'tree viewer' mode

Fig. 12 The SPLICR front-end: corpus contents and download mode

wants to work with a certain corpus outside of the platform using his or her own set of tools, the individual parts of a corpus can be downloaded as ZIP files by clicking on the disk icons. Furthermore, single files can also be downloaded by clicking the hyperlink 'source file'.

5 Concluding Remarks

The interoperability, reusability, and sustainability of electronic texts, multimodal data, and other digital resources is a crucial issue for all branches within Digital Humanities. Several initiatives have been launched—especially in the USA and in Europe—to address this topic, and the project 'Sustainability of Linguistic Resources', funded by the German Research Foundation (DFG), can be seen as one example of these recent developments. This article presents a practical result of this linguistic project, the sustainability platform SPLICR, a system that allows one to examine, query, and download

linguistic data sets. Although linguistic resources are special, most importantly because of the depth and complexity of their annotations, they are first and foremost a type of digital resource within the humanities. This is why we are convinced that other initiatives can benefit from our solutions to the problems we encountered in the course of this project.

The aspect of data normalization is rarely discussed in academic publications. This is mostly due to the fact that the conversion from one format into another is not regarded as a difficult or challenging task, because tools and specialized programming languages exist that support researchers in converting data sets from one format into another. In practice, however, it turns out that this rather time-consuming task is in fact of interest for researchers within Digital Humanities. The reason is that the specification of transformations may change the model according to which a text resource is annotated. Analogously to our experiences in dealing with data conversion, practical

work with heterogeneous linguistic resources revealed new insights for our work with metadata. While metadata for text resources is a topic that has been extensively discussed within Digital Humanities and, especially, in library science for decades, we had to devote a significant amount of effort to finding an appropriate solution for the task of representing metadata for normalized corpora. Our approach is based on the specification of metadata within the TEI framework but extends its methodology by specifying TEI headers for different components or layers of a linguistic resource. As a result, the metadata of a resource included in the sustainability platform are split over several header files, each of which deals with different aspects or components of the data that it describes, e.g. the setting describes a real-world situation the language resource is related to, whereas metadata for annotations describe, e.g. the names of the annotators or the name and version of the annotation software and the date and time at which the annotation was produced.

The detailed description of SPLICR, the sustainability platform for linguistic resources, can also be of relevance to other fields apart from linguistics. To give only one example, in order to store and maintain the more than fifty resources in the staging area, we developed a specification that contains, among others, strict naming conventions and a carefully designed directory tree. The automatic import of resources into a staging area that includes multiple consistency checks is an approach that we recommend to all projects that build large repositories of digital resources. To sum up, we are convinced that the results of this large linguistic project will not only help linguists dealing with digital resources but also researchers from other domains in the Digital Humanities.

References

Bird, S. and Liberman, M. (2001). A formal framework for linguistic annotation. *Speech Communication*, **33**(1/2): 23–60.

Bird, S. and Simons, G. (2003). Seven dimensions of portability for language documentation and description. *Language*, **79**: 557–82.

Burnard, L. and Bauman, S. (eds) (2007). TEI P5: *Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Initiative Consortium http://www.tei-c.org/release/doc/tei-p5-doc/html/ (last accessed on 6 March 2009).

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

Chiarcos, C. (2008). An ontology of linguistic annotations'. *LDV Forum*, **23**(1): 1–16.

Dipper, S., Götze, M., and Skopeteas, S. (eds) (2007). *Information Structure in Cross-Linguistic Corpora: Annotation Guidelines for Phonology, Morphology, Syntax, Semantics, and Information Structure*, ISIS-Working Papers of the SFB 632 vol. 7. Potsdam.

Dipper, S., Hinrichs, E., Schmidt, T., Wagner, A., and Witt, A. (2006). *Sustainability of Linguistic Resources*. In Hinrichs, E., Ide, N., Palmer, M., and Pustejovsky, J. (eds), *Proceedings of the LREC 2006 Satellite Workshop Merging and Layering Linguistic Information,* Genoa, Italy, 48–54.

Eckart, R. and Teich, E. (2007). *An XML-Based Data Model for Flexible Representation and Query of Linguistically Interpreted Corpora*. In Rehm, G., Witt, A., and Lemnitzer, L. (eds), *Datenstrukturen für linguistische Ressourcen und ihre Anwendungen – Data Structures for Linguistic Resources and Applications: Proceedings of the Biennial GLDV Conference 2007*. Tübingen: Gunter Narr, pp. 327–36.

Farrar, S. and Langendoen, T. (2003). A linguistic ontology for the semantic web. *GLOT International*, **3**: 97–100.

Garside, R., Leech, G., & McEnery, A. (eds) (1997). *Corpus Annotation – Linguistic Information from Computer Text Corpora*. London, New York: Longman.

Himmelmann, N. P. (2006). *Daten und Datenhuberei*. Keynote speech, 28th annual meeting of the DGfS, University of Bielefeld.

Ide, N., Bonhomme, P., and Romary, L. (2000). *XCES: An XML-based Standard for Linguistic Corpora, Proceedings of the Second Language Resources and Evaluation Conference (LREC)*, Athens, pp. 825–30.

Lehmberg, T., Chiarcos, C., Hinrichs, E., Rehm, G., and Witt, A. (2007a). Collecting Legally Relevant Metadata by Means of a Decision-Tree-Based Questionnaire System. In Schmidt, S., Siemens, R., Kumar, A., and

Unsworth, J. (eds), *Digital Humanities 2007*. ACH, ALLC, Urbana-Champaign, IL, USA: Graduate School of Library and Information Science, University of Illinois, Urbana-Champaign, pp. 164–6.

Lehmberg, T., Chiarcos, C., Rehm, G., and Witt, A. (2007b). *Rechtsfragen bei der Nutzung und Weitergabe linguistischer Daten*. In Rehm, G., Witt, A., and Lemnitzer, L. (eds), *Datenstrukturen für linguistische Ressourcen und ihre Anwendungen—Data Structures for Linguistic Resources and Applications: Proceedings of the Biennial GLDV Conference 2007*. Tübingen: Gunter Narr, pp. 93–102.

Lehmberg, T., Rehm, G., Witt, A., and Zimmermann, F. (2008). Digital text collections, linguistic research data, and mashups: notes on the legal situation. *Library Trends*, **57**(1): 52–71.

Lehmberg, T. and Wörner, K. (2009) Annotation Standards. In Lüdeling, A. and Kytö, M. (eds), *Corpus Linguistics*. Berlin, New York: de Gruyter, Handbücher zur Sprach- und Kommunikationswissenschaft (HSK).

Lezius, W. (2002). *Ein Suchwerkzeug für syntaktisch annotierte Textkorpora*. Ph.D. thesis, University of Stuttgart. Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung, vol. 8, No. 4.

Rehm, G., Eckart, R., and Chiarcos, C. (2007a). An OWL- and XQuery-based Mechanism for the Retrieval of Linguistic Patterns from XML-Corpora. In Angelova, G., Bontcheva, K., Mitkov, R., Nicolov, N., and Nikolov, N. (eds), *International Conference Recent Advances in Natural Language Processing (RANLP 2007)*. Borovets, Bulgaria, 510–514.

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

Rehm, G., Schonefeld, O., Witt, A., Chiarcos, C., and Lehmberg, T. (2008b). A web-platform for preserving, exploring, visualising and querying linguistic corpora and other resources. *Procesamiento del Lenguaje Natural*, **41**:155–162. (Proceedings of SEPLN 2008 – 24th edition of the Conference of the Spanish Society for Natural Language Processing, 10–12 September, Madrid).

Rehm, G., Schonefeld, O., Witt, A., Chiarcos, C., and Lehmberg, T. (2008c). SPLICR: A Sustainability Platform for Linguistic Corpora and Resources. In Storrer, A., Geyken, A., Siebert, A., and Würzner, K.-M. (eds), *KONVENS 2008 (Konferenz zur Verarbeitung natürlicher Sprache) – Textressourcen und lexikalisches Wissen,* Berlin, pp. 86–95.

Rehm, G., Schonefeld, O., Witt, A., Lehmberg, T., Chiarcos, C., Bechara, H., Eishold, F., Evang, K., Leshtanska, M., Savkov, A., and Stark, M. (2008d). *The Metadata-Database of a Next Generation Sustainability Web-Platform for Language Resources, Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008)*, Marrakech, Morocco.

Rehm, G., Witt, A., Zinsmeister, H., and Dellert, J. (2007b) Corpus Masking: Legally Bypassing Licensing Restrictions for the Free Distribution of Text Collections. In Schmidt, S., Siemens, R., Kumar, A., and Unsworth, J. (eds), *Digital Humanities 2007*. ACH, ALLC, Urbana-Champaign, IL, USA: Graduate School of Library and Information Science, University of Illinois, Urbana-Champaign, pp. 166–70.

Rehm, G., Witt, A., Zinsmeister, H., and Dellert, J. (2007c). *Masking Treebanks for the Free Distribution of Linguistic Resources and Other Applications, Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories (TLT 2007)*, Bergen, Norway, number 1 in Northern European Association for Language Technology Proceedings Series, pp. 127–38.

Schmidt, T. (2005). *Time Based Data Models and the Text Encoding Initiative's Guidelines for Transcription of Speech. Working Papers in Multilingualism, Series B 62.*

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

Soehn, J.-P., Zinsmeister, H., and Rehm, G. (2008). *Requirements of a User-friendly, General-purpose Corpus Query Interface*. In Burnard, L., Choukri, K., Rehm, G., Schmidt, T., and Witt, A. (eds), *Proceedings of the LREC 2008 Workshop Sustainability of Language Resources and Tools for Natural Language Processing,* Marrakech, Morocco.

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

Telljohann, H., Hinrichs, E., and Kübler, S. (2004) *The TüBa-D/Z Treebank—Annotating German with a*

Context-Free Backbone, *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal.

Trilsbeek, P. and Wittenburg, P. (2006). Archiving Challenges. In Gippert, J., Himmelmann, N. P., and Mosel, U. (eds), *Essentials of Language Documentation*. Berlin, New York: Mouton de Gruyter, pp. 311–35.

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

Wagner, A. (2005). Unity in Diversity: Integrating Differing Linguistic Data in TUSNELDA. In Dipper, S., Götze, M., and Stede, M. (eds), *Heterogeneity in Focus: Creating and Using Linguistic Databases*. Potsdam, vol. 2 of ISIS (Interdisciplinary Studies on Information Structure), Working Papers of the SFB 632, pp.1–20.

Witt, A., Rehm, G., Hinrichs, E., Lehmberg, T., and Stegmann, J. (2009) SusTEInability of linguistic resources through feature structures. *Literary and Linguistic Computing*.

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

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

Zimmermann, F. and Lehmberg, T. (2007). Language Corpora—Copyright—Data Protection: The Legal Point of View. In Schmidt, S., Siemens, R., Kumar, A., and Unsworth, J. (eds), *Digital Humanities 2007*. ACH, ALLC, Urbana-Champaign, IL, USA: Graduate School of Library and Information Science, University of Illinois, Urbana-Champaign, pp. 162–4.

Zinsmeister, H., Kübler, S., Hinrichs, E., and Witt, A. (2008). Linguistically Annotated Corpora: Quality Assurance, Reusability and Sustainability. In Lüdeling, A. and Kytö, M. (eds), *Corpus Linguistics*. Berlin, etc.: de Gruyter, HSK.

Notes

1 The authors would like to thank Christian Chiarcos (University of Potsdam) and Timm Lehmberg (University of Hamburg) for their work and commitment at the two remote project sites in Potsdam and Hamburg. Furthermore, we would like to thank Johannes Dellert, Kilian Evang, Magdalena Leshtanska, and Aleksandar Savkov (University of Tübingen) for taking care of extensive corpus data conversions and implementing part of the software described in this article. Finally, we would like to thank Richard Eckart (TU Darmstadt) for providing AnnoLab and supporting us in the integration of the software into our sustainability platform called SPLICR.

2 The screenshots shown in Figs 6–12 were made using a SPLICR installation that we set up for demonstration purposes; it does not comprise the complete list of corpora that we processed in the context of the project.

3 The value for the concept 'Part of Speech' shown in Fig. 7, '\$,', is used to annotate a comma in the treebank TüBa-D/Z (see the annotation guidelines for this resource available at http://www.sfs.uni-tuebingen.de/en_tuebadz.shtml). Other, probably more intuitive examples are 'adv', 'art', and 'vvfin'.

4 Figures 9–11, show some of the results this query produced with regard to the treebank TüBa-D/Z (see Section 3.1).