

***International Symposium on XML for the Long Haul:
Issues in the Long-term Preservation of XML***
Proceedings

Sustainability of Linguistic Resources Revisited

Georg Rehm

DFKI

<georg.rehm@dfki.de>

Oliver Schonefeld

Institute for the German Language (IDS), Mannheim

<schonefeld@ids-mannheim.de>

Thorsten Trippel

Tübingen University

<thorsten.trippel@uni-tuebingen.de>

Andreas Witt

Institute for the German Language (IDS), Mannheim

<witt@ids-mannheim.de>

***International Symposium on XML for the Long Haul: Issues in the Long-term Preservation
of XML***

August 2, 2010

Copyright © 2010 by the authors. Used with permission.

How to cite this paper

Rehm, Georg, Oliver Schonefeld, Thorsten Trippel and Andreas Witt. "Sustainability of Linguistic Resources Revisited."
Presented at International Symposium on XML for the Long Haul: Issues in the Long-term Preservation of XML, Montréal,
Canada, August 2, 2010. In *Proceedings of the International Symposium on XML for the Long Haul: Issues in the Long-term
Preservation of XML*. Balisage Series on Markup Technologies, vol. 6 (2010). DOI: 10.4242/BalisageVol6.Witt01.

Abstract

Table of Contents

Introduction

Case Study: The Project "Sustainability of Linguistic Resources"

 Normalization of Linguistic Resources

 Normalization of Metadata Records

 Architecture

 SPLICR: Concluding Remarks

XML and Sustainability: Problems and Solutions

 Problem: Stand-off Annotation

 Problem: Machine-Generated XML

 Problem: Proprietary Tag Sets

 Problem: Availability and Findability

 Problem: Selection and Qualification for Long-Term Archiving

 Additional Pitfalls

Conclusions

Introduction

This paper discusses work on the sustainability of linguistic resources as it was conducted in various projects, including the work of a three year project *Sustainability of Linguistic Resources* which finished in December 2008, a follow-up project, *Sustainable linguistic data*, and initiatives related to the work of the International Organization of Standardization (ISO) on developing standards for linguistic resources. The individual projects have been conducted at German collaborative research centres at the Universities of Potsdam, Hamburg and Tübingen, where the sustainability work was coordinated.

Today, most language resources are represented in XML. The representation of data in XML is an important prerequisite for long-term preservation but a reasonable representation format such as XML alone is not sufficient. Though XML is being said to be human-readable it is obvious that legibility is a rather problematic notion in terms of photos encoded in SVG, complex structures generated from data dumps of databases and other applications or even formats such as Office Open XML. In the linguistic data community, various flavours of stand-off annotation also demonstrate the complexity of the problem.

Usually these data formats are not meant to be read by humans, though the advantages mentioned in XML-introductions still hold, namely, that data modelled according to the standardized and continuously maintained XML formalism can be read and analysed by human users to re-engineer tools using simple parsers for validation and mental effort.

Case Study: The Project “Sustainability of Linguistic Resources”

This section briefly presents SPLICR, the Web-based Sustainability Platform for Linguistic Corpora and Resources aimed at researchers who work in Linguistics or Computational Linguistics: a comprehensive database of metadata records can be explored and searched in order to find language resources that could be appropriate for one’s specific research needs. SPLICR also provides a graphical interface that enables users to query and to visualise corpora.

The project in which SPLICR was developed aimed at sustainably archiving the language resources that were constructed in three 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 – corpora and treebanks mostly. According to our estimates it took more than one hundred person years to collect and to annotate these datasets. The project had two main goals: (a) To process and to sustainably archive the resources so that they are still available to the research community and other interested parties in five, ten, or even 20 years time. (b) To enable researchers to query the resources both on the level of their metadata as well as on the level of linguistic annotations. In more general terms, the main goal was to enable solutions that leverage the interoperability, reusability, and sustainability of a large collection of heterogeneous language resources.

One of the obstacles we were confronted with was 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 normalise the corpora and their metadata records.

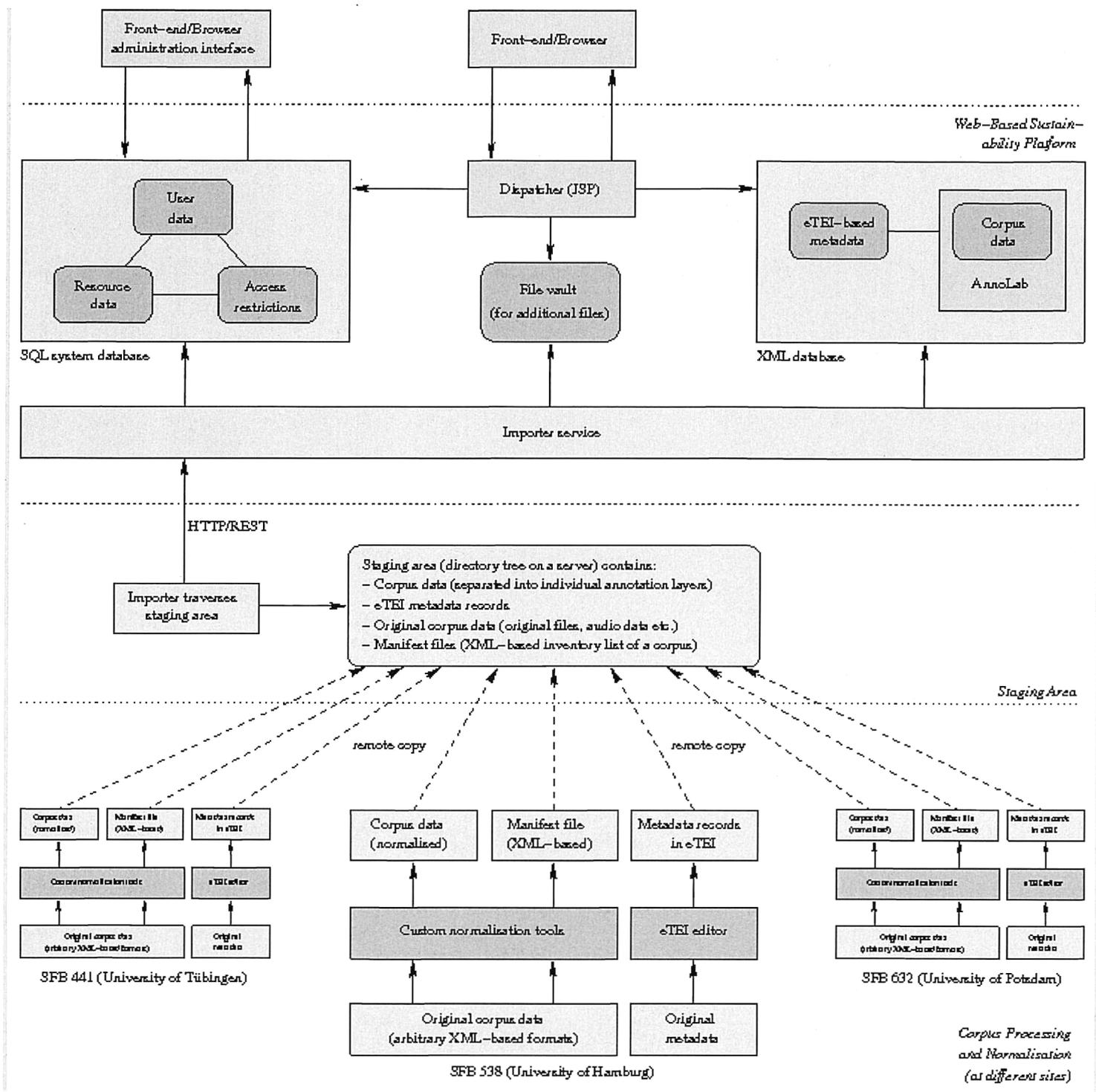
Normalization of Linguistic Resources

Language resources are nowadays usually built using XML-based representations and contain several concurrent annotation layers that correspond to multiple levels of linguistic description (e.g., part-of-speech, syntax, coreference). Our approach included 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 used 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.

A second reason for the normalization procedure was that both hierarchical and timeline-based corpora needed to be transformed into a common annotation approach, because we wanted our users to be able to query both types of resources at the same time and in a uniform way. The 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 files, so that each file represented the information related to a single annotation layer; this approach also guaranteed that overlapping structures can be represented straightforwardly. Timeline-based corpora were also processed in order to separate graph annotations. This approach enabled us to represent arbitrary types of XML-annotated corpora as individual files, i.e., individual XML element trees. These were encoded as regular XML document instances, but, as a single corpus comprises multiple files, there was 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. The normalization process is described in more detail in Witt et al. 2007.

Figure 1: Resource normalization and SPLICR's staging area.



Normalization of Metadata Records

The separation of the individual annotation layers contained in a corpus has serious consequences with regard to legal issues: due to copyright and personal rights specifics that usually apply to a corpus's primary data we provided a fine-grained access control layer to regulate access by means of user accounts and access roles. We had 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 they might be copyright-protected.

The generic metadata schema used for SPLICR, named *eTEI*, was based on the TEI P4 header and extended by a set of additional requirements. We decided to store both eTEI records and also the corpora in an XML database. The underlying assumption was that XML-annotated datasets are more sustainable than, for example, data stored in a proprietary relational DBMS. The main difference between eTEI and other approaches is that the generic eTEI metadata schema, formalized as a document type definition (DTD), can be applied to five different levels of

description. One eTEI file contains information on one of the following levels: (1) setting (recordings or transcripts of spoken language, 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; (5) a corpus (consists of primary data with one or more annotation levels). We devised a workflow that helps users to edit eTEI records. The workflow's primary components were the eTEI DTD and the Oxygen XML editor. Based on structured annotations contained in the DTD we automatically generate an empty XML document with embedded documentation and a Schematron schema. The Schematron specification is used to check whether all elements and attributes instantiated in an eTEI document conform to the current level of metadata description.

Architecture

The sustainability platform SPLICR consists of a front-end and a back-end. The front-end is the part visible to the user and is realized using JSP (Java Server Pages) and Ajax technology. It runs in the user's browser and provides functions for searching and exploring metadata records and corpus data. The back-end hosts the JSP files and related data. It accesses two different databases, the corpus database and the system database, as well as a set of ontologies and additional components. The corpus database is an XML database, extended by AnnoLab, an XML/XQuery-based corpus query and management framework that was specifically designed to deal with multiple possibly concurrent annotation layers, in which all resources and metadata are stored. 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. A specific user can only access a specific resource if the permissions for this user/resource tuple allow it.

SPLICR: Concluding Remarks

The corpus normalization and preprocessing phase in this project started in early 2007 and was finished in May 2008, the process of transforming the existing metadata records into the eTEI format was completed in June 2008. Work on the querying engine and integration of the XML database, metadata exploration and on the graphical visualization and querying front-end as well as on the back-end was carried out in the summer of 2008; a first prototype of the platform was finished in October 2008. Rehm et al. 2009 gives a more detailed description of the project.

XML and Sustainability: Problems and Solutions

Problem: Stand-off Annotation

Stand-off markup refers to the physical separation of annotations and text. Piotr Bański described this technique thoroughly at Balisage 2010 (Bański 2010). Stand-off annotation allows for marking up text without altering it by the inclusion of markup. It is the opposite approach to inline or embedded markup that was one of the principle ideas behind SGML and its successor XML. The term *stand-off annotation* was introduced by Henry Thompson and David McKelvie in 1997 (Thompson & McKelvie 1997), however the principles of this technique are even older, since, e.g., the linking mechanisms described in TEI P3 already allowed to mark up texts by linking annotations to text regions. Within the last couple of years the use of stand-off markup became predominant, especially for complex linguistic annotations.

Linguistically annotated corpora use stand-off markup extensively. Stand-off is also predominant within the forthcoming ISO standard "Linguistic Annotation Framework" (LAF, Ide & Romary 2007).

Figure 2: LAF based linguistic annotation

```
<!-- base segmentation -->
<region id="r42" a="24 35"/>
<!-- annotation over the base segmentation -->
```

```

<node id="n16">
  <f name="pos" value="NN"/>
</node>
<edge from="n16" to="r42"/>

<!-- annotation over another annotation -->

<node id="n23">
  <f name="synLabel" value="NP"/>
  <f name="role" value="-SBJ"/>
</node>
<edge from="n23" to="n16"/>

<!-- ... -->

```

Example of linguistic stand-off annotation (see Trippel et al. 2007)

Stand-off annotation has witnessed an increase in use due to the advantages of this approach (see Bański 2010 and Bański & Przepiórkowski 2009), but considering the sustainability and interoperability point of view, there are quite a few disadvantages (see Witt 2004):

- very difficult to read for humans
- the information, although included, is difficult to access using generic methods
- limited software support as standard parsing or editing software cannot be employed
- standard document grammars can only be used for the level which contains both markup and textual data
- new layers require a separate interpretation
- layers, although separate, often depend on each other

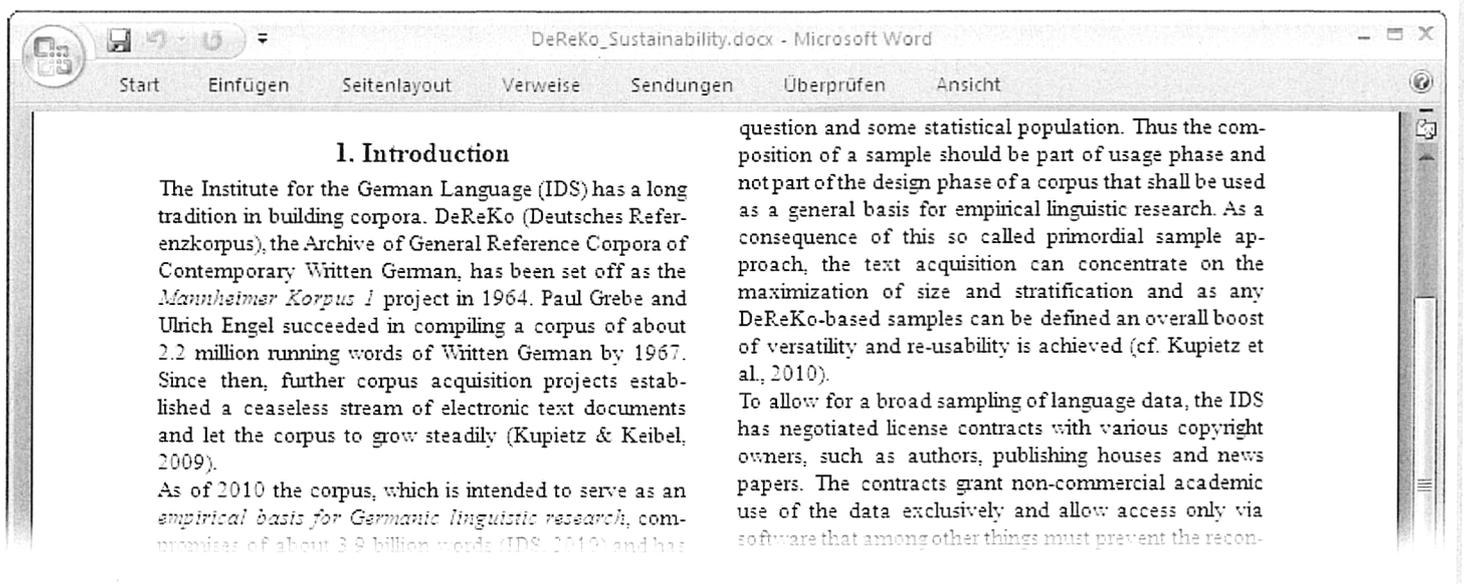
Our solution to overcome these problems is to process the standoff annotations and the annotated source text so that multiple annotations of the same text are created that are archived together with the original stand-off resources. This approach achieves sustainability through redundancy.

Problem: Machine-Generated XML

Today, a lot of XML data is generated by machines. Many of those XML documents are used for machine-to-machine communication, e.g., as SOAP-messages in web services. However, these messages are rather short-lived and will not be considered in this paper.

A growing number of applications use XML to store documents. These XML documents differ greatly from handcrafted XML and are rather complicated, especially with respect to the semantics of their tag sets, structure and code layout and therefore are difficult to comprehend by humans. Since users usually do not work with these documents directly this issue is not of a big concern. From a sustainability point of view these documents present a challenge though.

Figure 3: Screenshot of Microsoft Word 2007



As an example, the figure shows a conference paper created with Microsoft Word (see Figure 3). Since the 2007 version of Microsoft Office documents are saved by default in Office Open XML format (OOXML) (see ISO/IEC 29500:2008) and are – as the name suggests – encoded in XML. With regard to sustainability this is, in principle, a step in the right direction, but OOXML itself is not sufficient. ^[1] Without the corresponding application the generated XML document is very hard to understand or to use. Figure 4 shows an excerpt of the resulting OOXML document for the first heading and paragraph. The document is mostly structured by sections and paragraphs, but the OOXML structure does not show this structure in a transparent way. The following can be noted:

- There is no difference in markup used for headings and paragraph. Both are encoded by `w:p` elements. A heading made different from a paragraph by adding further information through the `w:pStyle` element. It's `w:val` attribute denotes whether the construct is a heading (“Heading1”) or a regular paragraph (“Textkorper”). More style information is encoded in additional XML files, but this still does not yield enough properties to resolving their role in structuring the text.
- The running text in the paragraph is heavily fragmented. For example the words “Referenzkorpus” or “established” are – for no apparent reason – both fragmented into 3 parts with a middle part which only contains a single character. The fragmentation could be the result of editing the document in MS Word's Track Changes mode.
- The markup contains rather complex constructs, e.g., the handling of italics. The words “Mannheimer Korpus 1” are set in italics. The formatting is applied to a text-run (`w:r` element) which has formatting information applied to it by means of a `w:rPr` element.

Figure 4: OOXML excerpt

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<w:document xmlns:ve="http://schemas.openxmlformats.org/markup-compatibility/2006"
  xmlns:o="urn:schemas-microsoft-com:office:office"
  xmlns:r="http://schemas.openxmlformats.org/officeDocument/2006/relationships"
  xmlns:m="http://schemas.openxmlformats.org/officeDocument/2006/math"
  xmlns:v="urn:schemas-microsoft-com:vml" xmlns:w10="urn:schemas-microsoft-com:office:word"
  xmlns:wp="http://schemas.openxmlformats.org/drawingml/2006/wordprocessingDrawing"
  xmlns:w="http://schemas.openxmlformats.org/wordprocessingml/2006/main"
  xmlns:wne="http://schemas.microsoft.com/office/word/2006/wordml">
<!-- ... -->
<w:p w:rsidR="00A77FB8" w:rsidRDefault="00A77FB8">
  <w:pPr>
    <w:pStyle w:val="Heading1"/>
    <w:numPr>
      <w:ilvl w:val="0"/>
      <w:numId w:val="5"/>
    </w:numPr>
  </w:pPr>
  <w:r>
    <w:lastRenderedPageBreak/>
    <w:t>Introduction</w:t>
  </w:r>
</w:p>
<w:p w:rsidR="00A77FB8" w:rsidRDefault="00A77FB8" w:rsidP="007357D1">
  <w:pPr>
    <w:pStyle w:val="Textkorper"/>
  </w:pPr>
  <w:r>
    <w:t>The Institute for the German Language (IDS) has a long tradition in building
    corpora. DeReKo (Deutsches Referenzkorpus), the Archive of General Reference Corpora of
    Contemporary Written German, has been set
  </w:r>
  <w:r>
    <w:t>r</w:t>
  </w:r>
  <w:r>
    <w:t xml:space="preserve">enzkorpus), the Archive of General Reference Corpora of
    Contemporary Written German, has been set
  </w:r>
  <w:r w:rsidR="003A1540">
    <w:t>off</w:t>
  </w:r>
  <w:r>
    <w:t xml:space="preserve"> as the
  </w:r>
  <w:r>
    <w:rPr>
      <w:i/>
    </w:rPr>
    <w:t>Mannheimer Korpus 1</w:t>
  </w:r>
  <w:r>
    <w:t xml:space="preserve"> project in 1964. Paul Grebe and Ulrich Engel succeeded in
    compiling a corpus of about 2.2 million running words of Written German by 1967. Since
    then, further corpus acquisition projects esta
  </w:r>
  <w:r>
    <w:t>b</w:t>
  </w:r>
</w:p>
```

```
<w:r>
  <w:t>lished a ceaseless stream of electronic text documents and let the corpus to grow
  steadily (Kupietz & Keibel, 2009).</w:t>
</w:r>
</w:p>
<!-- ... -->
</w:document>
```

An excerpt of an OOXML document produced by MS Word 2007 (the document was reformatted for readability).

Just having data encoded in XML does not automatically make the data sustainable. Especially very complex tag sets such as OOXML are of very limited use if one does not have an application which understands these formats. For almost any given application, obtaining and using such a piece of software will most probably pose a big problem a few years later. Sustainability of software is a whole different topic by itself and is not within the scope of this paper.

As a possible solution for this problem we propose to provide the machine-generated XML data in multiple formats. For example, the OOXML document can be stored in its native format, in plain text or in Portable Document format (PDF). Furthermore, filters can be used to remove those XML elements and attributes from the machine-generated code that are not necessary. It would be even better to transform the machine-generated XML data to established formats such as TEI (TEI P5).

Other than that, one should provide various descriptions and a thorough documentation of the data format, not only providing the schema but also tutorials, conceptual descriptions or similar documents for human reimplementations of tools operating on the machine-generated XML code.

Problem: Proprietary Tag Sets

In the document lifecycle, especially when taking long-term maintenance and archiving into account, it is a common problem that XML tag sets and document grammars are being used that are not well established outside the group defining the tagset. The use of XML tags following the insights and beliefs of the individual who wrote the schema as such does not pose the problem, but the interpretation of the schema by somebody reviewing the material later may cause problems, because no one else knows and understands the implicit logical constraints of tag and attribute names as well as data structures.

The usual answer to the use of proprietary tag sets would be not to use them at all and to replace them by standard annotation schemas and tagsets wherever possible. For example TEI (TEI P5), tagsets developed in the context of the standardization processes of ISO TC 37 SC 4 (“Language Resources”) or DocBook for technical articles and texts (see Walsh & Muellner 1999) come to mind. However, these tagsets do not always fit the given problem very well and using them often results in the well-known problem of tag abuse: tags are used in unintended ways or – even worse – users confuse the semantics of tags with their intended use. In these cases the results are bound to be more confusing than starting from an idiosyncratic tagset. Therefore, if users decide to use one of the established tagsets they should thoughtfully select the most appropriate one for a given problem.

More critical are those cases in which for various reasons no established tagset is used. Reasons for not selecting established tagsets range from not knowing about tagsets, not understanding tagsets, via policy reasons to the unavailability of appropriate tagsets. For example commercial terminological applications may use a data model that is consistent with established standards (such as ISO 16642:2003 in combination with ISO 30042:2008) but use a native XML format that is very similar but utilises different generic identifiers (for example SDL Trados MultiTerm 2009 shows this behaviour). The reason for this does not lie in the technology, but in management decisions. In each of these cases it is not sufficient to include the document grammar only to achieve valid XML, but further documentation is required. The basic idea is to document everything.

One way of approaching this problem is by providing a reference in the element description to an ontology or some other form of knowledge representation to define the data types with possible values. Data types here refer

both to XML elements and attributes, similar to data types used in XML schema. The reference to the external definition of the elements allows for a human user to evaluate the correctness of the semantic interpretation, possibly also to automatically evaluate the content using a parser. With external definitions the data types are unambiguously described according to available means.

The definition by reference is only one part of the definition, for human use it is advisable to use a documentation with the tag set that uses multiple examples. This *prototype semantics* of a tagset is intended to explain the meaning of tags and attributes as applied in a given domain or application. For human use it is also recommended to use names that bear a certain meaning, i.e., which are easily interpretable by a person reading them. Interpreting and understanding element and attribute names and values depends on a common background of the creator and user. For example, it is harder if both do not use the same script or language, because mutual intelligibility is important.

In the field of language resources this method has been implemented with ISOcat (ISO 12620:2009). ISOcat is a registry for data categories used in describing terminological databases and language resources. All data categories needed in these fields are allowed to be registered with a unique identifier, definition and name in various languages. Several data categories have been defined, but the list is open, hence it is possible to insert data categories that are needed but not available in the registry yet. The registry consists of two parts, a private and a public section. Every data category that is defined or used by a project or tagset is first defined in a private workspace that is nevertheless part of the registry and can be reused and referenced. Data categories that are important for various contexts can then be moved from private workspaces to the public area by domain experts. This promotion includes a quality assessment of the definitions as well as a check for possible redundancy in the registry. By this means consistency and documentation of data categories is fostered, together with persistent identifiers of the data categories, even in the case of the renaming of elements.

Based on the idea of persistent category definitions, the Component Metadata Infrastructure (CMDI, see Broeder et al. 2010) was designed. CMDI is intended for describing language resources. These resources are of various types and require different metadata schemas to appropriately describe the contents in a form that allows a human user to understand what kind of resources they have to expect. Most of these schemas are far more detailed than traditional metadata schemas from archivists containing bibliographical data, but also contain keywords, abstracts, subject fields, participants, annotation schemas, etc. For reusability reasons the data categories are clustered into *components*, and components are combined to other components or to a *profile*, which is more or less a metadata schema for a specific type of resource, the components also allowing the definition of a value schema for each data category. The data categories which are used in the components do not provide their own description, but refer to the data category registry, for example ISOcat or Dublin Core, using URIs. By this procedure, the concrete tag name becomes language, script and application independent, because the definition is given in a central repository. User interfaces are provided with the component registry web tool and the Arbil Metadata editor developed by the Max Planck Institute for Psycholinguistics (all available at the CMDI site).

Problem: Availability and Findability

Many researchers creating language resources are more than willing to share their resources with close colleagues upon request. However, for various reasons such as personal, privacy or property rights they tend to restrict public access to these resources. Furthermore, resources created in research contexts are usually designed for specific purposes such as the analysis of specific linguistic phenomena. The resource itself is mostly not visible, because research publications discuss the phenomena and their analysis, but usually do not describe the resource in great detail. However, these publications are often the only documentation for the existence of the resource and describe the rationale behind their creation. Hence, accessibility to language resources is a major problem to be dealt with.

Especially in fields with a large economic interest in linguistic resources, such as statistical language processing and machine learning, data centres or distribution agencies were created to address the problem of accessibility.

These data centres provide material in large quantities and they use rather flat structures for their data. In contrast, resources created by individual research projects and researchers are often deeply structured and tend to be much more detailed and complex. Data centres have standard procedures for intellectual property rights handling and cataloguing resources using bibliographical procedures. Language resources from commercially less interesting areas or resources that are deeply structured, can hardly be found in these data centres. Even if such resources are accessible elsewhere, they cannot be reliably located by general search engines. Most often they will only be part of the statistical noise of general search engine results. There are some specialized search engines, such as ODIN (see <http://www.csufresno.edu/odin>) for interlinear glossed text, but they usually do not provide users with knowledge about the text type and what kind of structures and content to expect in the resource.

The solution is well-known from the initial ideas around the semantic web: metadata descriptions of resources should be used that are based on standards, quasi-standards, best practice and which are used for specialized catalogues of resources. Providing exhaustive metadata records enables a possible user to understanding the structures and content of a resource, not necessarily the document grammar, but at least they would give a fair idea on the theory behind it.

Providing metadata refers to issues of proprietary tagsets and controlled vocabulary again. The keywords used to describe a resource ideally refer to a conceptual space, in which all concepts are well defined and classified according to superordinate and subordinate terms. The reference to the concept system or to an ontology requires standardized values. Standardized values means that a central, accessible structure needs to provide them, i.e., a kind of a registry such as ISOCat.

In the process of metadata creation different perspectives can be taken: the perspective of the author of the resource, the software engineer, the publisher and the person looking for a resource later, to name just a few. These different roles in relation to a resource are not mutually exclusive in terms of metadata categories, but in the creation process different areas are emphasized. For example the publisher will usually be more interested in making sure that the copyright is explicitly defined than the user searching for a specific resource to be employed for a specific use case. Software engineers will be interested in technical features, while archivist require bibliographical data.

For the creation of metadata it is essential to use the perspective of prospective users. Though it can be argued that not all possible users and their requirements can possibly be anticipated, the perspective of users, especially with other backgrounds, helps to include not only technically relevant metadata but also descriptive metadata relevant for human users. Technical metadata here means those bits of information required by someone implementing tools for processing the data, while descriptive metadata refers to those classifications that help a possible user to understand the content of a resource before actually seeing it.

Taking various user groups into account when selecting the descriptive detail of metadata also allows the design of structured search engines. Structured search engines here refer to search engines not only interpreting the textual content of pages but that take into account the structure of the metadata. The intention behind using the structure of metadata is to provide search results with a higher precision while providing a high recall at the same time, which is not necessarily achieved by full text based search engines.

Additionally it is recommended to make these datasets available through as many national and international catalogues and initiatives in the respective field or sub-field as possible (see section "Conclusions") and also to enable harvesting of metadata sets using OAI-PMH. With the help of these catalogues it is possible to announce the availability of the dataset to the scientific community using websites, blogs, etc.

Problem: Selection and Qualification for Long-Term Archiving

In the past, resources were either available or not, a lot of data was lost due to conversion problems, technical failure, etc. For each of these there are technical solutions, but a major problem remains in the question: what is

worth archiving? Resources undergo a life cycle and in general it is agreed that not every step in the life cycle is worth being archived. In contrast to that, some resources are supposed to be archived, even if they did not reach the archival phase of the intended life cycle. Finding formal criteria for deciding upon archiving or not is a major problem that still remains unsolved, one that might be unsolvable as such.

Criteria for deciding which resource should be archived fall into different categories: status, technical quality, organizational and institutional requirements, extent of use, quality evaluation and longevity. Some of these criteria depend on each other, but can be evaluated independently and therefore be used to measure the need for archiving a resource.

The status of a resource defines the formal editing status, starting from first draft versions to released or published versions, etc. Projects that work with a life cycle model in resource creation need to archive those documents that are in the archiving phase. Naming conventions and value schemas for the different phases vary greatly. However, the archive status cannot be the sole criterion, because in some projects resources get stuck in an earlier state and do not reach the publication phase, but considering other criteria, they nevertheless may qualify for or even require long-term archiving.

Especially for technological applications the technical quality can be of prime importance. For some testing environments it is sufficient to have a resource that is technically adequate and has the correct size, so it can serve as a reference point or for testing procedures, algorithms and technologies, even if the content and status as such are incomplete and still pending improvement. Consequently, the technical quality can be a decisive factor for long-term archiving.

Institutionalized requirements may force data providers to submit material, for example close to the end of a project life, while others are hesitant in providing data for various reasons, even if the quality is much higher. These requirements are usually negotiated with archivists and partners, but often result in archiving the resource regardless of other criteria.

A resource that is widely used by various groups needs to be archived regardless of other factors, because it is used as a reference. Ignoring other criteria such as quality and status. One reason could be that it is the only resource available or has unique properties. Though the use of a resource by a variety of users is complex to evaluate, this criterion seems to be obvious.

Quality is another factor in an evaluation matrix. In contrast to an approach which might be termed a take-whatever-you-can-get approach in archiving, archiving material without prior evaluation is not desired, as the information flood becomes unmanageable, if not for saving, then for retrieval and search. The assessment can be both formal by algorithmic processes that can also provide information on the technical quality mentioned before, or by a peer reviewing process. In the latter, experts decide on the quality of a resource and based on this judgment a resource is archived or disregarded.

Even more problematic but essential is the question of longevity of a resource. A resource that is most likely to be usable for a long period of time is supposed to be archived. The usability over a long period depends on the application of a resource. If the resource answers to demands that are continuously present, then the resource needs to be available, hence archived, even if the number of users might be small.

When measuring all of these criteria separately it is comparatively easy to define a threshold of criteria that need to be fulfilled in order for a resource to be archived. The threshold is selected in a way that each criteria can serve as an overriding criterion, that is, if one of these criteria mandates archiving, then the resource will be archived. But if there is no criteria with this requirement, the values can accumulate. If the threshold is not set too low, the resource will then be archived.

The ultimate goal for working with resources is of course to achieve a high quality resource, that is highly regarded by experts, used and usable for many years, and reaches a maturity level that is technically well

established, etc. However, for most resources there are limitations that are not supposed to interfere as knock-out criteria for long-term archiving.

Additional Pitfalls

Technical sustainability is one aspect of sustainability. Other major aspects are organizational sustainability and legal issues – two issues not to be underestimated. While the technical sustainability is an engineering task which seems to be solved in most cases with semi-automatic migration procedures for digital devices, this is not true for organizational and legal aspects.

Eide et al. 2008 claim that organizational sustainability may even be more important than technical sustainability, because valuable resources can easily be lost when an organization is shut down. They list several examples from cultural heritage management, where shutting down museums almost lead to the loss of resources, e.g., the Newham case where data was only saved because the staff acted quickly and dumped it to floppy disks. Sometimes, the resources also exist on paper and could be digitized again, but as there is a movement away from paper, this option will cease to exist soon.

Organizational sustainability is a rather fragile process because it correlates with funding and institutional commitment, which are rather soft and fragile factors. Due to the structure of funding organizations it is hardly possible to receive a statement of commitment for a very long period of time. For example, the duration of German collaborative research centres is limited to 12 years. Other long time programs exist, but it is virtually impossible to find a commitment for more than 20 years. Therefore, ventures in sustainability also need to consider the organizational aspect with a proper strategy how to guarantee taking care of resources in the years to come – either by securing continuity of the organization itself or by preparing and implementing a proper migration plan for resources to a different organization. Preparing for both cases would be even better.

Another issue are legal aspects. Especially in the field of linguistics, intellectual property rights create their own set of problems which have to be dealt with when thinking about sustainability (see Lehmborg et al. 2008 and Zimmermann et al. 2007). These issues are investigated in the context of international projects such as CLARIN and META-NET; the current direction is to work out licensing models (see Lindén et al. 2010 and Weitzmann et al. 2010). These intellectual property rights issues are especially tricky as linguistic resources often cross political and cultural borders, hence not only legal issues but also ethical implications are involved.

Conclusions

Sustainability of language resources is an aspect wanted and needed by data providers, users and funders alike. To be able to speak of sustainable resources it is necessary to make resources available according to defined processes, platforms or archives in a reproducible and reliable way. To this end, XML is an essential part of a complex approach which, additionally, also encompasses other standards on multiple levels. These are requirements, but tools and systems, accessible in a reliable manner and operating based on standards, are important as well.

With SPLICR there is a proof-of-concept implementation of large parts of the functionality required for sustainability platforms. A platform alone is a node in the sustainable web of trusted resource repositories, each repository providing organizational support, technical infrastructure with archiving technology, and being entrusted to use specified procedures to respect privacy and rights of data providers while providing non-discriminatory access to the resources according to stated procedures and rights holders restrictions. Part of this network is also the cooperation of various national and international initiatives. In cases of sustainability a certain amount of overlap between these projects is desirable to further foster interoperability and reliability of tools, data centres and increase redundant archives, avoiding major problems in disaster scenarios.

All in all it can be said that with a number of international projects such as CLARIN and META-NET along with

its META-SHARE open resource exchange facility, together with the initial implementations of various tools, the development of standards in the ISO Technical Committee 37, Subcommittee 4 “Language Resources” (see TC 37 SC 4) and establishment of de-facto procedures, the sustainability of language resources is no longer something that needs to be argued for. Instead, the situation has changed dramatically, as the very real problem of providing sustainable data sets is, by now, firmly anchored in academic as well as commercially oriented research centres. With raised awareness in the community, the continuation of language resource distribution projects and institutional support by academic libraries and institutions, chances are more than promising for providing sustainable resources, using XML technology and state of the art processes.

References

- [Bański & Przepiórkowski 2009] Bański, P. and Przepiórkowski, A. “Stand-off TEI annotation: the case of the National Corpus of Polish”. In: *Proceedings of the Third Linguistic Annotation Workshop (LAW III) at ACL-IJCNLP 2009*, Singapore, 2009, pages 64–67.
- [Bański 2010] Bański, P. “Why TEI stand-off annotation doesn't quite work and why you might want to use it nevertheless”. In: *Proceedings of Balisage 2010. Series on Markup Technologies*, vol. 6, 2010. doi:10.4242/BalisageVol15.Banski01.
- [Broeder et al. 2010] Broeder, D., Kemps-Snijders, M., Van Uytvanck, D., Windhouwer, M., Withers, P., Wittenburg, P. and Zinn, C. “A Data Category Registry- and Component-based Metadata Framework”. In: *Proceedings of LREC 2010*, Malta, 2010, pp. 43–47
- [Carletta et al. 2003] Carletta, J., Kilgour, J., O'Donnell, T., Evert, S., Voormann, H. “The NITE Object Model Library for Handling Structured Linguistic Annotation on Multimodal Data Sets”. In: *Proceedings of the EACL Workshop on Language Technology and the Semantic Web (3rd Workshop on NLP and XML)*.
- [Eide et al. 2008] Eide, Ø., Ore, C.-E. and Holmen, J. “Sustainability in Cultural Heritage Management”. In: *Proceedings of Digital Humanities 2008*, Oulu, Finland, pp. 22–23.
- [Ide & Romary 2007] Ide, N. and Romary, L. “Towards International Standards for Language Resources”. In: Dybkjær, L., Hensen, H., Minker, W. (eds.), *Evaluation of Text and Speech Systems*, Springer, pages 263–284.
- [ISO 12620:2009] ISO 12620:2009. “Terminology and other language and content resources – Specification of data categories and management of a Data Category Registry for language resources”.
- [ISO 16642:2003] ISO 16642:2003. “Computer applications in terminology – Terminological markup framework”.
- [ISO 30042:2008] ISO 30042:2008. “Systems to manage terminology, knowledge and content – TermBase eXchange (TBX)”.
- [ISO/IEC 29500:2008] ISO/IEC 29500:2008. “Information technology – Office Open XML formats”.
- [Lehmberg et al. 2008] Lehmberg, T., Rehm, G., Witt, A. and Zimmermann, F. “Digital Text Collections, Linguistic Research Data, and Mashups: Notes on the Legal Situation”. In: *Library Trends*, 57/1, pp. 52–71. doi:10.1353/lib.0.0023.
- [Lindén et al. 2010] Lindén, K., Oksanen, V. and Bruun, S. (eds) “CLARIN Classification Guide for Deposition Licenses – First comprehensive summary about licensing problems”. To appear.
- [Rehm et al. 2009] Rehm, G., Schonefeld, O., Witt, A., Hinrichs, E. and Reis, M. Sustainability of annotated resources in linguistics: “A web-platform for exploring, querying, and distributing linguistic corpora and other resources”. In: *Literary & Linguistic Computing (LLC) – Journal of the Association for Literary and Linguistic Computing*, 24 (2009) 2, pp. 193–210. doi:10.1093/llc/fqp003.
- [TEI P5] The TEI Consortium (ed.) “Guidelines for Electronic Text Encoding and Interchange (TEI P5)”. The TEI Consortium, 2007. <http://www.tei-c.org/Guidelines/P5/>.
- [Thompson & McKelvie 1997] Thompson, H. and McKelvie, D. “Hyperlink semantics for standoff markup of read-only documents”. In: *Proceedings of SGML Europe*, 1987. <http://www.ltg.ed.ac.uk/~ht/sgmleu97.html>.
- [Trippel et al. 2007] Trippel, T., Declerck, T. and Ide, N. “Interoperable Language Resources”. In: *SDV Sprache und Datenverarbeitung/International Journal for Language Data Processing*, Volume 31.1–2, pp. 101–113

- [Walsh & Muellner 1999] Walsh, N. and Muellner, L. “DocBook: The Definitive Guide”, Sebastopol, O'Reilly Media, 1999.
- [Weitzmann et al. 2010] Weitzmann, J. H., Rehm, G and Uszkoreit, H. “Licensing and Sharing Language Resources: An Approach Inspired by Creative Commons and Open Science Data Movements”. In: *LREC 2010 Workshop Legal Issues for Sharing Language Resources: Constraints and Best Practices*, May 17, Malta, 2010.
- [Witt 2004] Witt, A. “Multiple Hierarchies: New Aspects of an Old Solution.”. In: *Proceedings of Extreme Markup Languages 2004*, Montréal, Canada, 2004
- [Witt et al. 2007] Witt, A., Schonefeld, O., Rehm, G., Khoo, J. and Evang, K. “On the Lossless Transformation of Single-File, Multi-Layer Annotations into Multi-Rooted Trees”. In: *Proceedings of Extreme Markup Languages 2007*, Montréal, Canada, 2007.
- [Zimmermann et al. 2007] Zimmermann, F. and Lehmborg, T. “Language Corpora – Copyright – Data Protection: The Legal Point of View”. In: *Proceedings of Digital Humanities 2008*, Urbana-Champaign, United States, pp. 162–164.

[¹] In this aspect we are not arguing in favour or against the OOXML standard. Whether OOXML is a good or bad standard or whether it is well designed or lies not in the scope of this paper and is to be discussed elsewhere. We were only interested in the generated XML code and inspected it against the background of sustainability. Similar results hold for OpenDocument format (ODF) documents generated by OpenOffice.

Georg Rehm

<georg.rehm@dfki.de>

DFKI

Georg Rehm works at DFKI, the German Research Center for Artificial Intelligence, where he coordinates META-NET, a strategic pan-European research project on Machine Translation and multilingualism. He holds a PhD in Computational Linguistics and has been working with SGML and related technologies in the context of Natural Language Processing since 1995.

Oliver Schonefeld

<schonefeld@ids-mannheim.de>

Institute for the German Language (IDS), Mannheim

Oliver Schonefeld works at the Institut für Deutsche Sprache (Institute for the German Language) in Mannheim and is involved in the projects TextGrid and Clarin.

He studied computer science with specialization in text technology at Bielefeld University until 2005. After graduating he worked as a researcher at Bielefeld University and later at Tübingen University's collaborative research center Linguistic Data Structures.

His major research interests are the limitations of markup languages (especially overlapping markup) and the use of markup languages in linguistic description of language data.

Thorsten Trippel

<thorsten.trippel@uni-tuebingen.de>

Tübingen University

Thorsten Trippel works at Tübingen University in a project on sustainability of language resources called NaLiDa. This national project aims at providing a platform for linguists to locate resources they need and to enable them to produce long time usable data by introducing them to relevant metadata descriptions and standards. He is part of national and international standardization groups on language resources.

His major research interests are directed towards language resources in general and specifically in terminology and lexicography/lexicon theory (PhD Thesis: The Lexicon Graph Model: A generic Model for multimodal lexicon development) including other types of resources such as speech corpora and involving other modalities. He has conducted research in speech technology and textual corpus linguistics, has been working with (XML-)databases for information retrieval over highly structured data and run research projects on interface design for

such data.

Work at his previous affiliation Bielefeld University involved research projects in Brazil, transforming archives of handwritten texts into web-usable multi purpose sources for computational linguists and historians. Additionally he taught at Bielefeld University, and various institutions and summer schools, for example in introducing text technological and computational linguistic backgrounds to field linguists and language documentarists in West-Africa.

Andreas Witt

`<witt@ids-mannheim.de>`

Institute for the German Language (IDS), Mannheim

Witt received his Ph.D. in Computational Linguistics and Text Technology from the Bielefeld University in 2002 (dissertation title: “Multiple Informationsstrukturierung mit Auszeichnungssprachen. XML-basierte Methoden und deren Nutzen für die Sprachtechnologie”).

After graduating in 1996, he started as a researcher and instructor in Computational Linguistics and Text Technology. He was heavily involved in the establishment of the minor subject Text Technology in Bielefeld University's Magister and B.A. program in 1999 and 2002 respectively. After his Ph.D. in 2002 he became an assistant lecturer, still at the Text Technology group in Bielefeld. In 2006 he moved to Tübingen University, where he was involved in a project on “Sustainability of Linguistic Resources” and in projects on the interoperability of language data. Since 2009 he is senior researcher at Institut für Deutsche Sprache (Institute for the German Language) in Mannheim.

Witt is and was a member of several research organizations, amongst them the TEI Special Interest Group on overlapping markup, for which he was involved in the writing of the latest version of the chapter “Multiple Hierarchies”, which is included in TEI-Guidelines P5.

Witt's major research interests deal with questions on the use and limitations of markup languages for the linguistic description of language data.

Balisage Series on Markup Technologies