Balisage: The Markup Conference 2012

Proceedings

A standards-related web-based information system

Maik Stührenberg
Institut für Deutsche Sprache (IDS) Mannheim
Universität Bielefeld
`<maik.stuehrenberg@uni-bielefeld.de>`

Oliver Schonefeld
Institut für Deutsche Sprache (IDS) Mannheim
`<schonefeld@ids-mannheim.de>`

Andreas Witt
Institut für Deutsche Sprache (IDS) Mannheim
`<witt@ids-mannheim.de>`

Balisage: The Markup Conference 2012
August 7 - 10, 2012

How to cite this paper

Stührenberg, Maik, Oliver Schonefeld and Andreas Witt. "A standards-related web-based information system." Presented at Balisage: The Markup Conference 2012, Montréal, Canada, August 7 - 10, 2012. In *Proceedings of Balisage: The Markup Conference 2012*. Balisage Series on Markup Technologies, vol. 8 (2012). DOI: `10.4242/BalisageVol8.Stuhrenberg01`.

Abstract

This late breaking proposal introduces the prototype of a web-based information system dealing with standards in the field of annotation.

Table of Contents

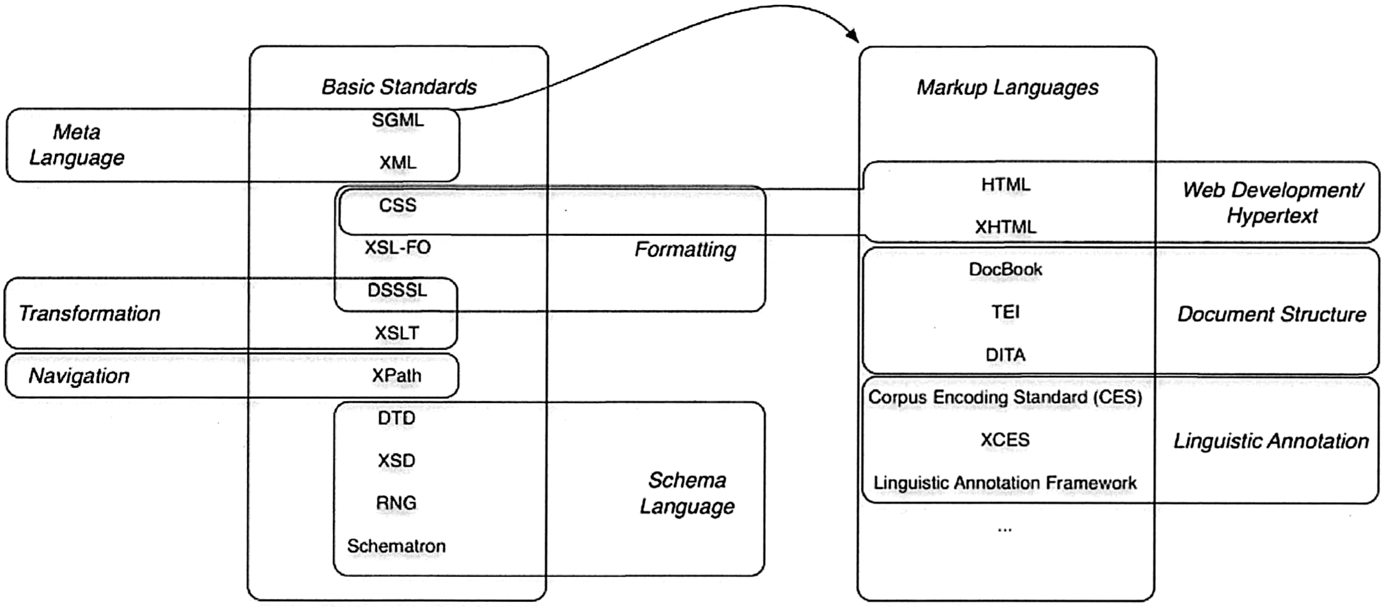
Introduction

This late breaking proposal is based on an ongoing effort started in the CLARIN project and which was presented briefly at the LREC 2012 Workshop on Collaborative Resource Development and Delivery. Initial point was the development of an easy to use information system for the description of standards developed in ISO/IEC TC37/SC4 *Language Resources Management*. Since these standards are heavily related to each other it is usually not feasible to adopt only a single standard for one's work but to dive into the standards jungle in full. Because of positive feedback after the presentation of the prototype at the Istanbul Workshop we decided to broaden the information accessible by the system to markup language related standards.

The problem with standards

Every year at Balisage's markup game or quiz difficult questions regarding markup have to be answered. Even the participants of this very special conference have sometimes problems to find the correct answer. An some of the younger generation have never heard terms like 'sosofo' or 'DSSSL' before. At current, an unmanageable number of standards is available for annotating data of various kinds. These standards can be divided into several groups, according to such different features such as standard body (W3C, ISO, OASIS, HL7, to name just a few), basic or derivative work (i. e., standards that are built upon basic specifications), the state (de jure vs. de facto standard), or the topic. Figure 1 gives a very rough overview about some of the named aspects. Other divisions would deal with a temporal aspect of standardization: specifications change over the time of development, some (almost) historical standards have been abandoned and have been replaced by other specifications.

Figure 1: A rough overview about the complexity in the fields of standardization.



Missing in the aforementioned list is another interesting aspect: the relationships between different standards. For example, meta languages such as SGML or XML are used to define (that is, syntax and possibly a schema formalism) markup languages. A given markup language is defined by a schema which in turn is defined by using a schema language (either a grammar-based or rule-based constraint language), and so on. Apart from these relations between basic standards and those that are built upon these, are other relations between members of the first group and of the second group. Some of the formal restrictions of XML instances and XML DTDs (and even XSD) are based on faits accomplis created during the development of SGML, DSSSL and HyTime &#8211; standards that are decades old. Features that were already present in DSSSL, have been improved and adapted for current W3C standards such as XPath, XSLT and XSL-FO. XPath is usually used via a host language such as XSLT. The Corpus Encoding Standard (CES) is an application of the SGML-based version of the TEI, P3. The are dozens of similar relations, some are only good for get bonus points at Balisage Bingo but others really help in understanding specific issues one may have when working with a given specification. At least the knowledge about these older standards and the decisions that were made during their design process and which still impact current technologies such as XML schema or XSLT runs the risk of being forgotten just because of the amount of time already passed.

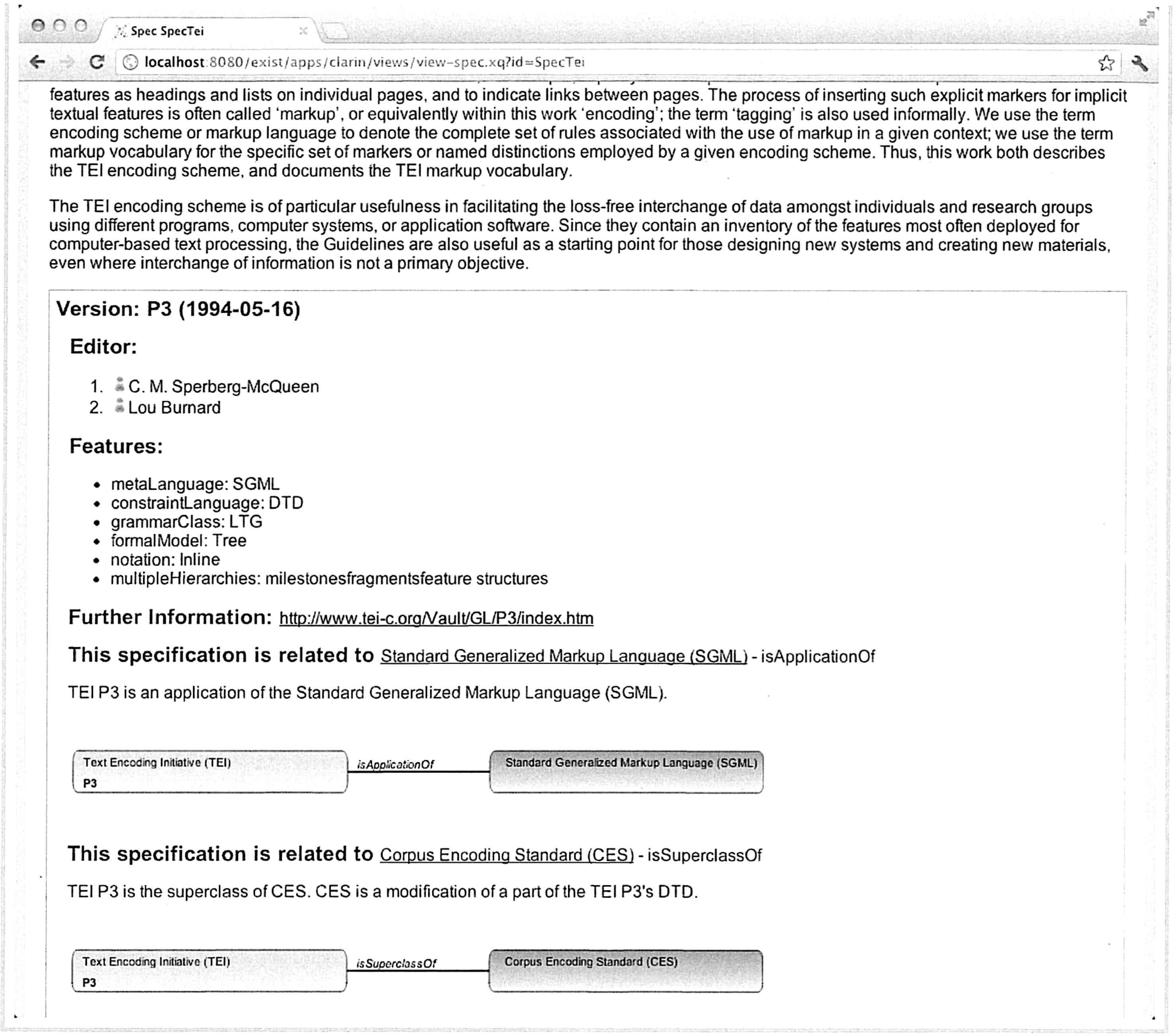
This holds especially when dealing with non-basic standards, for example with markup languages for annotating linguistic data. During the last couple of years a large number of specifications have been developed in ISO/IEC TC 37/SC 4 Language Resources Management. These standards are often released to the public during their various stages of the standardization process (either as publicly available specialisations, such as Draft International Standard or as topic of a research paper). However, these version may substantially differ from the final version of the standard which is usually not available without charge. In addition, the relations between these specifications are numerous which complicates their correct use.

Another problem is the fact, that often scholars and researchers are not even aware that standardized formats and models exist.

Providing guidance

We propose a community project to build up a platform providing guidance through the jungle that has been grown around the XML world. Starting with a very small set of standards and specifications and constructed as an XRX (XForms, Rest, XQuery) application we offer the starting point for a platform that allows Balisage's experts to share their knowledge with others. During the last months we have developed a prototypical web-based information system that serves as a starting point in providing guidance through the standards jungle as part of the *CLARIN* distributed project group.[1] Up to now, it contains a collection of topics such as *Meta Language*, *Metadata*, *Generic Corpus Annotation*, or *Constraint Language*, amongst others, standard bodies such as *ISO*, *W3C*, *OASIS*, and *HL7*, and 25 specifications at the time of writing. Figure 2 shows a partial screenshot of the current state. Relations between specifications (in this case between TEI P3 and SGML and P3 and CES) are described both in a textual and graphical way.

Figure 2: A screenshot of the current state of the information system (showing TEI's entry).



Spec SpecTei
localhost:8080/exist/apps/clarin/views/view-spec.xq?id=SpecTei

features as headings and lists on individual pages, and to indicate links between pages. The process of inserting such explicit markers for implicit textual features is often called 'markup', or equivalently within this work 'encoding'; the term 'tagging' is also used informally. We use the term encoding scheme or markup language to denote the complete set of rules associated with the use of markup in a given context; we use the term markup vocabulary for the specific set of markers or named distinctions employed by a given encoding scheme. Thus, this work both describes the TEI encoding scheme, and documents the TEI markup vocabulary.

The TEI encoding scheme is of particular usefulness in facilitating the loss-free interchange of data amongst individuals and research groups using different programs, computer systems, or application software. Since they contain an inventory of the features most often deployed for computer-based text processing, the Guidelines are also useful as a starting point for those designing new systems and creating new materials, even where interchange of information is not a primary objective.

Version: P3 (1994-05-16)

Editor:

1. C. M. Sperberg-McQueen
2. Lou Burnard

Features:

- metaLanguage: SGML
- constraintLanguage: DTD
- grammarClass: LTG
- formalModel: Tree
- notation: Inline
- multipleHierarchies: milestonesfragmentsfeature structures

Further Information: http://www.tei-c.org/Vault/GL/P3/index.htm

This specification is related to Standard Generalized Markup Language (SGML) - isApplicationOf

TEI P3 is an application of the Standard Generalized Markup Language (SGML).

Text Encoding Initiative (TEI)
P3
isApplicationOf
Standard Generalized Markup Language (SGML)

This specification is related to Corpus Encoding Standard (CES) - isSuperclassOf

TEI P3 is the superclass of CES. CES is a modification of a part of the TEI P3's DTD.

Text Encoding Initiative (TEI)
P3
isSuperclassOf
Corpus Encoding Standard (CES)

Information structure

The information system itself is based on standards as well. We have developed a lightweight format for structuring information about specifications, topics, and standard setting bodies defined by an XSD. Although it seems to be at least questionable to invent yet another annotation format especially for this kind of project we tried to stick as close to existing annotation formats such as the TEI as possible while streamlining the format and therefore keeping it small and simple. Figure 3 shows an excerpt of the storage format.

Figure 3: An excerpt of the storage format

```
<spec xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xml:id="SpecXML"
  standardSettingBody="W3C" topicRef="TopicMetaLanguage"
  xsi:noNamespaceSchemaLocation="http://localhost:8080/exist/apps/clarin/xsd/spec.xsd">
  <titleStmt>
    <title>Extensible Markup Language (XML)</title>
    <abbr>XML</abbr>
  </titleStmt>
  <scope>Meta language for Creating Markup Languages</scope>
  <description>
    <p>The Extensible Markup Language (XML) is a meta language specified by a W3C recommendation,
      containing both a syntax for describing and serializing data objects called XML documents (or
      XML instances) and a formalism for describing document grammars (DTD)
   <!-- [...] -->
   </p>
  </description>
  <version xml:id="SpecXML1">
    <versionNumber type="major">1.0</versionNumber>
    <date>1998-02-10</date>
    <respStmt>
```

```
      <resp>Editor</resp>
      <name type="person">Tim Bray</name>
      <name type="person">Jean Paoli</name>
      <name type="person">C. M. Sperberg-McQueen</name>
    </respStmt>
    <address type="URL">http://www.w3.org/TR/1998/REC-xml-19980210</address>
    <relation target="SpecSGML" type="isSubclassOf">
      <p>XML 1.0 is a subset of SGML.</p>
    </relation>
  </version>
  <version xml:id="SpecXML1-SE">
    <versionNumber type="major">1.0</versionNumber>
    <versionNumber type="minor">Second Edition</versionNumber>
    <date>2000-10-06</date>
    <!-- [...] -->
  </version>
  <version xml:id="SpecXML1.1">
    <versionNumber type="major">1.1</versionNumber>
    <date>2004-02-04</date>
    <description>
      <p>XML 1.1 differs from XML 1.0 in terms of naming conventions for elements and attributes
        (generic identifier) with respect to current and future version of Unicode. Whereas XML
        1.0 provided a rigid definition of names, wherein everything that was not permitted was
        forbidden, XML 1.1 names are designed so that everything that is not forbidden (for a
        specific reason) is permitted. </p>
    </description>
    <respStmt>
      <resp>Editor</resp>
      <name type="person">Tim Bray</name>
      <name type="person">Jean Paoli</name>
      <name type="person">C. M. Sperberg-McQueen</name>
      <name type="person">Eve Maler</name>
      <name type="person">François Yergeau</name>
      <name type="person">John Cowan</name>
    </respStmt>
    <address type="URL">http://www.w3.org/TR/2004/REC-xml11-20040204/</address>
    <relation target="SpecXML1" type="isVersionOf">
      <p>XML 1.1 is a refined version of XML 1.0.</p>
    </relation>
  </version>
  <version xml:id="SpecXML1.1-SE">
    <!-- [...] -->
  </version>
</spec>
```

The information is stored in a native XML database system (we have chosen the Open Source eXist database as starting point but try to do not use any application-dependant features). Queries on the data are performed via XQuery scripts and forms will be implemented by XForms (which in turn will be processed by XSLTForms supported by eXist). The goal of the information system is not to replicate information that is already available (but may not be traceable anymore), but to connect pieces of information and enrich these pieces with small amounts of additional data. In the example above (Figure 3) we only refer to the information available at the W3C (and respective places). Information about standard bodies and topics is stored in a similar way.

Representation

The output of the information system is based on HTML5's XML syntax. Although the textual representation is feasible for providing detailed information about a given standard, our plan is to establish an additional graphical representation format. This second representation should provide a better overview of the relations between a large number of standards at once. Minor demonstrations have been done with D3 Bostock et al., 2011, an Open Source JavaScript library that directly manipulates objects in the DOM tree to interactively visualize data[2]. However, since JavaScript and XQuery share some syntax elements (e.g. the curly brackets) there are some minor traps to avoid. In the current prototype we therefore only use the non-interactive SVG graphics that are included in the standard's description page (see Figure 2). Since browser support for interactive 2- and 3-dimensional graphics is getting stronger and stronger, other options such as WebGL-based representations are possible as well (see Jettka and Stührenberg, 2011 for example).

Current state and future work

The current state of the information system is still quite rough. What we want to propose is allow Balisage's participants to get involved in the process of collecting and sharing information about the standards they work with and have knowledge of. In addition, Balisage is the place to find experts in SVG (which is used to display relations between two or more specialisations), XQuery and XForms. If we manage to bring these people together to establish a community that is willing to share its knowledge the final product could be of much use to scholars and researchers around the world. As a first starting point we will publish the XML schema defining the annotation format to receive comments and add further enhancements. After a stable format has been established, interesting parties could create specification sheets and upload them into the platform. In addition, we will open the platform for reading access for other people to give feedback on a less technical way. The current prototype is made available at http://clarin.ids-mannheim.de/standards/index.xq.

References

[Bostock et al., 2011] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D3: Data-driven documents. IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis), 2011.

[Jettka and Stührenberg, 2011] Daniel Jettka, and Maik Stührenberg. Visualization of concurrent markup: From trees to graphs, from 2d to 3d. In Proceedings of Balisage: The Markup Conference, volume 7 of Balisage Series on Markup Technologies, Montreal, 2011. doi:10.4242/BalisageVol7.Jettka01.

[Stührenberg et al., 2012] Maik Stührenberg, Antonina Werthmann, and Andreas Witt. Guidance through the standards jungle for linguistic resources. In Nancy M. Ide, Collin Baker, Christiane Fellbaum, and Rebecca Passonneau, editors, Proceedings of the LREC 2012 Workshop on Collaborative Resource Development and Delivery, pages 9–13, 2012.

[1] The system was presented in May 2012 at the LREC 2012 Workshop on Collaborative Resource Development and Delivery and is described in more detail in Stührenberg et al., 2012.

[2] Additional information and downloads are available at http://d3js.org/.

Maik Stührenberg
<maik.stuehrenberg@uni-bielefeld.de>
Institut für Deutsche Sprache (IDS) Mannheim
Universität Bielefeld

Maik Stührenberg received his Ph.D. in Computational Linguistics and Text Technology from Bielefeld University in 2012. After graduating in 2001 he worked four years as research assistant at Giessen University in different text-technological projects together with Henning Lobin and Georg Rehm. Afterwards, he worked together with Andreas Witt, Dieter Metzing, Daniela Goecke and Daniel Jettka in the Sekimo project of the Research Group 437 Text-technological Modelling of Information funded by the German Research Foundation. During 2011 and 2012 he was employed at the Institut für Deutsche Sprache (IDS, Institute for the German Language) in Mannheim as member of the CLARIN-D project group and is currently employed as research assistant at Bielefeld University.

His main research interests include specifications for structuring multiple annotated data, query languages, and query processing.

Oliver Schonefeld
<schonefeld@ids-mannheim.de>
Institut für Deutsche Sprache (IDS) Mannheim

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

Andreas Witt
<witt@ids-mannheim.de>

Institut für Deutsche Sprache (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.

Balisage Series on Markup Technologies