Journal of the Text Encoding Initiative

Issue 3 | November 2012
TEI and Linguistics

A TEI P5 Document Grammar for the IDS Text Model

Harald Lüngen and C. M. Sperberg-McQueen

Electronic version
URL: http://journals.openedition.org/jtei/508
DOI: 10.4000/jtei.508
ISSN: 2162-5603

Publisher
TEI Consortium

Electronic reference
Harald Lüngen and C. M. Sperberg-McQueen, « A TEI P5 Document Grammar for the IDS Text Model », *Journal of the Text Encoding Initiative* [Online], Issue 3 | November 2012, Online since 15 October 2012, connection on 21 April 2019. URL : http://journals.openedition.org/jtei/508 ; DOI : 10.4000/jtei.508

A TEI P5 Document Grammar for the IDS Text Model

Harald Lüngen and C. M. Sperberg-McQueen

1. Introduction

1 The Institut für Deutsche Sprache (IDS) in Mannheim, Germany, hosts the German Reference Corpus (DEREKO), the largest archive in the world of corpora of contemporary written German. With over 5 billion word tokens,[1] DEREKO contains fiction, scientific texts, newspaper articles, and a wide variety of other text types. The corpora in DEREKO have been collected since 1964 and are licensed for academic use via the IDS corpus access platform COSMAS II.[2] They are used by linguistics researchers at the IDS and at other institutions around the world.

2 All corpora within DEREKO are marked up with metadata and annotations according to the IDS text model, which is currently realized in IDS-XCES, an IDS-specific adaptation of the XCES corpus encoding standard (Ide et al. 2000). This paper describes the features of the IDS text model and our ongoing project, named I5 (short for "*IDS-TEI P5*"), in which we are preparing a TEI P5 ODD document for this text model. Since DEREKO is not available for direct download, a migration to TEI P5 had not been highly prioritized (since no one outside the IDS would directly benefit from such a conversion). However, it is hoped that a TEI P5 document grammar for the IDS text model will facilitate the building and maintenance of quality assurance tools, will enable the IDS to abandon the older in-house annotation format and therefore enable new project members to familiarize themselves more quickly and easily with the model. In the long run, we hope that the migration to TEI P5 will contribute to a harmonization and standardization process in which tools will be produced that are able to deal with large-scale TEI data (cf. Kupietz et al. 2010).

3 This paper begins with background on the nature and purposes of the corpora collected at IDS and the motivation for the I5 project (section 1). It continues with a description of the origin and history of the IDS text model (section 2), and a description (section 3) of

the techniques used to automate, as far as possible, the preparation of the ODD file documenting the IDS text model. It ends with some concluding remarks (section 4). A survey of the additional features of the IDS-XCES realization of the IDS text model is given in an appendix.[3]

2. Origin and History of the IDS Text Model

4 IDS researchers at different locations in Germany initially created corpora for specific research purposes, each encoded using a home-grown encoding scheme. Examples of these early corpora are the *Wendekorpus*,[4] the *Bonner Zeitungskorpus*,[5] and the *Mannheimer Korpora 1* and *2*,[6] all of which are available to this day as part of the German Reference Corpus, DEREKO. A unified text and annotation format for all IDS corpus texts was first introduced in 1991.

2.1. BOT

5 The first attempt at a unified pivot format for IDS corpora was called *BOT* (an acronym formed out of the initials of "Beginning of text"). It grew out of the COSMAS (*Corpus Search, Management and Analysis System*, later known as COSMAS I) project, which lasted from 1991 to 2003. Among its goals were the integration of the various existing single corpora into a common representation format, the centralization of corpus acquisition and encoding activities at the IDS, and the development of corpus access software for linguistic research. The first version of BOT was defined by Cyril Belica of IDS in 1992 and remains the basis of the IDS text model. It is a character-based format, of which the header part contains bibliographic metadata expressed in seven data fields (see table 1 and the example below it), which form the minimum of bibliographic data. Each field is represented in a single line in the file and exhibits a binary structure (*field_name: value_string*).

Field name	**Semantics**
BOTC	*corpus identifier*
BOTD	*document identifier*
BOTT	*text identifier*
BOTd	*resolved document identifier*
BOTt	*elaborated bibliographic reference*
BOTi	*reduced bibliographic reference*
BOTP	*processing information: is page numbering encoded in the corpus text as in the source or not?*

Table 1: The seven fields of the BOT minimum

```
BOTC:DIV
BOTD:WC4
BOTT:WC4.04004
BOTd:Christa Wolf: Essays/Gespräche/Reden/Briefe
 1959-1974
BOTt:DIV/WC4.04004 Wolf, Christa: Das siebte Kreuz,
 [Nachwort], (Entstehung: 1963), In: Wolf, Christa:
 Werke, Bd. 4, Essays/Gespräche/Reden/Briefe 1959-
 1974, Hrsg.: Hilzinger, Sonja. - München:
 Luchterhand Literaturverlag, 1999, S. 24-41
BOTi:DIV/WC4.04004 Wolf, Christa: Das siebte Kreuz,
 [Nachwort], (Entstehung: 1963), In: Wolf, Christa:
 Werke, Bd. 4, Essays/Gespräche/Reden/Briefe 1959-
 1974, Hrsg.: Hilzinger, Sonja. - München:
 Luchterhand Literaturverlag, 1999, S. 24-41
BOTP:1
```

Example of corpus text header as BOT minimum. The line breaks within the fields are not present in the original.

6 The fields BOTC, BOTD, and BOTT reflect a three-level hierarchical structure in this model: *corpus*, *document*, and *text*. A corpus contains one or more documents, and a document contains one or more texts, and each corpus, document, or text would have such a header. In the model, a text is defined as *a relatively independent, coherent sequence of natural language utterances that has emerged from natural communicative situations.*[7] A text may comprise, for example, one or sometimes several newspaper articles, a journal article, a short story, or an extract of a literary work. Texts are combined to form a document according to certain aspects such as source, chronological sequence, topic, or text type—for example, texts from one edition of a particular day's newspaper would form one document. However, not every document contains more than one text: a corpus of the collected works of one author would contain one document per novel, each of which would include a single text.

7 The fields BOTd, BOTt, and BOTi contain the bibliographic reference in different degrees of detail,[8] each of which was needed for different presentation modes (for example, as part of a corpus overview or of a KWIC view of query results).

8 Later, the field BOT+ent (for "Entstehungszeit", the time of creation, if known, or otherwise of the first edition) was also included in the BOT minimum because the (approximate) year when a literary work was actually written can differ considerably from the year of publication of the source used in the composition of the corpus. The collected works of Thomas Mann, for instance, all have 1960 as their year of publication, while his first novel, *Buddenbrooks*, first appeared in 1901. If only the date of publication were recorded, discrepancies between the date of composition and the date of publication would distort linguistic analyses of language variation over time.

9 BOT also included a number of "surrounding tags" for inline annotations such as b+…+b for a caption or u+…+u for a heading, using what is sometimes known as the "Mannheim Conventions" ("Mannheimer Konvention", MK). This convention was based on markup as used in several of the earlier corpora (see, for example, Kolvenbach 1989).

10 Within the COSMAS project, all existing IDS corpora (about 28 million tokens) were converted into the BOT/MK format using a set of conversion scripts,[9] by 1993 they were accessible via the new corpus research system, also named COSMAS, to researchers at the IDS, and by 1996 to researchers all over the world via a web interface.

11 While the P2 version of the TEI Guidelines had been published in 1992, IDS staff chose not to adopt the TEI at that time, both because the IDS was not yet receiving any text data in SGML and because COSMAS had already been designed to use BOT/MK syntax.

12 Many useful types of information were only implicit in BOT/MK or were missing entirely and therefore unavailable for researchers to use in queries. Moreover, the follow-up project COSMAS II had started in 1995, and one of its goals was to allow the creation of virtual corpora (cf. Kupietz and Keibel 2009), but the original BOT/MK format did not contain all the fields necessary to do this. Consequently, in the years 1993–1998 many more fields were added to the BOT header, in particular new fields for the components of bibliographic information (this information had been included as an unparsed string in the first version of BOT). The more recent field names all start with 'BOT+', e.g. BOT+a (author), BOT+ti (title), BOT+u (subtitle), BOT+X (text type), BOT+b (volume), BOT+in (title of a collection in which the document or text was contained), and the above-mentioned BOT+ent. Altogether the revised BOT header has 38 fields available. Moreover, two basic templates of a BOT header were defined, each specifying a subset of the full set of BOT fields: Template 1 was used for independent works and dependent works contained in collections, and Template 2 was used for newspaper and journal articles.[10]

13 New texts to be added to the corpora were encoded according to the new version of BOT/MK. The values of the fields BOTd, BOTt, and BOTi, which contained the various versions of a bibliographic text string, were now automatically assembled at a later stage of the conversion from the fields that contained the components. By 1998, the IDS corpora comprised approximately 260 million tokens.

2.2. Conversion to IDS-CES

14 The year 1999 saw the start of DEREKO, a project for the acquisition and annotation of a German Reference Corpus (Deutsches Referenzkorpus),[11] conducted in cooperation with the universities of Stuttgart and Tübingen and lasting until 2002, when it reached 1.8 billion tokens.[12] Two important goals of DEREKO were, first, mass acquisition of texts by obtaining licenses from publishing houses and individuals, and, second, the use of CES, a new corpus encoding standard (Ide 1998) based on TEI. Between 1998 and 2003, a mapping of all BOT/MK fields and inline markup into the CES structure of elements and attributes was specified. Certain features of the BOT/MK markup, however, could not be rendered within the CES markup; therefore, additional elements and attributes were defined on top of CES, yielding IDS-CES, the IDS-specific adaptation of CES. As far as possible, the additional elements and attributes were taken from the TEI P3 Guidelines (ACH/ACL/ALLC 1994), but several had to be defined totally outside CES and TEI, with care taken to name and define them in the style of CES. In particular, it was decided that the three-way hierarchical structure with the units *corpus*, *document*, and *text* should be

retained, although CES/TEI provided only <cesCorpus> and <cesDoc>. Hence, <idsCorpus>, <idsDoc>, and <idsText> were defined to replace these. Another element that was newly introduced in IDS-CES is <creatDate> for the time of creation, i.e. for the value of the field BOT+ent.[13] Initially, IDS-CES was used as an exchange format only, as COSMAS still employed BOT/MK internally. Newly acquired texts (some of which arrived in SGML) were first encoded in BOT/MK, and a converter (TRADUCES[14]) was developed to transform the new and old BOT texts into IDS-CES.

15 Starting in 2001, the BOT/MK format was extended again, this time under the name "BOTX". For BOTX, new markup was defined: u+zz+, u+zzz+, etc. for sub-headings at different levels, li+ for list items, and other tags for tables, preface, table of contents, footnotes (which had previously not been marked up or had even been removed), and more textual features. The idea behind this extension was that all the features of a document—including not only previously unrepresented layout features but also tables of contents and imprint information—should be representable within the IDS text model so that the source document's layout would be reconstructible. The endeavor was also inspired by the many elements and attributes offered by CES for document features that were not captured by BOT/MK. Many of these features were in fact already marked up in the SGML source documents that the IDS received but then dropped in the BOT/MK representation, so it seemed worthwhile to make an effort to retain them. For some time, all incoming texts were converted to BOTX using small specialized programming routines that a programmer designed by checking the original layout in the corresponding hardcopy edition of the text. Since all the older corpus texts remained in plain BOT/MK, all BOTX texts first had to be converted to BOT/MK (that is, some markup had to be removed automatically) for their integration in COSMAS I.

16 BOTX and BOT/MK still had some flaws. For instance, the order of MK annotations (the inline annotations) is not fixed and was sometimes unclear—for example, when a passage is in a foreign language, is a quotation, and is printed in italics. Moreover, some tags are ambiguous character sequences that happen to appear in the source text, so since around 2004 several alternative tags taken from the TEI have been introduced in BOT/MK, such as <line>...</line> instead of '.../'. Incoming texts were marked up with the new tags and then converted to IDS-CES, but the existing corpora were not retroactively changed.

17 For completeness, we would like to mention that around 2007, some more new markup was added to BOT/MK only, namely the three fields BOT+D, BOT+V, and BOT+R for specifying the results of the IDS duplicate detection module, and the field BOT+th for results from the IDS thematic classification module.[15] These fields are mapped to <classDecl> and sub-elements in IDS-CES.

18 IDS-CES was introduced as the internal corpus representation format in COSMAS II, the successor of the research software COSMAS I, which was finally taken out of service in 2003. Under COSMAS II, the BOTX texts were directly converted to IDS-CES without loss of information.[16]

2.3. Conversion to IDS-XCES

19 In 2000, the first XCES specification was released (Ide, Bonhomme, and Romary 2000), in which the SGML-based Corpus Encoding Standard was redefined on an XML basis. In 2006,

an IDS-XCES DTD was developed, consisting of the XCES DTD with the addition of those elements and attributes that had already been added to CES to form IDS-CES.

20 The corpus archive (containing around 2.4 billion tokens at that time) was converted in 2006. The mapping from IDS-CES to IDS-XCES was entirely automated using XSLT.[17] (The differences between XCES and IDS-XCES are described in the appendix.) In 2008, IDS-XCES was introduced as the internal corpus data format in COSMAS.

21 All incoming texts are still initially encoded in the IDS pivot format BOT/MK or its extension BOTX. So the chain of conversions for new text data to be integrated in COSMAS II is currently *original format* → BOT(X) → IDS-CES → IDS-XCES.

22 The following diagram illustrates the long and complex development of IDS-XCES described above, by which this format was derived from TEI P3 and TEI P4 (TEI Consortium 2001), through CES and XCES and the local changes at IDS. As a result of the long chain of derivation, the relationship between the text model of the TEI Guidelines and the IDS text model (and the corresponding differences between the two markup systems) are hard to take in at a glance.

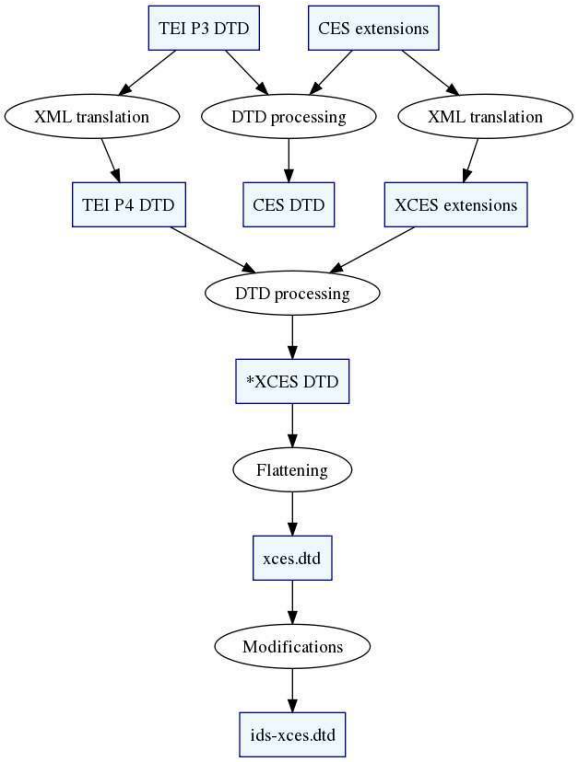


Figure 1: Development of the IDS-XCES DTD

23 In fact, one motivation for preparing a TEI P5-based ODD file for the IDS text model is to make the relation between the two text models simpler and clearer.

24 The Appendix gives a brief account of IDS-XCES, mainly by specifying its differences to the original XCES.

3. Preparation of an ODD File

25 As the summary above has made clear, elements and attributes from TEI P3 and TEI P4 come into IDS-XCES through two channels: some are inherited through XCES, while others are retroactively added in IDS-XCES.

26 The goal of the I5 project is to reorganize the definition of the IDS vocabulary as a single set of modifications taking TEI P5 as its base and using the new customization mechanism specified in TEI P5, which uses an ODD ("one document does it all") file to specify a particular customization instead of relying on the customization mechanisms built into a particular schema language. TEI P5 defines a specific XML tag set for use in ODD files and prescribes an algorithm for processing ODD files to generate customized versions of the TEI encoding scheme. This prescribed algorithm is implemented by software available from the TEI Consortium under the name *Roma*. As indicated in the diagram below, Roma reads the TEI P5 specification of the vocabulary and the ODD file provided by the user and generates on demand from them a DTD, a schema document in Relax NG or XSD notation, or reference documentation for the elements and attributes included in the specified customization.

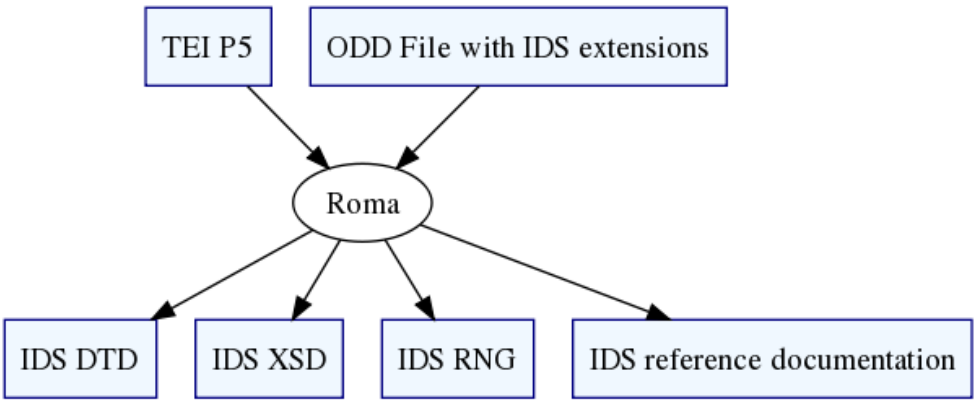


Figure 2: Generating document grammars and documentation using Roma

27 The immediate concrete goal of the I5 project, therefore, is to prepare an ODD file which will, when processed by Roma, produce document grammars suitable for use by IDS in processing the DEREKO archive.

3.1. Conditions for the Language as a Set of Documents

28 What language should those document grammars describe? The core requirements for the language to be defined can be summarized schematically in this way:

- I5 : P5 ≃ IX : P3
- L(I5) ±⊆ L(P5)
- L(I5) ≡ L(IX)

29 In these and subsequent formulae, the following abbreviations are used for brevity:

- *P3*, *P4*, and *P5* denote the document grammars (or in some cases the languages defined by those grammars) of TEI P3, TEI P4, and TEI P5 respectively.
- *CES* and *XCES* denote the document grammars of the original Corpus Encoding Standard and its XML revision.

- *IX* denotes the document grammar of the current IDS-XCES DTD (as realized in file ids-xces.dtd).
- *I5* denotes the grammar developed by this project.
- For any document grammar *x*, *L(x)* denotes the language recognized (or defined) by *x*.
- *E(x)* denotes the set of elements defined in an XML vocabulary or document grammar *x*.
- *DRK* denotes the German Reference Corpus (Deutsches Referenzkorpus) DEREKO, viewed as a set of documents (and thus as a language defined by enumeration of its sentences).

30 With these notational conventions in place, we can reformulate the core requirements for the I5 project.

31 First, the I5 vocabulary should stand in roughly the same relation to TEI P5 as the current IDS-XCES vocabulary stands to TEI P3 (or TEI P4):

- I5 : P5 ≃ IX : P3

32 Second, the language should be (more or less) a subset of the language defined by TEI P5:

- L(I5) ±⊆ L(P5)

33 Third, the language defined by the I5 document grammar should at least in principle be equivalent, or nearly equivalent, to that defined by IDS-XCES:

- L(I5) ≡ L(IX) (?)

34 Since one goal of the project is to incorporate in I5 the improvements on TEI P3 incorporated in TEI P5, absolute equivalence in all details is probably not in fact desirable (hence the addition of the question mark vis-à-vis the initial formulation).

35 Some further requirements or desiderata can also be identified and expressed formulaically. If absolute equivalence is not the goal, then the language of I5 needs to be constrained in other ways. The language of I5 should be similar to that of IDS-XCES, even if not absolutely equivalent. That is, perhaps strict equivalence (≡) should be replaced by similarity (≃):

- L(I5) ≃ L(IX)

36 Every document in DEREKO should be legal against the new document grammar:

- DRK ⊆ L(I5)

37 Empirically, it can be observed that DEREKO exercises only a proper subset of the current document grammar IDS-XCES: for example, there are a number of attributes defined in the grammar which do not in fact occur in the corpus. It is a design decision (still being considered) whether to retain such currently unused constructs, in the expectation that they may be used later, or to eliminate them so as to make the language of I5 be more nearly equivalent to that of DEREKO:

- DRK ⊊ L(I5) ?
- DRK = L(I5) ?

3.2. Realization of a Grammar in ODD

38 To specify a particular customization of the TEI vocabulary, an ODD file must specify the inclusion or exclusion of individual:

- TEI modules
- elements within a module
- attributes within a module

39 Some examples may help make it clear how this is done.

40 For example, the following ODD-file fragment includes the `tei` and `core` modules in a customization:

```
<TEI ...> ...
  <specGrp xml:id="specgroup-core">
    <moduleRef key="tei"/><!-- required -->
    <moduleRef key="core"/>
    <!--* abbr address analytic author bibl biblScope biblStruct
        * corr date distinct editor foreign gap gloss head
        * hi imprint item l label lb lg list measure mentioned
        * monogr name note num orig p pb ptr pubPlace publisher
        * q quote ref reg respStmt sp speaker stage term time title
        *-->
      <p>Delete unneeded elements.</p>
      <specGrpRef target="#specgroup-core-deletions"/>
      <p>Rename some elements.</p>
      <specGrpRef target="#specgroup-core-renamings"/>
  </specGrp>
  ...
</TEI>
```

41 Individual elements may be excluded by specifying `mode="delete"` on an appropriate `<elementSpec>` element:

```
<elementSpec ident="add" module="core" mode="delete"/>
<elementSpec ident="addrLine" module="core" mode="delete"/>
<elementSpec ident="binaryObject" module="core" mode="delete"/>
<elementSpec ident="cb" module="core" mode="delete"/>
<elementSpec ident="choice" module="core" mode="delete"/>
```

42 I5 must deal with several different sets of elements:

- Some elements should be incorporated from TEI P5. TEI P5 elements not present in IDS-XCES, on the other hand, should be excluded.
- Elements present in XCES but not present in TEI P5 must be defined. (They could be taken over from an XCES ODD file, if one existed, but there is not currently any ODD-defined version of XCES.)
- Additional elements found in IDS-XCES but not in XCES or TEI P5 must be defined.

43 That is, E(I5) =

$$(E(IX) \cap E(P5)) \cup (E(IX) \cap (E(XCES) \setminus E(P5)) \cup (E(IX) \setminus (E(XCES) \cup E(P5))$$

44 Note that the elements in the last group are not necessarily IDS extensions to XCES: they may also include elements in TEI P3 which are inherited by IDS-XCES from TEI P3, but which are no longer included in the TEI vocabulary in version P5.

45 It is possible to identify the elements which belong in each of the subsets described manually, given sufficient patience and capacity for tedious detail. It is significantly more convenient, however, to make the machine help us in the task. This can be done in a three-step process:

- Encode the relevant document grammars as XML documents.
- Compare them using XQuery.
- Generate the appropriate ODD declarations automatically.

46 A number of tools exist which can provide XML representations of DTDs. For the work described here, we have used a simple application based on SWI Prolog (Wielemaker n.d.), which loads a DTD and emits an XML representation of the DTD. The following example shows a fragment of the IDS-XCES document grammar in this representation:

```
<dtd>

  <desc>This document
  (<code>2011/blackmesatech/IDS/interim/ids_xces_onefile.v3.xml</code>)
  is an XML representation of
  <code>2011/blackmesatech/IDS/interim/onefile.dtd</code> made by
  dtdxml.pl on <date value="2011-09-11">2011-09-11</date></desc>

  <elemdecl gi="gloss">
    <star>
      <or>
        <elem>#pcdata</elem>
        <or>
          <elem>abbr</elem>
          <or>
            <elem>date</elem>
            <or>
              <elem>num</elem>
              <or>
                <!--* ... *-->

              </or>
            </or>
          </or>
        </or>
      </or>
    </star>
  </elemdecl>
  <attlist gi="gloss">
    <att>
      <name>id</name>
      <type>id</type>
      <dft>
        <implied/>
      </dft>
    </att>
    <att>
      <name>n</name>
      <type>cdata</type>
      <dft>
        <implied/>
      </dft>
    </att>
    <att>
      <name>xml:lang</name>
      <type>cdata</type>
      <dft>
        <implied/>
      </dft>
```

```
    </att>
    <!--* ... *-->

  </attlist>
</dtd>
```

47 It is then a straightforward task to use XQuery to identify the first set of elements: IDS elements which appear in TEI P5:

```
(: find the IDS elements that appear in the TEI Guidelines :)

declare namespace TEI = "http://www.tei-c.org/ns/1.0";

declare variable $dir.TEI := "file:/home/TEI";
declare variable $dir.IDS := "file:/Users/cmsmcq/2011/blackmesatech/IDS";
declare variable $P5 := doc(
  concat($dir.TEI,
        "/P5/Source/Guidelines/en/guidelines-en.xml"
        ));
declare variable $ids-xces := doc(
  concat($dir.IDS,
        '/interim/ids_xces_onefile.v3.xml'
        ));

<elements>{
       for $e in $ids-xces/dtd/elemdecl
       let $gi := string($e/@gi),
           $elemspec := $P5//TEI:elementSpec
                        [@ident = $gi]
     where $elemspec
  order by $gi
    return <e gi="{$gi}" module="{$elemspec/@module}"/>
}</elements>
```

48 After setting up variables for the text of TEI P5 and the XML encoding of the IDS-XCES DTD (`$P5` and `$ids-xces`, respectively), the query identifies each element name in the IDS-XCES DTD (`$gi`, for *generic identifier*) and then finds (`$elemspec`) the specification for that element in TEI P5, if there is one. If such an element specification exists in TEI P5, then an XML element is returned giving the name of the element and its module.

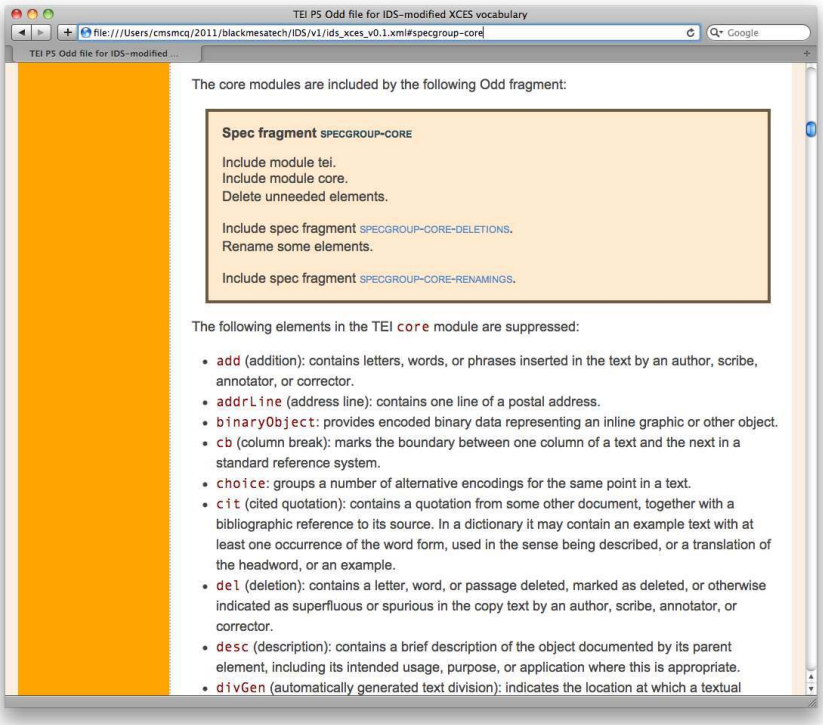
49 Once the basic query is formulated, it is simple to modify the `return` statement to return instead the appropriate ODD declaration for the element:

```
(: ... :)
    return <elementSpec  module="{$module}"
              ident="{$teigi}" mode="change">
              <altIdent>{$idsgi}</altIdent>
           </elementSpec>
(: ... :)
```

50 Similar queries can be constructed to generate appropriate ODD declarations for elements to be suppressed from TEI P5 or added to it.

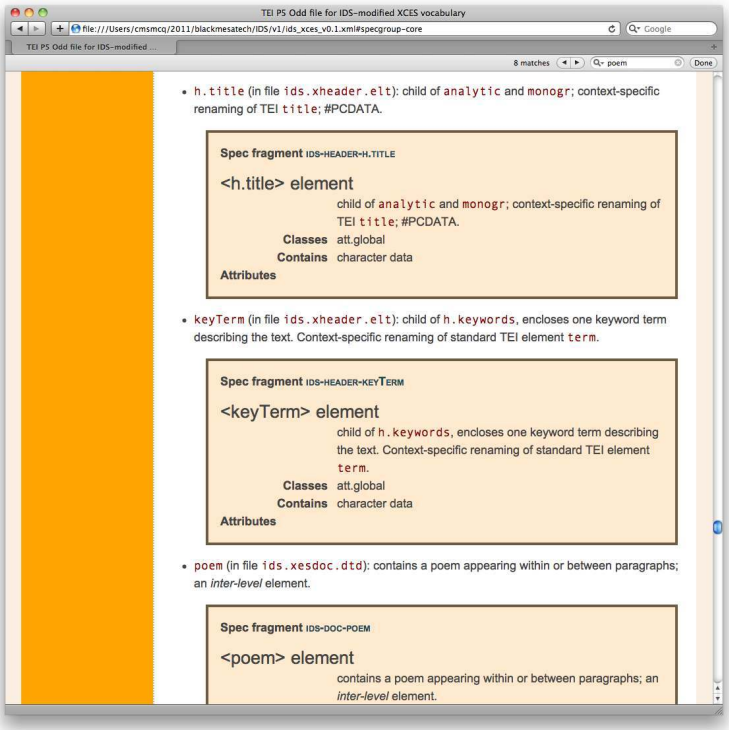
3.3. Documentation

51 The ODD file is designed as a form of literate program which allows us to embed the formal declarations of the document grammar in a human-readable document and intertwine the schema with the documentation. The I5 project endeavors to make the relation of TEI P5 to the IDS customization of TEI P5 easier to understand by treating the ODD file itself not, as is sometimes done, primarily as input to Roma but primarily as a document intended for human readers. The screen shot below illustrates the principle: it shows the part of the document beginning with the ODD fragment given above, which embeds the `tei` and `core` modules of TEI P5, in a style derived from the IDS house style for Web pages.



52 A significant part of the effort in the I5 project is the preparation of appropriate tag-set documentation for IDS-specific elements and attributes and for IDS-specific usages for

standard TEI and XCES constructs. Descriptions of the elements and attributes of the document grammar are taken in part from TEI P5, in part transcribed from the XCES documentation, and in part written from scratch. The individual element specifications are embedded in the ODD file, as can be seen in the screen shot below; they can also be extracted by Roma and integrated with documentation for standard TEI elements in the form of reference documentation.



4. Conclusions

53 The history of the IDS text model and its markup reflects, in its individual way, several important trends in the processing of natural language data generally and in corpus linguistics more particularly.

54 From the earliest beginnings, corpus data at IDS used markup to record important information about the text and to make explicit certain information within the text which would otherwise have been inaccessible to automatic processing. The early collections, however, all used idiosyncratic markup. Because of the large (for the times) volume of data it had collected, IDS became convinced earlier than some projects of the need to develop a standardized system of text representation. Like many who are early to perceive the need for standardization, IDS developed its own standard format, in the form of BOT. This standardization effort paid off: it made feasible the significant investment in infrastructure represented by the COSMAS I project.

55 During COSMAS II, TEI markup was introduced in the form of CES. The broad coverage, non-prescriptive approach, and sheer size of TEI P2, P3, and P4 made them daunting to many prospective users: hard to understand and thus hard to adopt. CES and XCES, which took a more focused, domain-specific approach, were more prescriptive, smaller, and easier to understand; in consequence, they were easier for IDS to adopt as the basis for its

SGML and XML formats. Experience showed, however, that some TEI constructs omitted from CES as unnecessary for corpus-linguistic work were needed, after all, to handle the broad variety of texts and textual constructs which turn up in large corpora like DEREKO. So IDS-CES and IDS-XCES found it necessary to bring some elements and attributes back from TEI P3 and P4.

56 With I5, the IDS text model is directly derived from the TEI text model; the relation of I5 to TEI, defined as it is by a single ODD file, will be somewhat easier to discern than the relation of IDS-CES to TEI P3 or of IDS-XCES to TEI P4. The relation to XCES will still be relatively easy to identify: by comparing the I5 ODD file to the extension files of XCES, any reader will be able to see which TEI elements are retained in one customization but not the other, and which additional elements and attributes are common to the two.

Appendix: Features of IDS-XCES

57 As indicated in Section 2.3, the format IDS-XCES is based on the XCES document grammar as defined in Ide, Bonhomme, and Romary (2000) and the XCES DTD files.[18] These DTD files have been taken and modified as necessary for the IDS text model as described below. The IDS-XCES document grammar comprises the files `ids-xcesdoc.dtd`, `ids-lat1.ent`, `ids.xcustomize.ent`, and `ids.xheader.elt`;[19] the XCES DTD files `xcesAlign.dtd` and `xcesAna.dtd` have no equivalents among the IDS-XCES DTD source files. The former file has no equivalent because there is no need to align corpus data in the monolingual German reference corpus. The latter has no equivalent because linguistic annotations, apart from the sentence segmentation, play (almost) no role in the IDS text model. Instead, several layers of linguistic markup are provided as standoff annotation in separate files.[20] Still, IDS-XCES allows the specification of morphosyntactic annotations in attributes of the `<w>` element: for a small number of corpora, there are versions of IDS-XCES documents with inline linguistic annotations added to the element `<w>`.

58 In IDS-XCES, some IDS-specific elements and attributes have been added to the original XCES, and in doing so, some of the XCES content models have been modified. These additional elements and attributes can be grouped into those that are essentially (context-dependent) renamings of XCES elements, those that have been taken from the TEI P3 (or P4) specification (such as `<textDesc>` and `<front>`) and those that are neither in XCES nor in TEI P3 (or P4).

59 In the following sections, we will give a summary of the most important features of IDS-XCES compared with XCES. We give examples of elements and their characteristics, without presenting complete content models including all attributes. The complete changes are documented formally in a synopsis at http://www.ids-mannheim.de/kl/projekte/korpora/idsxces.html.

A.1. Corpus Structure and Header

Element name	Possible parents	Modeled on	Meaning	Example
`<idsCorpus>`	xml document root	XCES: `cesCorpus`	*corpus*	
`<idsDoc>`	`<idsCorpus>`	XCES: `cesDoc`	*document*	

<idsText>	<idsDoc>	XCES: cesDoc	*text*	
<idsHeader>	<idsCorpus>, <idsDoc>, <idsText>	XCES: cesHeader	*header*	
<korpusSigle>	<titleStmt>	IDS-specific	*corpus ID (formerly BOTC)*	<korpusSigle>DIV</korpusSigle>
<dokumentSigle>	<titleStmt>	IDS-specific	*document ID (formerly BOTD)*	<dokumentSigle>DIV/SGP</dokumentSigle>
<textSigle>	<titleStmt>	IDS-specific	*text ID (formerly BOTT)*	<textSigle>DIV/SGP.00000</textSigle>
<c.title>	<titleStmt>	IDS-specific	*corpus title*	
<d.title>	<titleStmt>	IDS-specific	*document title*	
<t.title>	<titleStmt>	IDS-specific	*text title*	
<pagination>	<editorialDecl>	IDS-specific	*whether page numbering is present or not (processing info; formerly BOTP)*	<pagination type="yes"/>*

* Pagination information is included in a @type attribute, which is available for many elements in both XCES and TEI.

Table 2: Examples of elements added for the description of the corpus structure and header

60 Those elements that are essentially renamings of XCES elements are the high-level components <idsCorpus>, <idDoc>, and <idsText>—representing the three-level corpus structure of the IDS text model (from <cesCorpus> and <cesDoc>)—and <idsHeader> (from <cesHeader>).

61 In the content model for the <idsHeader>, the CES element <titleStmt> has been substantially revised to contain one of <korpusSigle>, <dokumentSigle>, or <textSigle> and one of <c.title>, <d.title>, or <t.title> to mark the ID and title of a corpus, document, or text (respectively).

A.2. Front and Back Matter

Element name	**Possible parents**	**Modeled on**	**Meaning**	**Example**

<front>	<text>	TEI P3/P4: <front>	*front matter*	
<back>	<text>	TEI P3/P4: <back>	*back matter*	
<titlePage>	<front>	TEI P3/P4: <titlePage>	*title page*	
<docTitle>	<titlePage>	TEI P3/P4: <docTitle>	*document title as part of the source*	<docTitle><titlePart type="main"><s>Jacques Hilarius Sandsacks Psychoschmarotzer</s></titlePart><titlePart type="desc"><s>Roman</s></titlePart></docTitle>*
<docImprint>	<front>	TEI P3/P4: <docImprint>	*imprint*	<docImprint>Aufbau-Verlag</docImprint>

* Since <docTitle> contains the title as it occurs printed in the source, it is part of the object text, and can be divided into sentence-like divisions marked by <s>.

62 One group of non-XCES elements found in IDS-XCES includes <front> and <back> and their child elements, all of which were taken from the TEI P3 (or P4) specifications.

A.3. Drama

Element name	Possible parents	Modeled on	Meaning	Example
<stage>	, <sp>, <s>	TEI P3/P4: <stage>	*stage direction or extra-linguistic event in debate*	<stage>(Beifall bei der CDU/CSU und der FDP)</stage>

63 For the encoding of drama and records of parliamentary debates, the element <stage> (for stage directions) was adopted from TEI P3 (or P4).

A.4. Page Breaks and Pointers

Element name	Possible parents	Modeled on	Meaning	Example
<pb>	non-header elements with mixed content	TEI P3/P4: <pb>	*page break*	<pb n="38" TEIform="pb"/>
<lb>		TEI P3/P4: <lb>	*line break*	<lb TEIform="pb"/>

<ptr>		TEI P3/P4: <ptr>	*pointer to xml id*	<ptr rend="number" targType="note" targOrder="u" target="shs.00000-n2-f2"/>
<xptr>		TEI P3/P4: <xptr>		<xptr targType="pb" targOrder="u" doc="korpref.mk2" from="WF1.00004-168-PB168" to="DITTO" TEIform="xptr"/>

64 A group of "milestone" elements—<pb>, <lb>, <ptr>, and <xptr>—has been added to IDS-XCES as part of almost all mixed content models. They are adopted from the TEI P3 (or P4) to mark page breaks, line breaks, and references to other corpora, documents, texts (e.g. from a bibliography section), sections, pages etc.

A.5. Corrections and Completions

Element name	Possible parents	Modeled on	Meaning	Example
<orig>	non-header elements with mixed content	TEI P3/P4: <orig>	*spelling variant or morphological ellipsis*	<orig reg="Ferienheime">Ferien-</orig> und Kinderheime

65 The <orig> element with its attribute @reg has also been adopted from the TEI P3 (or P4). In some corpora it is used to mark and complete morphological ellipsis and sometimes used to mark spelling variants.

A.6. Morphosyntactic Inline Annotations

Element name	Possible parents	Modeled on	Meaning	Example
<w>	non-header elements with mixed content	TEI P3/P4: <w>	*wordform*	<w ana="NOU com sg n dat">Telefon</w>

66 The <w> element with its attribute @ana has been adopted from the TEI P3 (or P4) to mark word forms and to provide morphosyntactic analyses for them. Only a handful of the IDS corpora, however, contain such inline annotations.

A.7. Time of Creation

Element name	Possible parents	Modeled on	Meaning	Example
<creatDate>	<creation>	IDS-specific	*time of creation and (short version of) reference to first edition*	<creation><creatDate>2001</creatDate><creatRef> (Erstveröffentlichung: Frankfurt a.M., 2001)</creatRef><creatRefShort> (Erstv. 2001)</creatRefShort></creation>

<creatRef>	<creation>	IDS-specific		
<creatRefShort>	<creation>	IDS-specific		

67 The elements under <creation> are used to encode available information about the time of creation of a text and the publication date of the first edition, if known. In TEI P3 and P4, the contents of <creation> can be marked up using generic <bibl> and <date> elements, but TEI does not provide an unambiguous way to indicate that a particular bibliographic reference and date inside a <creation> element are for the first edition. In CES and XCES, <creation> contains only character data with no substructure at all.

A.8. Text Description

Element name	Possible parents	Modeled on	Meaning	Example
<textDesc>	<profileDesc>	TEI P3/P4: <textDesc>	*wrapper for text description*	
<textType>	<textDesc>	IDS-specific	*text type according to type inventory (BOT+x)*	<textType>Roman</textType>
<textTypeRef>	<textDesc>	IDS-specific	*text type as to appear in bibliographic string (BOT +X)*	<textTypeRef>Tageszeitung</textTypeRef>
<textTypeArt>	<textDesc>	IDS-specific	*text type of a specific article (BOT +xa)*	<textTypeArt>Interview</textTypeArt>
<textDomain>	<textDesc>	IDS-specific	*subject area (BOT+r)*	<textDomain>Regionales / Unterhaltung/Kultur</textDomain>
<column>	<textDesc>	IDS-specific	*original label of newspaper section as in the source (BOT+ress)*	<column>FERNSEHEN</column>

68 IDS-specific elements under <textDesc> are used to encode genre, text type, newspaper section or subject area according to different classification schemes.

A.9. Edition Information

Element name	Possible parents	Modeled on	Meaning	Example
<further>	<edition>	IDS-specific	*further edition of the same source with year (BOT+gg)*	<further>5. Auflage 1998 (1. Auflage 1997)</further>
<kind>		IDS-specific	*kind of edition of the source (BOT+g)*	<kind>Taschenbuch</kind>
<appearance>		IDS-specific	*"physical" appearance of the source (BOT+e)*	<appearance>Microfiche</appearance>

69 IDS-specific elements under <edition> are used to encode information about other existing editions or the range of existing editions, the kind of edition (paperback, special edition etc.), and the kind of object that was used as the source (photocopy, microfiche, etc.).

A.10. Bibliographic Reference

Element name	Possible parents	Modeled on	Meaning	Example
<reference>	<sourceDesc>	IDS-specific	*bibliographic reference string*	<reference type="short" assemblage="regular">DIV/SGP.00000 Szendrödi: Jacques Hilarius Sandsacks Psychoschmarotzer, 2001</reference>

70 The element <reference> may appear multiple times under <sourceDesc>, with different values of its @type attribute specifying various versions of the bibliographic reference string required for different modes of display and information in the @assemblage attribute about whether it has been automatically assembled from other elements or not.

BIBLIOGRAPHY

Association for Computers and the Humanities (ACH), Association for Computational Linguistics (ACL), and Association for Literary and Linguistic Computing (ALLC). 1999. *Guidelines for Electronic Text Encoding and Interchange (TEI P3)*, edited by C. M. Sperberg-McQueen and Lou Burnard. Chicago and Oxford: Text Encoding Initiative. First published 1993. http://www.tei-c.org/Vault/GL/P3/index.htm.

al-Wadi, Doris and Irmtraud Jüttner. 1996. "Textkorpora des Instituts für Deutsche Sprache: Zur einheitlichen Struktur der bibliographischen Beschreibung der Korpustexte." In *LDV-INFO 8. Informationsschrift der Arbeitsstelle Linguistische Datenverarbeitung*, edited by IDS, 1–85. Mannheim.

Belica, Cyril, Marc Kupietz, Andreas Witt, and Harald Lüngen. 2011 "The Morphosyntactic Annotation of DEREKO: Interpretation, Opportunities, and Pitfalls." In *Grammatik und Korpora 2009. Dritte Internationale Konferenz. Mannheim, 22.4.–24.9.2009*, edited by Marek Konopka, Jacqueline Kubczak, Christian Mair, František Šticha, and Ulrich Hermann Waßner, 451–469. Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache 1. Tübingen: Narr.

Ide, Nancy. 1998. "Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora." *Proceedings of the First International Language Resources and Evaluation Conference*, 463–470. Grananda, Spain.

Ide, Nancy, Patrice Bonhomme, and Laurent Romary. 2000. "XCES: An XML-based Standard for Linguistic Corpora." In *Proceedings of the Second Language Resources and Evaluation Conference (LREC)*, 825–830. Athens, Greece.

Kolvenbach, Monika. 1988/1989. "Schreibkonventionen für IDS-Korpora." In *LDV-INFO 7. Informationsschrift der Arbeitsstelle Linguistische Datenverarbeitung*. Edited by Tobias Brückner.

Kupietz, Marc. 2005. *Near-Duplicate Detection in the IDS Corpora of Written German*. Technical Report KT-2006-01. Institut für Deutsche Sprache, Mannheim.

Kupietz, Marc and Holger Keibel. 2009. "The Mannheim German Reference Corpus (DEREKO) as a basis for empirical linguistic research." *Working Papers in Corpus-based Linguistics and Language Education* 3, edited by Makoto Minegishi and Yuji Kawaguchi, 53–59. Tokyo: Tokyo University of Foreign Studies (TUFS).

Kupietz, Marc, Oliver Schonefeld, and Andreas Witt. 2010. "The German Reference Corpus: New developments building on almost 50 years of experience." In *Language Resources: From Storyboard to Sustainability and LR Lifecycle Management*, edited by Victoria Arranz and Laura van Eerten. http://www.lrec-conf.org/proceedings/lrec2010/workshops/W20.pdf

Perkuhn, Rainer, Cyril Belica, Doris al-Wadi, Meike Lauer, Kathrin Steyer, and Christian Weiß. 2005. "Korpustechnologie am Institut für Deutsche Sprache." In *Korpuslinguistik deutsch: synchron – diachron – kontrastiv*, edited by Johannes Schwitalla and Werner Wegstein, 57–70. Tübingen, Germany.

TEI Consortium. 2001. *TEI P4: Guidelines for Electronic Text Encoding and Interchange: XML-Compatible Edition*, edited by C. M. Sperberg-McQueen and Lou Burnard. N.p.: TEI Consortium. http://www.tei-c.org/release/doc/tei-p4-doc/html/.

TEI Consortium. 2012. *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* Version 2.1.0. Last updated June 17. N.p.: TEI Consortium. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html.

Wielemaker, Jan. n.d. "SWI-Prolog SGML/XML parser." SWI-Prolog. http://www.swi-prolog.org/pldoc/package/sgml.html.

NOTES

1. http://www.ids-mannheim.de/kl/projekte/korpora/

2. http://www.ids-mannheim.de/cosmas2/

3. We would like to thank Doris al-Wadi, Cyril Belica, Marc Kupietz, and Eric Seubert for their enormous help regarding our questions about the history of the IDS text model.

4. Texts from 1989–1990 that document the political change that led to reunification, prepared by the IDS and the former *Zentralinstitut für Sprachwissenschaft*.

5. Bonn newspaper corpus, from various years between 1949 and 1974, prepared in the 1970s.

6. Mannheim corpus 1 and 2, with texts from 1949 to 1974.

7. Cf. Perkuhn 2005 et al., p. 61 (our translation).

8. Since the document identifier consists of three capital letters usually derived from the initials of the author and/or initials of content words from the title of the document, the resolved document identifier (field BOTd) also corresponds to an abbreviated version of the bibliographic reference.

9. Like BOT itself, the conversion scripts were prepared by Cyril Belica.

10. The many additional BOT fields and the two basic templates were all specified by Doris al-Wadi and Irmtraud Jüttner (al-Wadi and Jüttner 1996).

11. The name DEREKO (Deutsches Referenzkorpus) has been in use since then for the archive of contemporary written-language corpora at the IDS.

12. The project would later be called DEREKO-I.

13. The specification of the mapping and the definition of IDS-specific elements were prepared by Doris al-Wadi of IDS.

14. TRADUCES was developed by Eric Seubert of IDS.

15. These fields were added by Marc Kupietz of IDS (Kupietz 2005).

16. BOTX was defined by Eric Seubert.

17. The specification of the mapping and the conversion script were prepared by Marc Kupietz.

18. These may be downloaded from http://www.xces.org/dtds.html.

19. These files may be downloaded from http://corpora.ids-mannheim.de/idsxces1/DTD/
.

20. See Belica et al. 2011.

ABSTRACTS

This paper describes work in progress on I5, a TEI-based document grammar for the corpus holdings of the Institut für Deutsche Sprache (IDS) in Mannheim and the text model used by IDS in its work. The paper begins with background information on the nature and purposes of the corpora collected at IDS and the motivation for the I5 project (section 1). It continues with a description of the origin and history of the IDS text model (section 2), and a description (section 3) of the techniques used to automate, as far as possible, the preparation of the ODD file documenting the IDS text model. It ends with some concluding remarks (section 4). A survey of the additional features of the IDS-XCES realization of the IDS text model is given in an appendix.

INDEX

Keywords: corpora, ODD, DTD, CES, XCES

AUTHORS

HARALD LÜNGEN

Harald Lüngen has been a researcher in the area of corpus linguistics at the Institut für Deutsche Sprache in Mannheim, Germany, since 2011, specialising in the construction and maintenance of the German Reference Corpus DEREKO and in methods of corpus analysis. Before that, he worked as a computational linguist and project scientist in the fields of computational lexicology and morphology, text parsing, and text technology.

C. M. SPERBERG-MCQUEEN

C. M. Sperberg-McQueen (Black Mesa Technologies LLC) is a consultant specializing in helping memory institutions solve information management problems and preserve cultural heritage information for the future by using descriptive markup: XML, XSLT, XQuery, XML Schema, and related technologies. He co-edited the XML 1.0 specification and the first versions of the TEI Guidelines.