Text metadata

What the header of a text item looks like[1]

DK-CLARIN WP 2.1 Technical Report
Jørg Asmussen, DSL, with input from other WP 2 members
Final version of May 5, 2015[2]

Deliverables concerned

D13 TEI transducer The original plan for WP 2.1 was based on the assumption that the repository of potential corpus texts – the corpus text bank – most likely would have a non-XML structure (relational db). In order to make interchange of texts easy and in order to make them fit into the intended resource repository of DK-CLARIN, the development of a transducer that could reshape the texts and metadata stored in the corpus text bank to valid TEI XML seemed necessary. However, during the course of the project, it became clear that the text bank itself should be implemented as an XML database so that the texts could be stored in their final TEI XML format. Therefore, the task of developing a transducer became a task of defining an appropriate subset of TEI in order to suit the metadata and text format needs of DK-CLARIN. **Outcome:** Report.

[1] A header/text template can be downloaded from:
http://ctb.dsl.dk/templates/formatsample.xml
The corresponding XML schema is available at:
http://dkclarin.dk/schemas/WP2
Schema read-me file available at:
http://dkclarin.dk/schemas/WP2/README_TEIP5DKCLARIN_validation.pdf

[2] A more recent version may be available at:
http://korpus.dsl.dk/clarin/corpus-doc/text-header.pdf

Outline of this document

This technical report describes how the metadata part of text items can be expressed by means of a TEI P5 header whereas Asmussen (2013b) describes the text part proper. One major aim of the header design described in this technical report is to integrate header information from text items in existing corpora of Danish language, i.e. the Corpus of the Danish Dictionary and PAROLE-DK, KORPUS 2000, other corpus-relevant material from DOT/DSL, as well as the LGP and LSP corpora of written Danish which are compiled as part of DK-CLARIN.

Guide to reading this document

The structure of the header is oriented towards that one used by the BNC Burnard (2007) and PAROLE-DK Keson (1998b) but tries to avoid idiosyncrasies not covered by TEI P5 as well as modifications of the TEI header schema.

Section 1 summarizes some corpus linguistic concepts used throughout the DK-CLARIN project, which are described in further detail in Asmussen (2013a).

Section 2 gives a general account of the header structure of headers of text items to be included in the *Corpus Text Bank*, CTB.[3] The description of the CTB header structure is in its starting point strongly inspired by that one given in Burnard (2007). This section constitutes the major part of this report.

Section 3 starts with a complete header template and describes in detail the sets of values that have to be used to fill in the header. It can be used as a manual

[3] The CTB is a text repository of written texts that are candidates to be included in a linguistic corpus. The CTB has been developed by WP 2.1 in order to better process and organize potential corpus text material. It must not be confused with the general DK-CLARIN repository developed in WP 5 that is supposed to support various data types (e.g. texts, images, lexicons) and various formats.

for those who have to fill in text headers with appropriate information, either manually or automatically by converting and mapping existing material. This section is probably too detailed for those readers who just want the more general lines of how the CTB header is composed and may therefore be skipped by most readers.

1 Concepts

A *text item* consists of a *text* potentially to be included in a corpus, and of some metadata about the text. The metadata is typically contained in a *header* which precedes the text proper.[4] A text item is the smallest chunk of text plus metadata in a repository of potential corpus texts – a *corpus text bank* – from which text items are selected for inclusion in a specific corpus. Thus, a text item is the smallest corpus-compositional unit. The text part of a text item is either a complete text (usually a shorter one) or a sample taken from a longer text, e.g. a chapter from a book, see Asmussen (2013a). Longer texts, e.g. novels, are divided into smaller parts, e.g. chapters, before they are included in a corpus text bank. A corpus text bank may be considered as a somewhat more specialized kind of text archive, intended to contain all kinds of corpus-relevant text chunks. The reason why longer texts are chopped into smaller chunks is that this subsequently makes corpus composition more precise as text-typological fine-tuning becomes easier – a novel, for instance, is less likely to skew the intended balance of a corpus if it can be selected from the text bank in smaller quantities, e.g. chapter-wise.

This technical report describes the header structure of text items collected in the *Corpus Text Bank* (CTB) – a corpus text bank for all kinds of written corpus-relevant texts collected as part of the DK-CLARIN project's work package 2.1: "Basic written language resources — Reference corpus of general language". Text items from the CTB may be included in one or more specific corpora intended for linguistic research. A *corpus* is a more organized collection of texts compiled on the basis of the text bank for a specific – i.e. linguistic – purpose. Text material being collected for literary purposes or as part of an electronic library (archive) may stress other features of the TEI header proposal. Here, the header structure is adopted to the specific needs of *corpus* texts.

Text item headers are structured by means of TEI P5. In the following, this structure adapted to the needs of structurally integrating various existing corpora or text collections is described in detail. The collections to be structurally integrated are the *Corpus of the Danish Dictionary* (DDOC, Norling-Christensen and Asmussen (1998)), *PAROLE-DK* (Keson (1998a) and Keson (1998b)), *KORPUS 2000* (Andersen et al. (2002)), other corpus-relevant material from DOT/DSL and Dansk Sprognævn (DSN), as well as the LGP and LSP corpora of written Danish which are compiled as part of DK-CLARIN.[5]

[4] Another solution would be to store the metadata in a separate database and establish a link between text and metadata.

[5] Text material from the *Arkiv for Dansk Litteratur* (ADL) and other archives may at a later stage

The TEI header structure provides extremely flexible means of expressing textual metadata. A wealth of information can be given in a more or less fine-grained way. The following Section 2 describes a header that exactly accommodates the needs of potential corpus texts. In many cases, TEI allows the header to be modified either by augmenting or simplifying it. However, a header with more or less information is still compatible with the model described here as long as its structure does not conflict with TEI P5 syntax (and semantics) requirements.

Therefore, the following section does not describe a TEI header in general, but the specific header of a potential corpus text in the Corpus Text Bank of WP 2.1, expressed by means of TEI.[6]

2 Header structure

The header of a text item provides a structured description of the text contents, analogous to the title page and front matter of a book. Every separate text item in the Corpus Text Bank has its own header `<teiHeader type="text">`. In addition, a corpus itself may have a header `<teiHeader type="corpus">` containing information which is applicable to the whole corpus. The corpus header is not part of this description. To a large extent, a corpus header would be an abridged and slightly modified version of a text header. Furthermore the corpus header should contain the declaration of value sets for various elements (e.g. a domain taxonomy for LSP texts). The Corpus Text Bank contains value declarations in form of a collection of certain value set files which may be referenced by the CTB header. The content structure of the Corpus Text Bank is described in detail in **?** The value set files proper are described in detail in Section 3.2.

The remainder of this section describes the components of the `<teiHeader type="text">` element as used within the Corpus Text Bank. A TEI header contains a file description (Section 2.1), an encoding description (Section 2.2), a profile description (Section 2.3), and a revision description (Section 2.4), represented by the following four elements:

`<fileDesc>` (file description) contains a full bibliographic description of an electronic text as well as the source from which it was derived.

`<encodingDesc>` (encoding description) documents the relationship between an electronic text and the source from which it was derived.

`<profileDesc>` (text-profile description) provides a detailed description of non-bibliographic aspects of a text, specifically the languages and sublanguages used, the situation in which it was produced, the participants and their setting.

be integrated as well, if the header structure of their texts can be mapped to that one described here.

[6] The header design has been adopted for text resources to be included in the DK-CLARIN repository developed by WP 5.

`<revisionDesc>` (revision description) summarizes the revision history for a file.

2.1 The file description

The file description `<fileDesc>` is the first of the four main constituents of the header. It is intended to document a digital file. It contains the following four subdivisions:

`<titleStmt>` (title statement) groups information about the title of a work represented in the electronic text sample and those responsible for its intellectual content.

`<extent>` specifies the size of the electronic text sample in number of words and paragraphs (and other countable units).

`<publicationStmt>` (publication statement) groups information concerning the publication or distribution of the electronic text sample.

`<notesStmt>` (notes statement) collects together any notes providing information about a text additional to that recorded in other parts of the bibliographic description.

`<sourceDesc>` (source description) supplies a description of the source text from which the digital text sample was derived.

Further detail for each of these is given in the following subsections.

2.1.1 The title statement

The title statement `<titleStmt>` element of a text item contains one `<title>` element, followed by one `<sponsor>` and one `<respStmt>` element as shown in the following pattern:

```
<titleStmt>
  <title>
    samplingDeclaration textTitle
  </title>
  <sponsor>sponsorName</sponsor>
  <respStmt>
    <resp>Data capture</resp>
    <name>organizationName
      <note type="method">captureMethod</note>
      <date when="captureYear"/>
    </name>
  </respStmt>
</titleStmt>
```

The content of the <title> element is an initial caption (*samplingDeclaration*), e.g. "CTB version of:",[7] followed by the title of the source text (*textTitle*). Thus, the contents of the title element resemble that one used in PAROLE-DK: "Tagged sample of: '*textTitle*' ". As the CTB virtually can contain both tagged (even differently tagged) and untagged text, any statements about whether the text is tagged in some respect or not must not be made in the <title> element but should be given as *application information*, see Section 2.2.3.

The <title> element is followed by a <sponsor> element indicating the name of the sponsoring organization or institution.[8] According to the TEI guidelines, sponsors give their intellectual authority to a project; they are to be distinguished from funders, who provide the funding but do not necessarily take intellectual responsibility. The <sponsor> content of material captured as part of the DK-CLARIN project is "DK-CLARIN". Texts which were captured in other projects (and made available to DK-CLARIN) have their own specific <sponsor> content.

A <respStmt> element is used to indicate each institution responsible for any significant effort in the creation of the electronic text sample. The CTB header has only one responsibility statement indicating the responsibility for original data capture. The name of the responsible institution is given as an constant string for each institution in a <name> element. The <note> element of type "method", subordinate to <name> gives an indication of how the text was captured, e.g. by scanning or typing. Finally, the year of data capture is given as a four-digit date (or a complete date) as the value of the *when* attribute in the <date> element subordinate to <name>.

PAROLE-DK's header does neither include sponsor nor responsibility information, whereas the BNC uses lots of <respStmt> elements with great verbosity. In PAROLE-DK, this information instead is virtually part of the <publicationStmt> assuming that the distributor is always the same as the organization responsible for data capture (and is the sponsor). Here, it is assumed that the sponsor, the collector, and the distributor are of central importance and that it cannot be taken for granted that these decisive roles are played by one organization only. However, it is assumed that these roles are fully sufficient to describe the institutional background of a potential corpus text. Additional roles may come into play for a whole corpus or text collection and must be part of the headers of these resources.

OBS! Author and editor information for the source from which a text is derived (e.g. the author of a book) is not included in the <titleStmt> element but in the <sourceDesc> element discussed below in Section 2.1.5.

[7] Other *samplingDeclaration* captions are acceptable as well. A complete list is given in Section 3. The chosen caption must always be identical to the string value given in the <samplingDecl> element, see Section 2.2.1. In the example given, *CTB* stands for *Corpus Text Bank*.

[8] An alternative (and probably more appropriate) expression instead of *sponsor* would be *initiative*.

2.1.2 The extent statement

The <extent> element is used in each text header to specify the size of the text to which it is attached. The size is given as the number of words in the <num> element, the *n* attribute is set to "words". In another <num> element with the n attribute set to "paragraphs" the number of paragraphs is stated.[9] Other <num> elements measuring extent in other units may be added, but must be registered as part of the legal inventory described in Section 3:

```
<extent>
  <num n="words">numberOfWords</num>
  <num n="paragraphs">numberOfParagraphs</num>
</extent>
```

The count given does not include the size of the header itself. The number of words and paragraphs must be mechanically computed prior to insertion of the text into the text bank.

2.1.3 The publication statement

The <publicationStmt> element is used to specify publication and availability information for an electronic text. It contains the following three elements:

<distributor> supplies the name of a person or agency responsible for the distribution of a text.

<availability> supplies information about the availability of a text, for example any restrictions on its use or distribution, its copyright status, etc.

<idno> (identifying number) supplies an identifying code for a text.

```
<publicationStmt>
  <distributor>organizationName</distributor>
  <idno type="textIdType">textId</idno>
  <availability status="availStatus">
    <ab type="availGroup">availDesc anonymisationDesc</ab>
    <ab type="availGroup">availDesc anonymisationDesc</ab>
    <ab type="availGroup">availDesc anonymisationDesc</ab>
  </availability>
</publicationStmt>
```

The <distributor> element contains the name of the organization[10] responsible for the distribution of the electronic text sample. Usually there can only be one

[9] This is a necessary extent information particularly for texts which are to be included in parallel corpora.

[10] In DK-CLARIN this will typically be a member of the DK-CLARIN consortium.

distributor for each text even though TEI allows to repeat this element as often as needed. The inventory of strings denoting distributors should be invariant, i.e. one name only per distributor.

The obligatory CTB text id is given as contents of an `<idno type="ctb">` element. Some dialects of TEI introduce an attribute *id* of the `<TEI>` element which is illegal according to strict TEI. Other types of text, project-, or institution-internal identifications may be given in additional `<idno>` elements whose type attributes indicate the specific type of id.

The text strings in `<ab>` ('anonymous block')[11] elements given under `<availability>` for both restricted (attribute *status* is set to "restricted") and free (attribute *status* is set to "free") give availability information for three fixed user categories: academic users, non-commercial users, and all types of users.

Academic users are defined as users who are affiliated with the DK-CLARIN consortium.

Non-commercial users are academic users not affiliated with the DK-CLARIN consortium, users from educational or governmental institutions.

All users are any type of users including commercial users.

The DK-CLARIN license committee has finally, i.e. at the end of the project, concluded that the types of licenses should be employed: public, academic and restricted and that licenses are to be managed outside text headers. However, WP 2.1 will stick to the categories and values described above.

The following pattern shows the substructure of the `<availability>` element:[12]

```
<availability status="restricted">
  <ab type="academic">
    <seg type="availDesc">availDesc</seg>
    <seg type="anonymDesc">anonymDesc</seg>
  </ab>
  <ab type="nonCommercial">
    <seg type="availDesc">availDesc</seg>
    <seg type="anonymDesc">anonymDesc</seg>
  </ab>
  <ab type="all">
    <seg type="availDesc">availDesc</seg>
    <seg type="anonymDesc">anonymDesc</seg>
```

[11] This type of elements is preferred to the alternative `<p>` which is semantically misleading – these are no paragraphs but blocks of information.

[12] The `<availability>` element requires subordinate `<p>` or `<ab>` elements thus inhibiting more meaningfully structured availability information. The cumbersome typed `<ab>` and `<seg>` elements thus seem to be the only way of expressing structured availability information, unless TEI P5 is modified.

```
    </ab>
  </availability>
```

The various values are defined in Section 3. Two types of values are given in two subordinate <seg> elements: The availability description *availDesc* and a description of how to anonymize private information associated with the text, *anonymDesc*. If availability for any user category is other than "full" or any kind of anonymization is required, that is if *anonymDesc* is other than "nothing" (i.e. value "0"), the availability *status* attribute is set to "restricted", otherwise it is set to "free".

TEI allows a <date> element as part of <publicationStmt>; however, it is left out here, as the CTB version of a text cannot be said to having been published at a given time. Text bank texts may undergo changes (e.g. annotations are modified, more detailed info is given in the header) some of which are time-stamped in the revision description of the header, see Section 2.4, so the texts can never be said to be final, but they are available at all times in the shape they have at a given point in time. However, they may be published as part of a corpus, hence the <date> element under <publicationStmt> should be part of the corpus header.

2.1.4 The notes statement

The <notesStmt> contains one or more <note> elements, each containing a single piece of descriptive information, which does not fit into other parts of the header. Each <note> element carries an obligatory *xml:lang* attribute indicating the language of the note as well as a *resp* attribute denoting the organization responsible for this note, that is, the organization that has authored this note:

```
<notesStmt>
  <note xml:lang="languageId"
  resp="organizationName">note</note>
</notesStmt>
```

2.1.5 The source description

The <sourceDesc> element is used to supply bibliographic details for the original source material from which an electronic text sample derives. In the case of DK-CLARIN corpus texts, this may be a book, pamphlet, newspaper, etc. or an electronic source of some (non-TEI) format. Within the <sourceDesc> element several sub-structures are available according to TEI. Here, the <biblStruct> sub-structure is used in almost the same way as in PAROLE because it imposes a fixed structure on the bibliographic description and, most importantly, because it allows to distinguish between information concerning the text proper and information concerning the edition (e.g. book, newspaper) from which the text was drawn:

```
<sourceDesc>
  <biblStruct>
    [...]
  </biblStruct>
</sourceDesc>
```

The <biblStruct> element contains the following three elements:

<analytic> (analytic level) contains bibliographic elements describing an item (e.g. an article or poem) published within a monograph or journal and – according to the TEI guidelines – not as an independent publication. In the CTB headers, though, it is used for independent publications as well, see below.

<monogr> (monographic level) contains bibliographic elements describing an item (e.g. a book or journal) published as an independent item (i.e. as a separate physical object).

<idno> (identifying number) supplies any standard or non-standard number used to identify a bibliographic item.

<relatedItem> may contain a reference to some other bibliographic item related to the present one in some specified manner, for example as a translation of it. However, the use of this element is deprecated as the quality and quantity of relationships between texts may vary depending on the perspective of the user, therefore they should not be treated as a fixed information in the header of a text. Instead, various relation files should be introduced that relate any number of texts to each other in any way. The format of these relation files should be defined in a technical report. The substructure of the deprecated <relatedItem> is:

```
<relatedItem type="relatedType">
  <bibl>
    <title xml:lang="languageId">relatedTitle</title>
    <idno type="ctb">relatedId</idno>
  </bibl>
</relatedItem>
```

It must be placed as last element in <biblStruct> and it may be repeated as many times as necessary.

The complete substructure of <biblStruct> looks as follows:

```
<biblStruct>
  <analytic>
    <title xml:lang="languageId"
      level="titleLevel">textTitle</title>
    <author>
```

```
        <name ref="#personId">surname, forename</name>
        <note xml:lang="languageId" resp="organizationName" >
          note
        </note>
      </author>
      <respStmt n="translators">
        <resp>Translated by</resp>
        <name ref="#personId">
          surname, forename
        </name>
      </respStmt>
    </analytic>
    <monogr>
      <title xml:lang="languageId">editionTitle</title>
      <editor>
        <name ref="#personId">surname, forename</name>
      </editor>
      <imprint>
        <publisher n="publId">publHouse</publisher>
        <date when="publDate" cert="certainty"/>
        <biblScope type="issue">edIssue</biblScope>
        <biblScope type="sect">edSect</biblScope>
        <biblScope type="vol">edVolume</biblScope>
        <biblScope type="chap">edChapter</biblScope>
        <biblScope type="pp">edPages</biblScope>
      </imprint>
    </monogr>
    <idno type="uri">textUri</idno>
    <idno type="file">textFileName</idno>
    <relatedItem type="relatedType">
      <bibl>
        <title xml:lang="languageId">relatedTitle</title>
        <idno type="ctb">relatedId</idno>
      </bibl>
    </relatedItem>
  </biblStruct>
```

According to the TEI guidelines,

> [in] common library practice a clear distinction is usually made between an individual item within a larger collection and a free-standing book, journal, or collection. Similarly a book in a series is distinguished sharply from the series within which it appears. An article forming part of a collection which itself appears in a series thus

> has a bibliographic description with three quite distinct levels of information: the analytic level, giving the title, author, etc. of the article; the monographic level, giving the title, editor, etc. of the collection; the series level, giving the title of the series, possibly the names of its editors, etc. and the number of the volume within that series.[13]

The aim of the bibliographic information for texts which are intended to be included in a corpus, that is the type of texts collected in the Corpus Text Bank, is not to imitate the precision of a librarian but to give an easy way of referring to texts and to probably use bibliographic information in some corpus searches as well. This requires a rather fixed and to some extent rigid structure of the bibliographic part of the header which is the reason why the <biblStruct> structure is used here and not one of the other (less structured) possibilities of TEI. The <biblStruct> structure can be used to distinguish between the three information levels discussed above in the TEI guideline snippet. Here, only two of the levels are used, namely the analytic and the monographic level. The <monogr> element in the <biblStruct> structure is obligatory. According to TEI, it seems that in the case of a text being monographic, the <analytic> part of the structure should be left out and the text title and author information should be given within the <monogr> part of the structure. However, in the CTB headers, the <analytic> part is considered *obligatory*, no matter whether the text is part of a collection of some kind, i.e. analytic, or a stand-alone publication, i.e. monographic. This is to ensure that all <biblStruct> elements in CTB headers have the same structure, that text title and author information is always found in the same place, that is in the obligatory <analytic> part of the structure.

Within the <analytic> structure, <title> always gives the title of the text. If the text is part of a collection, e.g. a newspaper article which is part of a newspaper, the *level* attribute of <title> is set to "a" which means *analytic*, whereas the <title> element in <monogr> gives the title of the collection, e.g. the name of a newspaper. If the text is a free-standing book, e.g. a novel, the *level* attribute is set to "m" meaning *monographic*; in such cases the <title> element in the <monogr> part is left empty. All <title> elements carry the obligatory attribute *xml:lang* indicating the language of the title.

The author of a text is always given in <author> in the <analytic> part of <biblStruct>. There is one <author> element for each author who has contributed to the document. The name of the author is given in a <name> element. If the name has been decomposed into forename and surname, the information is given as *surname, forename(s)*, otherwise the comma is left out. If the name of the author is unknown, the <name> element is filled in with an *unknown* symbol (see Section 3), if an author for some reason is anonymous, the <name> element is filled in with the string "NN". A <name> element should have a *ref* attribute giving an XML reference to a corresponding <person> element in the <profileDesc>

[13] See http://www.tei-c.org/release/doc/tei-p5-doc/en/html/CO.html.

part of the header where additional info concerning the author(s) is given, see Section 2.3.5.[14] If texts are converted from existing corpora, e.g. the *Corpus of the Danish Dictionary, DDOC*, having a more elaborate description of the authors, e.g. place of birth, education, profession, there is no other way of expressing this information in the header structure other than by putting it into the <note> element together with the xml:lang and resp attributes giving the person or organization responsible for this note and the language of this note content.[15]

PAROLE has no participant description as part of the profile description. Instead, PAROLE augments TEI by adding two arguments (*gender* and *born*) to the <author> element. In contrast to PAROLE, the CTB header defers from altering the TEI proposal.

The <author> element is followed by a <respStmt> with an obligatory attribute *n* carrying the constant value "translators" that contains the name(s) of the person(s) who has/have translated this text if it is a translation, otherwise <respStmt> is filled in with the *empty* symbol, see Section 3. The <respStmt> element contains an obligatory <resp> element with the fixed string "Translated by" and a subsequent <name> element of *type* "translator" gives the name of the translator. If there is more than one translator, additional <name> elements are used.[16] If the translation has been carried out by a company or the like, the name of the company is given. The <name> elements may carry a *ref* attribute giving a reference to a corresponding <person> element in the <profileDesc> part of the header where additional info concerning the translator(s) may be given. This <name> element is of special relevance to texts which may be included in parallel corpora. More on translated texts can be found under the description of the <derivation> element in Section 2.3.3.

In the <monogr> part, the title of the collection is given if the text is part of a collection, otherwise it is left empty. The name of the editor is given in a <name> element as *surname, forename(s)*; if it is undeterminable how to decompose the name into forename(s) and surname, the comma is left out. If there are more than one editor, each of them is given in its own <editor> element. If there is no editor, the <name> element of <editor> carries an *empty* symbol, see Section 3. The <name> elements may carry a *ref* attribute giving a reference to a corresponding <person> element in the <profileDesc> part of the header where additional info

[14] It may seem odd that the *ref* attribute is given on the <name> element and not on the <author> element which would have been an option. However, as *ref* attributes also are used with translators and editors and neither the <respStmt> element used for translators nor the <editor> element are allowed to carry a *ref* attribute, it is instead attached to the <name> element in all these cases.

[15] The <note> element was added early 2015 in order to cope with extra author information in the Corpus of The Danish Dictionary in order to preserve it within the CTB header structure.

[16] It may seem inconsequent to repeat the <name> element for each translator whereas in case of the author and editor, the corresponding <author> and <editor> elements are repeated. However, as there obviously is no <translator> element in TEI, and as <respStmt> cannot carry a *type* attribute, repetition of the semantically rather empty <respStmt> element with its obligatory subordinate <resp> element (giving the semantics) seems much too awkward and would furthermore increase the complexity of queries.

concerning the editor(s) may be given.

In the <imprint> part of <monogr>, the name of the publishing house is given in the element <publHouse>,[17] the obligatory date of publishing as value of the *when* attribute of <date>, either the year or – in the case of newspapers – the year, month, and day according to the pattern *yyyy-mm-dd*. The *cert* attribute of <date> tells the certainty of the date which can either be "high" or "low". If the exact date is not known, an estimate is given and the *cert* attribute is set to "low". <imprint> includes five <biblScope> elements of different types which have to be filled in with the appropriate types of information, see Section 3. If a certain type of information does not apply to the publication described, it is left empty.

The <monogr> part of the structure is followed by an <idno> element of type "uri"[18] where a web pointer to the text can be given, i.e. the location from which it can be or has been downloaded. Other possible types are "isbn" and "issn". If it for some reason seems necessary to register the ISBN or ISSN, <idno> elements of the corresponding types can be added as well.

Another <idno> element of type "file" follows. As texts in most cases are delivered as electronic files, a back-reference to this source file is made by stating its filename and if necessary the path to it in this element. The file itself should be kept in an archive maintained by the organization which collected that particular text.[19] It may be necessary to leave out some information from material delivered, e.g. formatting, figures, tables, etc. In other cases, one single source file may contain a longer text that has to be chopped into smaller chunks. Being able to locate the source file ensures that certain completions or corrections can be made to the CTB file at a later point in time, if necessary.

2.2 The encoding description

The second major component of the TEI header is the encoding description <encodingDesc>. This contains information about the relationship between an encoded text and its original source.

The CTB <encodingDesc> element has the following sub-elements:

<samplingDecl> (sampling declaration) contains a description of the method used in sampling the text.

<projectDesc> (project description) describes the aim or purpose for which an electronic file was encoded.

<appInfo> (application information) records information about the applications which have processed the text of the TEI file.

[17] This element may be repeated if more publishers are to be listed.

[18] It might seem weird to place the URI of a text here. However, as there does not seem to be another adequate element to put this information, common practice obviously is to do it in this manner, see http://colab.mpdl.mpg.de/mediawiki/TEI_Bibliographic_Information.

[19] In the case of DK-CLARIN WP 2.1 all original texts are kept on the ja-korpus.dsl.lan server under /Volumes/Data/textrepository.

2.2.1 The sampling declaration

The <samplingDecl> element gives an indication of how the text was sampled, the indication is put in an <ab> element. The indication is a string from a fixed set. It must always be completely identical to the initial caption given in the <title> of <titleStmt>, see Section 2.1.1.

```
<samplingDecl>
  <ab>CTB version of:</ab>
</samplingDecl>
```

2.2.2 The project description

The <projectDesc> element gives an indication of the aim of collecting and encoding that particular text, i.e the corpus or text collection project or process:

```
<projectDesc>
  <ab>projectIdentifier</ab>
</projectDesc>
```

In the case of new texts captured by WP 2.1 of the DK-CLARIN project, the value of *projectIdentifier* is "DK-CLARIN-WP2.1". Similar fixed contents are defined for other relevant DK-CLARIN projects and for other finished projects like DDOC or KORPUS 2000, see Section 3.

2.2.3 Application information

The <appInfo> element gives information about all applications or other (manual) procedures by which the text sample has been enriched with markup. The header itself may also be manipulated by such applications or procedures, but this is not registered in the <appInfo> element – this may however be recorded under <revisionDesc>, see Section 2.4. The application information helps determining whether texts are structurally comparable, i.e. texts that have been processed by the same bundle of applications and procedures should be structurally identical.

The <appInfo> element should be filled in with one empty dummy-application if the file just contains the default-segmented (i.e. pre-tokenized) version of the text, the so-called *base version*, however the whole <appInfo> structure may be left out in this case as well.[20] The following example shows an <appInfo> with one empty dummy-application. The values given are explained further in Section 3.2.

```
<appInfo>
  <application xml:id="app_nil"
    type="nil"
```

[20] Leaving <appInfo> out is recommended by DK-CLARIN WP 5.

```
      subtype="nil"
      ident="nil"
      version="99999999"
      n="nil"
      when="99999999">
      <desc>nil</desc>
      <ptr target="#app_nil"/>
      <ref target="#opt_nil"/>
    </application>
  </appInfo>
```

Otherwise, there is one <appInfo> element for each *annotation layer* belonging to the text in the file, see Asmussen (2013b). The general structure is as follows:

```
<appInfo>
  <application xml:id="appXmlId"
    type="appType"
    subtype="appTool"
    ident="appId"
    version="appVersionNumber"
    n="appMode"
    when="appDate">
    <desc>appDesc</desc>
    <ptr target="#appXmlId"/> (may be left out)
    <ref target="#appOptionFile"/> (optional)
  </application>
</appInfo>
```

The <application> element has the following attributes:

xml:id unique XML identifier which is referenced by the corresponding annotation layer in the text.

type specifies both the task (segmentation, annotation) and whether it was performed by an automatic application or a manual procedure (or a combination of both).

subtype gives a further description of the applied tool taken from a fixed list of options.

ident supplies a unique identifier for the application/procedure.

version supplies a version number for the application/procedure. The version specification may contain other characters than digits, however it must

match the following regular expression :
`[\d]+[a-z]*[\d]*(\.[\d]+[a-z]*[\d]*){0,3}`.[21]

n gives supplementary info about the applied tag set or tokenization mode.

when gives the date when the application was executed on the text.

The `<application>` element contains an element `<desc>` giving a free-text description of the application.

The element `<ptr>` within `<application>` references that/those application/applications whose output has been used as input for the application in question as annotations can be added as layers on each other, cf. Asmussen (2013b). This element is left out if an annotation refers to the base version of the text and not to another annotation layer.

Finally, the optional `<ref>` element may reference certain resources a given tool has been using in cases where this is important.

2.3 The profile description

The third component of a TEI header is the profile description `<profileDesc>`. In the CTB, this is used to provide the following elements:

<creation> contains information about the creation of a text.

<langUsage> (language usage) describes the languages, sublanguages, registers, dialects etc. represented within a text.

<textDesc> (text description) provides a description of a text in terms of its situational parameters.

<textClass> (text classification) groups information which describes the nature or topic of a text in terms of a standard classification scheme, thesaurus, etc.

<particDesc> (participation description) describes the identifiable speakers, voices, or other participants in a linguistic interaction.

2.3.1 Text creation

The element `<creation>` is provided to record details of a text's creation, in the CTB header just the date it was *composed*, i.e. writing on it was finished; it should not be confused with the `<imprint>` element, where the date of the publication of the (source) text is recorded. In many cases the date, that is the year when a text was finished, is not known: in these cases the date is set to the same as under `<imprint>` and the value of the attribute *cert* is set to "low" instead of "high". Here is the patten:

[21] It may seem weird to apply version numbers to manual procedures. However, the *version* attribute is mandatory in TEI and also manual procedures may alter over time and should in any case be thoroughly documented – that is versioned.

```
<creation>
  <date when="textCreationYear" cert="certainty"/>
</creation>
```

2.3.2 Language usage

The <languageUsage> element contains the element <language> where the (dominant) language of the text is indicated by the attribute *ident*. Language codes are constructed as defined in BCP 47[22], the language notation standard to use should be ISO 639-1[23].[24] Particularly for sublanguages, an informal prose characterization should be supplied as content for the element. Language usage is expressed by the following XML pattern:

```
<langUsage>
  <language ident="languageId">
    languageCharacterization
  </language>
</langUsage>
```

2.3.3 Text description

The overall intention of using this part of the TEI proposal is to establish a structure that can contain text descriptions which can be applied to *every* potential corpus text. The structure is considered general and mandatory for every text in the CTB and information from this structure can be used to extract corpora from the CTB. Specialized textual information, which only may apply to *some* texts, is gathered in the <textClass> part of the header, see Section 2.3.4. Also, the amount of specialized textual information may vary from text to text.

The <textDesc> element characterizes each text according to the following eight situational parameters, each represented by one of the following eight elements:

<channel> (primary channel) describes the medium or channel by which a text is delivered or experienced. For a written text, this might be print, manuscript, e-mail, etc.; for a spoken one, radio, telephone, face-to-face, etc. The *mode* attribute describes the mode of the channel with respect to speech or writing.

[22] http://tools.ietf.org/html/bcp47

[23] http://www.sil.org/iso639-3/codes.asp. OBS! Select *View by 639-1*.

[24] At first glance, ISO 639-3 may seem a better choice as it provides more than 6900 language codes, also for dialects and historic languages. However, Danish seems only weakly represented in this standard. Danish authorities should probably get more involved in this standardization work. For DK-CLARIN purposes some of the private areas of this standard could be utilized. Maybe an issue for DK-CLARIN WP 1? Therefore, in the current headers, additional linguistic information may be given in a private BCP 47 extension with regional and historical tags (which needs to be defined).

<constitution> describes the internal composition of a text or text sample, for example as fragmentary, complete, etc.

<derivation> describes the nature and extent of originality of this text, that is, in the CTB header, just an indication of whether it has been translated from another language.

<domain> (domain of use) describes the most important social context in which the text was realized or for which it is intended, for example education, religion, business etc.

<factuality> describes the extent to which the text may be regarded as imaginative or non-imaginative, that is, as describing a fictional or a non-fictional world.

<interaction> describes the number of those producing and experiencing the text.

<preparedness> describes the extent to which a text may be regarded as prepared or spontaneous

<purpose> characterizes a single purpose or communicative function of the text, e.g. whether it is informative, expressive, etc.

By default, a text description will contain each of the above elements, supplied in the order specified. In the CTB, the <textDesc> pattern looks as follows:

```
<textDesc>
  <channel mode="tdChannelMode">tdChannel</channel>
  <constitution type="tdConstitutionType"/>
  <derivation type="tdDerivationType">
    <lang>languageId</lang>
  </derivation>
  <domain type="tdDomainDiscourse">tdDomain</domain>
  <factuality type="tdFactualityType"/>
  <interaction active="tdInteractActive"
    passive="tdInteractPassive">
    <note type="interactRole">tdInteractRole</note>
    <note type="interactAge">tdInteractAge</note>
  </interaction>
  <preparedness type="tdPrepType"/>
  <purpose type="tdPurposeType"/>
</textDesc>
```

Some of the elements given in the <textDesc> pattern contain further specified information:

The <derivation> element has a subordinate element <language> which indicates the original language of the text; if the text is not translated, the original language is identical to that indicated under <langUsage>, see Section 2.3.2.

The <interaction> element contains two subordinate <note> elements, one of them indicating the roles of the participants in the communication, that is, whether they are experts or laymen; the other <note> element gives the ages of addressor and addressee. Using a <note> element for giving further interaction-related information is not an optimal solution. A straighter way is to use special elements for the needed purposes or to augment the attribute list of the <interaction> element. However, this would require a modification of the TEI grammar.

More info on this part of the header can be found in Section 3.

2.3.4 Text classification

Texts may be described along many dimensions, according to many different taxonomies. No generally accepted consensus as to how such taxonomies should be defined has yet emerged. To accommodate special needs, TEI allows to express more specialized text characteristics by the following elements:

<catRef> (category reference) provides either a list of codes or one single code identifying the categories to which the text has been assigned, each code referencing a category element declared in the corpus header or under a separate, invariant URL. In CTB, there is one <catRef> element for each dimension, the type of dimension is indicated by the (referencing) value of the attribute *scheme*. CTB does not use lists of codes.

<classCode> contains the classification code used for the text in some standard classification system. There is one <classCode> element for each classification system.

Using <catRef> is the preferred way to give additional textual classifications in all cases where the classification system follows a CTB-internal standard. The pattern to be applied is as follows:

```
<textClass>
  <catRef scheme="myClassification" target="myValue"/>
</textClass>
```

The <catRef> element is repeated for each classification dimension used. If several values are given within the same classification dimension, <catRef> elements with the same classification scheme are repeated.

In cases where an official classification system is applied, the <classCode> element is used instead. More values within the same scheme are given by repeating <catRef> elements. The <catRef> and <classCode> elements should be used according to the following, invented, example:

```
<textClass>
  <catRef scheme="dk-clarin.eu/ctb/agerel" target="#a-c"/>
  <catRef scheme="dk-clarin.eu/ctb/domain" target="#med"/>
  <catRef scheme="dk-clarin.eu/ctb/domain" target="#bio"/>
  <catRef scheme="dk-clarin.eu/ctb/genre" target="#ad"/>
  <classCode scheme="official.classfication.eu">xyz</classCode>
</textClass>
```

2.3.5 The participant description

The participant description (<particDesc>) element is used to provide additional information about authors (or speakers) of texts. The element itself is considered obligatory in the CTB header, however, its contents may just be an empty <person> element which is given as a placeholder to ensure that the header has a valid TEI structure. If additional personal info is given, one <person> element for each participant having been involved in creating the text is inserted into <particDesc>.[25] The <person> element carries a number of attributes which are used to provide encoded values for some key aspects of the person concerned, see the following example:[26]

```
<particDesc>
  <person xml:id="personId"
    role="creatorRole"
    age="creatorAge"
    sex="creatorSex">
    <birth>
      <date when="creatorBirth" cert="certainty"/>
    </birth>
  </person>
</particDesc>
```

The DDOC material mentioned in Section 1 has a lot more information on each text creator, e.g. his place of birth which could be expressed as an element <placeName> under <birth>, his place of residence which could be put into an element <residence> as sibling to <birth>, and so on. However, corpus-linguistic practice has shown that this type of information hardly ever is used (nor useful if it is not given according to clear-cut classification schemes). Therefore, new material should not be marked-up with this kind of information that is also extremely costly to gather. For DDOC (and other material) which already carries this type of information, appropriate structural elements of <person> should be included into the header to allow keeping this information for possible future investigation, see Asmussen (2009).

[25] A possible empty placeholder <person> element may then be deleted.

[26] More details of which values to fill in can be found in Section 3.

2.4 The revision description

A list of typical revisions which a document will undergo should be created, i.e. values for *revisionType*. At least the revision type "Document created" seems important. Others, which deal with the completeness of the header may be useful as well. The pattern of the revision description is as follows:

```
<revisionDesc>
  <change when="revisionDate"
    who="organizationName">revisionType
  </change>
</revisionDesc>
```

The revision description must not be confused with the application information discussed in Section 2.2.3.

3 Filling in the header

3.1 Full header template

In the following, a complete version of the CTB header template is shown. Its four main constituents and their subdivisions are separated by horizontal lines to facilitate orientation:

```
<teiHeader type="text">
```

<fileDesc>

```
  <fileDesc>
```

<titleStmt>

```
    <titleStmt>
      <title>samplingDeclaration textTitle</title>
      <sponsor>sponsorName</sponsor>
      <respStmt>
        <resp>Data capture</resp>
        <name>organizationName
          <note type="method">captureMethod</note>
          <date when="captureYear"/>
        </name>
      </respStmt>
    </titleStmt>
```

<extent>

```
    <extent>
      <num n="words">numberOfWords</num>
      <num n="paragraphs">numberOfParagraphs</num>
    </extent>
```

<publicationStmt>

```
    <publicationStmt>
      <distributor>organizationName</distributor>
      <idno type="textIdType">textId</idno>
      <availability status="availStatus">
```

```
    <ab type="academic">
      <seg type="availDesc">availDesc</seg>
      <seg type="anonymDesc">anonymDesc</seg>
    </ab>
    <ab type="nonCommercial">
      <seg type="availDesc">availDesc</seg>
      <seg type="anonymDesc">anonymDesc</seg>
    </ab>
    <ab type="all">
      <seg type="availDesc">availDesc</seg>
      <seg type="anonymDesc">anonymDesc</seg>
    </ab>
  </availability>
</publicationStmt>
```

<notesStmt>

```
<notesStmt>
  <note xml:lang="languageId"
    resp="organizationName">note</note>
</notesStmt>
```

<sourceDesc>

```
<sourceDesc>
  <biblStruct>
    <analytic>
      <title xml:lang="languageId"
        level="titleLevel">textTitle</title>
      <author>
        <name ref="#personId">surname, forename</name>
        <note xml:lang="languageId"
          resp="organizationName">note</note>
      </author>
      <respStmt n="translators">
        <resp>Translated by</resp>
        <name ref="#personId">surname, forename</name>
      </respStmt>
    </analytic>
    <monogr>
      <title xml:lang="languageId">editionTitle</title>
      <editor>
        <name ref="#personId">surname, forename</name>
      </editor>
      <imprint>
        <publisher n="publId">publHouse</publisher>
        <date when="publDate" cert="certainty"/>
        <biblScope type="issue">edIssue</biblScope>
        <biblScope type="sect">edSect</biblScope>
        <biblScope type="vol">edVolume</biblScope>
        <biblScope type="chap">edChapter</biblScope>
        <biblScope type="pp">edPages</biblScope>
      </imprint>
    </monogr>
    <idno type="uri">textUri</idno>
    <idno type="file">textFileName</idno>
    <relatedItem type="relatedType ">
```

```
        <bibl>
          <title xml:lang="languageId">relatedTitle</title>
          <idno type="ctb">relatedId</idno>
        </bibl>
      </relatedItem>
    </biblStruct>
  </sourceDesc>
</fileDesc>
```

<encodingDesc>

```
<encodingDesc>
```

<samplingDecl>

```
  <samplingDecl>
    <ab>samplingDeclaration</ab>
  </samplingDecl>
```

<projectDesc>

```
  <projectDesc>
    <ab>projectIdentifier</ab>
  </projectDesc>
```

<appInfo>

```
  <appInfo>
    <application xml:id="appXmlId"
      type="appType"
      subtype="appTool"
      ident="appId"
      version="appVersion"
      n="appMode"
      when="appDate">
      <desc>appDesc</desc>
      <ptr target="#appXmlId"/>
      <ref target="#appOptionFile"/>
    </application>
  </appInfo>
</encodingDesc>
```

<profileDesc>

```
<profileDesc>
```

<creation>

```
  <creation>
    <date when="textCreationYear" cert="certainty"/>
  </creation>
```

<langUsage>

```
  <langUsage>
    <language ident="languageId">
      languageCharacterization
    </language>
  </langUsage>
```

<textDesc>

```
  <textDesc>
    <channel mode="tdChannelMode">tdChannel</channel>
    <constitution type="tdConstitutionType"/>
    <derivation type="tdDerivationType">
      <lang>languageId</lang>
    </derivation>
```

```
      <domain type="tdDomainDiscourse">tdDomain</domain>
      <factuality type="tdFactualityType"/>
      <interaction active="tdInteractActive"
        passive="tdInteractPassive">
        <note type="interactRole">tdInteractRole</note>
        <note type="interactAge">tdInteractAge</note>
      </interaction>
      <preparedness type="tdPrepType"/>
      <purpose type="tdPurposeType"/>
    </textDesc>
```

<textClass>

```
    <textClass>
      <catRef scheme="myClassification" target="myValue"/>
      <classCode scheme="theirClassification">theirValue</classCode>
    </textClass>
    <particDesc>
      <person xml:id="personId"
        role="creatorRole"
        age="creatorAge"
        sex="creatorSex">
        <birth>
          <date when="creatorBirth" cert="certainty"/>
        </birth>
      </person>
    </particDesc>
  </profileDesc>
```

<revisionDesc>

```
  <revisionDesc>
```

<change>

```
    <change when="revisionDate"
      who="organizationName">revisionType
    </change>
  </revisionDesc>
</teiHeader>
```

3.2 Value sets for header standard information

When filling in the header with standard information about the text, some types of information may be undetermined or non-existent, e.g. the name of an author may be simply missing in the header for some reason, that is, it is *undetermined*, or a text may not have a title, that is, its title is *non-existent*. Such incomplete parts of the header could be left out in these cases if permitted by TEI, however, leaving out such parts would obscure whether the information is missing because it is undetermined or because it is non-existent. If the information is undetermined, efforts should be undertaken to occasionally add it, otherwise, if it is non-existent, such efforts would be waste of time. In order to distinguish these two cases, it is recommended to always explicitly state non-existent information by filling in *empty* for string and symbol values, *0* (= zero) for integers, and *1000* in the case of years

(and dates),[27] in other words never to leave these parts of a header out. However, if the information is undetermined, these parts of a header may be left out indicating that the missing information occasionally should be added or be marked as non-existent if that is the case.

So in the case of undetermined information, it is legal to skip the respective part of the header if allowed by TEI; however, for the sake of completeness, it is strongly recommended to state *nil* in case of string values and *99999999*[28] in the case of integers and dates to indicate that this particular information obviously is missing and should be added if it does exist or, if it turns out that the information definitely does not exist, it should be marked as non-existent. To sum up, the following constant symbols are used as values for header elements and attributes, unless otherwise stated further below in this section:[29]

Symbol	Type	Meaning
empty	String	Info is non-existent
0	Integer	Info is non-existent
1000	Date/Year	Info is non-existent
nil	String	Info has not been determined yet
99999999	Integer and Date/Year	Info has not been determined yet

In all other cases, that is in cases where the desired information is available, the values listed in Section 3.2.1 are used replacing the header variables indicated in the full header template above. For each of these variables a description is given followed by an overview of its properties and – in the case of enumerated sets – a list of legal values. In cases where these lists are too comprehensive, they are replaced by a link to an XML version of them. All value sets are also accessible as XML files and may be referenced automatically or manually when filling in headers. All value set files are found under the path `http://korpus.dsl.dk/clarin/corpus-doc/text-header/`. The filenames themselves are given below.[30] The

[27] The value *1000* for dates is necessary in order to comply with the TEI data type `date` that does not allow a value of *0*.

[28] In former versions of the documentation the 'undetermined' value was *1* (minus one). However, TEI does not always allow a negative value for some of its integer datatypes which is the reason why it has been replaced.

[29] In cases where TEI does not allow the undetermined/non-existent values defined here, the elements of the value sets are restricted to those that are accepted by TEI. This is the case for the following attributes: `cert` in `<date>`, `sex` in `<person>`, `mode` in `<channel>`, `type` in `<factuality>`, `level` in `<title>`.

[30] As these are XML files, a web browser may not show them well formatted. Viewing them as HTML *source* may help though.

structure of the XML value set files is as shown in the following extract. The structure has been designed for this specific purpose (i.e. it is not TEI) and it should be fairly self-explanatory:

```
<?xml version="1.0" encoding="UTF-8"?>
<valuesetCollection
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:noNamespaceSchemaLocation=
        "http://korpus.dsl.dk/clarin/corpus-doc/
        text-header/valuesetCollection.xsd">
    <set name="captureMethod" type="symbol">
        <element>
            <value>nil</value>
            <desc>Info has not been determined yet</desc>
        </element>
        <element>
            <value>empty</value>
            <desc>Info is irrelevant, non-existent, or undeterminable</desc>
        </element>
        <element default="true">
            <value>file</value>
            <desc>The source of the text is an electronic file</desc>
        </element>
        <element>
            <value>ocr-raw</value>
            <desc>The text is OCR-scanned but not proof-read</desc>
        </element>
        <element>
            <value>ocr-proof</value>
            <desc>The text is OCR-scanned and proof-read</desc>
        </element>
        <element>
            <value>keyed-raw</value>
            <desc>The text is manually keyed but not proof-read</desc>
        </element>
        <element>
            <value>keyed-proof</value>
            <desc>The text is manually keyed and proof-read</desc>
        </element>
        [...]
    </set>
</valuesetCollection>
```

The following properties are given for each value set:

1. The ***value set type*** gives an indication of whether the set of values is meant to be augmented or not. It may be

   enumerated, closed, which means that no further values should be added to it

   enumerated, open, meaning that one can add further values if necessary

Open and *closed* is a distinction only relevant to enumerated, i.e. extensionally defined sets, whereas sets whose contents are intentionally defined, i.e. by description, as a matter of fact always are open:

descriptive can contain any description that observes the definition of the set

2. The ***XML URL*** is a URL that points to an XML version of the value set (only applicable for extensional value sets)

In some cases, properties are indicated as "undetermined" which means that this information still is missing for some reason and should be added in a future version of this document.

In other cases, properties are indicated as "n/a" as not applicable.

3.2.1 Alphabetical list of value sets

Note that some value sets are still empty as the properties they describe have not been relevant meta-info yet. Many others may still be augmented with additional values. Please refer to the most recent version of this document which can be downloaded as a technical report from http://korpus.dsl.dk/clarin/corpus-doc/text-header.pdf.

► ***anonymDesc***

Indicator specifying what type(s) of private text information must be made anonymous (= must not be shown).

Properties

Value set type	enumerated, closed
XML name	vs_anonymDesc.xml

Legal values

Value	Description
nil	Info has not been determined yet
empty	Info is irrelevant, non-existent, or undeterminable
0	Nothing in the text or associated with the text must be made anonymous**. Default**
I	Names of individuals must not be shown
P	Names of places must not be shown
A	Name(s) of the author(s) must not be shown
T	Text title must not be shown

The values can be combined if more of them apply to a specific user group, e.g. "IA" means that names of individuals and of the author(s) must be made anonymous.

- ***appDate***

  The date a particular markup application/procedure was applied to the text.

  Properties

Value set type	descriptive
XML name	n/a

  Legal values Dates must follow the pattern *yyyy-mm-dd*.

- ***appDesc***

  Free-text description of the application/procedure that has operated on the text.

  Properties

Value set type	descriptive
XML name	n/a

  Legal values Any string.

- ***appId***

  Unique version name-independent identifier of an application/procedure that has operated on the text.

  Properties

Value set type	enumerated, open
XML name	vs_appId.xml

Legal values

Value	Description
nil	Info has not been determined yet**. Default**
empty	Info is irrelevant, non-existent, or undeterminable
LocalInfoMediaConvert	Converts Infomedia text to CTB base format with simple headers
DoConvertK2000cqp2tx	Converts K2000 text to CTB base format with simple headers
DoSplitDDOC	Processes DDOC SGML-files by splitting them into CTB textfiles and mapping DDOC metadata to CTB
DoSplitBerling	Converts preprocessed Berling CD ROM files 1995-2000 into CTB textfiles and maps metadata to CTB

- ***appMode***

Info about the applied tag set, tokenization mode, or configuration.

Properties

Value set type	enumerated, open
XML name	vs_appMode.xml

Legal values

Value	Description
99999999	Info has not been determined yet
0	Info is irrelevant, non-existent, or undeterminable
da-001	Raw HHM Danish language model derived from Parole 2

- ***#appOptionFile***

XML pointer to information on the setup of the tool that has processed the text.

Properties

Value set type	descriptive
XML name	n/a

Legal values Any string that can be used for unique XML-referencing.

▶ ***appTool***

Describes the (automatic or manual) tool that has operated on the text.

Properties

Value set type	enumerated, closed
XML name	vs_appTool.xml

Legal values

Value	Description
nil	Info has not been determined yet**. Default**
empty	Info is irrelevant, non-existent, or undeterminable
pretokenizer	Splits a text into word-like segments. A pretokenizer is only applied once, all other applications are based on the pretokenized version of the text
tokenizer	Splits a text into word-like segments
s-splitter	Sentence splitter. Splits the text into sentences, i.e. a segment between two full stops or some similar type of punctuation. Inserts `<s>` and `</s>` tags around sentence-like text segments
p-splitter	Paragraph splitter. Splits the text into paragraphs. Inserts `<p>` and `</p>` tags around paragraph-like text segments
regularizer	Tags a token with a regularised version of its surface representation, i.e. its orthography
lemmatizer	Tags a token with its lemma form
pos-tagger	Tags a token with part-of-speech info
morph-tagger	Tags a token with morphological/inflectional info
term-tagger	Tags a token with some indication of whether it is a term (in texts to be included in LSP corpora)
multi-processor	Multifunctional tool that performs various tasks like tokenizing, lemmatizing, tagging as one complex process
other	Tool performing tasks not yet listed

- ***appType***

  Specifies whether an application or procedure that operated on the text was automatic (or a combination of both) as well as the type of task of the application/procedure in terms of segmentation or annotation.

  Properties

Value set type	enumerated, closed
XML name	vs_appType.xml

  Legal values

Value	Description
nil	Info has not been determined yet
empty	Info is irrelevant, non-existent, or undeterminable
a_segmentation	Text split into smaller segments (e.g. sentences, tokens) by an automatic process**. Default**
c_segmentation	Text split into smaller segments (e.g. sentences, tokens) by a combined automatic-manual process
m_segmentation	Text split into smaller segments (e.g. sentences, tokens) by a manual process
a_annotation	Text segments annotated with info by an automatic process
c_annotation	Text segments annotated with info by a combined automatic-manual process
m_annotation	Text segments annotated with info by a manual process

- ***appVersion***

  Version specification of an application/procedure that has operated on the text.

  Properties

Value set type	descriptive
XML name	n/a

  Legal values The version specification must start with at least one digit but may contain other characters than digits. It must match the following regular expression:
  `[\d]+[a-z]*[\d]*(\.[\d]+[a-z]*[\d]*){0,3}`.

- ***appXmlId***

  Unique XML identifier which is referenced by the corresponding annotation layer (`<spanGrp>` element, see Asmussen (2013b)) in the text.

  Properties

Value set type	descriptive
XML name	n/a

  Legal values Valid XML IDs are constructed by concatenating the *appId*, an underscore, and the *appVersion* where dots are replaced by underscores.

- ***availDesc***

  Tells how this text may be used in terms of copyright and other restrictions.

  Properties

Value set type	enumerated, closed
XML name	vs_availDesc.xml

  Legal values

Value	Description
nil	Info has not been determined yet
empty	Info is irrelevant, non-existent, or undeterminable
full	The user has free access to the complete material, but is not allowed to redistribute it
partial	The user can search and view text contents limited to what is specified in Danish citation law**. Default**
limited	Access only upon written agreement between the DK-CLARIN consortium and the user. Details of this agreement are to be further specified
none	No acces for users not affiliated with the DK-CLARIN consortium

- ***availStatus***

  Attribute of the `<availability>` element indicating whether the text is freely available for all user categories (cf. the header template above) or not.

Properties

Value set type	enumerated, closed
XML name	vs_availStatus.xml

Legal values

Value	Description
nil	Info has not been determined yet
empty	Info is irrelevant, non-existent, or undeterminable
free	The text is freely available for all user categories
restricted	The text is not freely available for at least one user category**. Default**
DSL only until YYYY	Access for The Danish Dictionary at DSL only until the year specified

- ***captureMethod***

  The method of data capture.

Properties

Value set type	enumerated, closed
XML name	vs_captureMethod.xml

Legal values

Value	Description
nil	Info has not been determined yet
empty	Info is irrelevant, non-existent, or undeterminable
file	The source of the text is an electronic file**. Default**
file-manually	The source of the text is an electronic file that has been edited or processed manually
corpus	The source of the text is an existing corpus
ocr-raw	The text is OCR-scanned but not proof-read
ocr-proof	The text is OCR-scanned and proof-read
keyed-raw	The text is manually keyed but not proof-read
keyed-proof	The text is manually keyed and proof-read
double-keyed	The text is double-keyed, i.e. keyed in two versions by two individual typists, both versions are automatically compared and manually corrected
pdf-converted-acrobat9	Converted from PDF by Acrobat 9
pdf-converted-pdf2xml	Converted from PDF by pdf2xml

- ***captureYear***

The year of data capture. In cases where *captureMethod* is *corpus*, the captureYear may be set to the year of the original corpus creation.

Properties

Value set type	descriptive
XML name	n/a

Legal values Four-digit years which may be extended to full dates following the pattern *yyyy-mm-dd*.

- ***certainty***

The degree of certainty of how precise some data, typically dates, are.

Properties

Value set type	enumerated, closed
XML name	vs_certainty.xml

Legal values

Value	Description
empty	Info is irrelevant, non-existent, or undeterminable. **Default**
high	The given dates are definitely correct
low	The given dates are an estimate

▶ ***creatorAge***

The age group to which a particular author belonged at the time he/she produced the text.

Properties

Value set type	enumerated, closed
XML name	vs_creatorAge.xml

Legal values The age intervals are inevitably arbitrary. The "teen" interval is consciously extended to the age of 25 to be able to better indicate young people's language in general. See also TEI P5.[31]

Value	Description
nil	Info has not been determined yet
empty	Info is irrelevant, non-existent, or undeterminable
infant	A person aged 0–5
child	A person aged 6–12
teen	A person aged 13–25
adult	A person aged 26–60. **Default**
senior	A person aged 61 and above

▶ ***creatorBirth***

The year a particular author was born.

[31] http://www.tei-c.org/release/doc/tei-p5-doc/html/ref-person.html

Properties

Value set type	descriptive
XML name	n/a

Legal values Four-digit date following the pattern *yyyy*.

▶ ***creatorRole***

The role of a particular author in terms of his or her influence on the language of the text.

Properties

Value set type	enumerated, closed
XML name	vs_creatorRole.xml

Legal values For written texts:[32]

Value	Description
nil	Info has not been determined yet
empty	Info is irrelevant, non-existent, or undeterminable
major	Assigned to one single autor, translator, or editor who is assumed to have had major impact on the language of the text**. Default**
minor	Assigned to all other textual contributors

There should only be one author, translator, or editor with “major” influence. All other contributors should be classified “minor”.

▶ ***creatorSex***

The sex of a particular author.

Properties

Value set type	enumerated, closed
XML name	vs_creatorSex.xml

[32] The list may be augmented with values for spoken texts from the DDOC.

Legal values From ISO 5218:1977 Representation of Human Sexes to comply with TEI, see http://www.tei-c.org/release/doc/tei-p5-doc/html/ref-data.sex.html. OBS! The values for *undetermined* ("nil") and *n/a* ("empty") differ from the CTB standard values.

Value	Description
0	Unknown. **Default**
1	Male
2	Female
9	Not applicable

▶ ***edChapter***

The chapter of a book or similar edition from which the text sample is taken.

Properties

Value set type	descriptive
XML name	n/a

Legal values Any integer.

▶ ***edIssue***

The issue of a newspaper or journal from which the text sample is taken.

Properties

Value set type	descriptive
XML name	n/a

Legal values Any string.

▶ ***edPages***

The range of pages the text sample spans over in the edition from which it is taken.

Properties

Value set type	descriptive
XML name	n/a

Legal values Any integer or an interval of integers according to the pattern: *x*–*y* where *y* > *x*. Groups of intervals are not allowed. Each text sample in the CTB must be coherent. If several samples are taken from the same text source, each of them has to be put into a CTB file of its own.

- ***edSection***

  The section of a newspaper from which the sample is taken.

  Properties

Value set type	descriptive
XML name	n/a

  Legal values Any string.

- ***edVolume***

  The volume of a book from which the text sample is taken.

  Properties

Value set type	descriptive
XML name	n/a

  Legal values Any integer.

- ***editionTitle***

  The title of the edition (e.g. book, newspaper) in which the text appeared.

  Properties

Value set type	descriptive
XML name	n/a

  Legal values Any string.

- ***fileCreationYear***

  The year the electronic text sample was created.

  Properties

Value set type	descriptive
XML name	n/a

Legal values Four-digit date which may be extended to a full date following the pattern *yyyy-mm-dd*.

▶ ***forename***

First name(s) of a text's author/editor/translator.

Properties

Value set type	descriptive
XML name	n/a

Legal values Any string. Names are always given as a string of pattern *surname, forename* in `<name>` elements. If the name cannot be decomposed into forename and surname, the name is stated without a comma. If the text has been written/translated/edited by a company or organization, the name of that company/organization is stated. If the name for some reason must be kept anonymous, the `<name>` element is filled in with the string "anonymous".

▶ ***languageCharacterization***

Prose description of the language indicated by *languageId*.

Properties

Value set type	descriptive
XML name	n/a

Legal values Comma-separated list of the descriptions associated with the values applied in *languageId*, e.g. "Danish" if *languageId* is "da". See *languageId*.

▶ ***languageId***

Code that identifies the language used in the text sample or in a `<note>` or `<title>` tag.

Properties

Value set type	enumerated, open
XML name	vs_langSubId.xml

Legal values Values follow BCP 47[33] and ISO 639-1.[34] The language code is constructed according to BCP 47 as follows:

langSubId [- x [- *langSubHist*] [- *langSubRegion*]]

It consists of an obligatory part with a language code *langSubId* according to ISO 639-1[35] and an optional private extension, prefixed by the BCP 47 subtag *x*[36] that holds a code *langSubHist* for the historic period of the language in question, and another optional part with a regional code *langSubRegion*. If both optional parts are present, they must come in the order specified.

Legal values for *langSubId* are defined in the following subset of the ISO 639-1 standard, however the non-standard value "xx" has been added to indicate formal or constructed language that may occur in the content of `<note>` elements.

Value	Description
nil	Info has not been determined yet (not part of ISO 639-1). **Default**
empty	Info is irrelevant, non-existent, or undeterminable (not part of ISO 639-1)
da	Danish
de	German
en	English
es	Spanish
fr	French
xx	Formal or constructed (not part of ISO 639-1)

For each *langSubId*, that is for each language, a set of *langSubHist* and *langSubRegion* codes can be defined; for each language the name of the *langSubHist* and *langSubRegion* variables is extended with the ISO 639-1 code of the language in question, e.g. *langSubHistDa* or *langSubRegionDa* for Danish. Legal values must be defined according to the pattern "hCode" for historic codes and "rCode" for region codes, the "h" and the "r" indicating *historic* and *region* respectively, whereas the "Code" part contains the code to be used for a certain period or region. Currently, there are no such "hCode" codes are defined for any language within the CTB framework, however, the following "rCode" codes are defined for Danish.[37]

[33] `http://tools.ietf.org/html/bcp47`

[34] `http://www.sil.org/iso639-3/codes.asp`. OBS! Select *View by 639-1*.

[35] A list is available at www.loc.gov.

[36] A quick introduction on the standard and on using private x-extensions of the tag can be found at w3.org.

[37] They are only used in the Corpus of the Danish Dictionary, DDOC.

langSubHist..

Properties

Value set type	enumerated, open
XML name	langSubHist

Legal values Currently, no values defined.

langSubRegionDa

Properties

Value set type	enumerated, open
XML name	vs_langSubRegion.xml

Legal values The following values are defined (Danish only):

Value	Description
nil	Info has not been determined yet
empty	Info is irrelevant, non-existent, or undeterminable
rStd	Standard (rigssprog)**. Default**
rReg	Regional (regionalsprog)

▶ ***myClassification***

URL of a user-defined text classification.

Properties

Value set type	enumerated, open
XML name	vs_myClassification.xml

Legal values Any valid URL pointing to a classification scheme. Currently, the following classification scheme URLs are defined:

Value	Description
nil	Info has not been determined yet
empty	Info is irrelevant, non-existent, or undeterminable
http://ctb.dsl.dk/class/catRef/DDOC/RePr.xml	Synsvinkel (produktion, reception)
http://ctb.dsl.dk/class/catRef/DDOC/Medi.xml	Medium, channel
http://ctb.dsl.dk/class/catRef/DDOC/Genr.xml	Genre, text type
http://ctb.dsl.dk/class/catRef/DDOC/GnTy.xml	Genre type (simplified genre classification)
http://ctb.dsl.dk/class/catRef/infomedia/PSIN.xml	Infomedia PSIN topic labels

► ***myValue***

Value given in a user-defined text classification.

Properties

Value set type	enumerated, open
XML name	n/a

Legal values Legal values according to the user-defined classification.

► ***note***

Any note giving additional information about the parent element which cannot be expressed by other elements in the header.

Properties

Value set type	descriptive
XML name	n/a

Legal values Any string.

► ***numberOfParagraphs***

The number of paragraphs in the text.

Properties

Value set type	descriptive
XML name	n/a

Legal values Any integer.

▶ ***numberOfWords***

The number of word-like units, i.e. `<w>` elements, in the text.

Properties

Value set type	descriptive
XML name	n/a

Legal values Any integer.

▶ ***organizationName***

The name of (a person at) an organization who carried out some particular piece of work or had some particular responsibility related to the electronic text sample.

Properties

Value set type	enumerated, open
XML name	vs_organizationName.xml

Legal values

Value	Description
nil	Info has not been determined yet**. Default**
empty	Info is irrelevant, non-existent, or undeterminable
cst.ku.dk	Center for Sprogteknologi, KU
dsl.dk	Det Danske Sprog- og Litteraturselskab
ja@dsl.dk	Jørg Asmussen hos Det Danske Sprog- og Litteraturselskab
dsn.dk	Dansk Sprognævn
dsl-dsn.dk	DSL og DSN i fællesskab
duds.nordisk.ku.dk	Digitale Undersøgelser af Dansk Sprog, INSS, KU

- ***#personId***

  Id linking between the name of an author and the `<person>` element in `<textDesc>` giving additional author information.

  Properties

Value set type	descriptive
XML name	n/a

  Legal values Any string that can be used for unique XML-referencing. The string should contain a sequence of digits.

- ***publDate***

  The publishing date of the edition in which the text appeared.

  Properties

Value set type	descriptive
XML name	n/a

  Legal values Values are given either as the year as a four-digit number, or the year, month, and day given according to the pattern *yyyy-mm-dd*.

- ***publHouse***

  The name of the publisher (company, or if self-published, the author) of the edition in which the text appeared, or the name of the text supplier.

  Properties

Value set type	enumerated, open
XML names	vs_publId.xml

  Legal values String denoting a publisher/supplier taken from the description part of the lists referred to under *publId* below.

- ***publId***

  Unique identifier of either publisher or text supplier pointing to an external database of publishers.

  Properties

Value set type	enumerated, open
XML names	vs_publId.xml

Legal values Integer according to specified lists maintained by WP 2.1.

Additional publisher/supplier info is found in the resource

- `/db/ctb/suppliers/ctb-suppliers.xml`

in the eXist-db on the `ja-korpus.dsl.lan` server. The *publId*s given in the list above can be seen as pointers to the records with additional supplier info.

► ***projectIdentifier***

Unique identifier of the text collection project in which this electronic text was captured and prepared.

Properties

Value set type	enumerated, open
XML name	vs_projectIdentifier.xml

Legal values

Value	Description
nil	Info has not been determined yet**. Default**
empty	Info is irrelevant, non-existent, or undeterminable
DK-CLARIN-WP2.1	LGP corpus project under DK-CLARIN, 2008-2010
DK-CLARIN-WP2.2	LSP corpus project under DK-CLARIN, 2008-2010
DK-CLARIN-WP2.3	Renaissance corpus project under DK-CLARIN, 2008-2010
DK-CLARIN-WP2.4	JVJ/ADL corpus project under DK-CLARIN, 2008-2010
DK-CLARIN-WP2.5	Nationalmuseet's corpus project under DK-CLARIN, 2008-2010
DK-CLARIN-WP2.6	Parallel corpus project under DK-CLARIN, 2008-2010
DSL-DOT	Ongoing DSL-DOT gathering
DSL-DOT-IM	Ongoing DSL-DOT gathering via InfoMedia
DDOC-spoken	Corpus of The Danish Dictionary, transcribed speech
DDOC-written	Corpus of The Danish Dictionary, written
K2000	Material collected in the Korpus 2000 project
DDO	Material collected in The Danish Dictionary project

► ***relatedTitle***

Title of a text related to the current one.

Properties

Value set type	descriptive
XML name	n/a

Legal values Any string denoting a text title.

► ***relatedType***

Value stating how the text possibly is related to another text.

Properties

Value set type	enumerated, closed
XML name	

Legal values

Value	Description
nil	Info has not been determined yet**. Default**
empty	Info is irrelevant, non-existent, or undeterminable
noRelated	No related text exists
original	The related text is the original from which the current text has been translated
parallel	It is not known whether the related text is the original or the translation, as may be the case for texts from the EU

- ***revisionDate***

  Date when a revision was performed on the text item.

  Properties

Value set type	descriptive
XML name	n/a

  Legal values Year, month, and day given according to the pattern *yyyy-mm-dd*.

- ***revisionType***

  Standardized type of revision applied to the text item.

  Properties

Value set type	enumerated, open
XML name	vs_revisionType.xml

  Legal values

Value	Description
nil	Info has not been determined yet
empty	Info is irrelevant, non-existent, or undeterminable
created	First version of CTB file created**. Default**

► ***samplingDeclaration***

Indicates the amount of original text included in the CTB version.

Properties

Value set type	enumerated, closed
XML name	vs_samplingDeclaration.xml

Legal values

Value	Description
nil	Info has not been determined yet
empty	Info is irrelevant, non-existent, or undeterminable
CTB sample	It is unknown whether the text is complete or abridged**. Default**
CTB version	Complete text is included
ctbTextUnit version	DEPRECATED: Use "CTB version" instead
CTB excerpt	Continuous excerpt from the original text

► ***sponsorName***

The name of the initiative (or organization) that intellectually has supported or initiated the collection of a particular text.

Properties

Value set type	enumerated, open
XML name	vs_sponsorName.xml

Legal values

Value	Description
nil	Info has not been determined yet
empty	Info is irrelevant, non-existent, or undeterminable
DK-CLARIN	The DK-CLARIN Consortium, 2008-2010**. Default**
ordnet.dk	The Ordnet.dk Project at dsl.dk, 2006-2013
Korpus 2000	The Korpus 2000 Project at dsl.dk, 2000-2002
DDO	Den Danske Ordbog at dsl.dk, 1991-2005

▶ ***surname***

Last name of a text's author/editor/translator.

Properties

Value set type	descriptive
XML name	n/a

Legal values Names are always given as a string of pattern *surname, forename* in `<name>` elements. If the name cannot be decomposed into forename and surname, the name is stated without a comma. If the text has been written/translated/edited by a company or organization, the name of that company/organization is stated. If it for some reason is anonymous, the `<name>` element is filled in with the string "anonymous".

▶ ***tdChannel***

The primary channel/medium by which a text is delivered or experienced.

Properties

Value set type	enumerated, open
XML name	vs_tdChannel.xml

Legal values Generally, a text can either be written or spoken. If it is written, it can either be distributed electronically, e.g. on the Internet, or on paper, e.g. as a book. The following table is only rudimentary, but shows the principle of coding: The first digit from the left indicates the general channel which can be further specified by adding further digits, e.g. "2" means written, "22" means written using an electronic channel, "221" might mean email, etc.

Value	Description
99999999	Info has not been determined yet. **Default**
0	Unknown channel
1	Spoken
121	Radio
122	TV
123	Movie
124	Audio recording
125	Speaker
126	Speech
127	Theatre
128	Telephone
129	Video recording
2	Written
21	Paper
211	Magazine
212	Book
213	Newspaper
214	Local paper
215	Labour paper
216	Ephemeron
217	Journal
22	Electronic

- ***tdChannelMode***

  Describes the channel/medium of a text with respect to speech or writing.

  Properties

Value set type	enumerated, closed
XML name	vs_tdChannelMode.xml

  Legal values Values follow the TEI specifications:[38]

[38] http://www.tei-c.org/release/doc/tei-p5-doc/html/ref-channel.html

Value	Description
w	Written**. Default**
s	Spoken
sw	Spoken recorded by writing it down
ws	Written meant to be spoken
m	Mixed
x	Unknown or inapplicable. OBS! TEI mixes two cases which usually are kept apart in CTB

▶ ***tdConstitutionType***

Describes the internal composition of a text or text sample, for example as fragmentary or complete.

Properties

Value set type	enumerated, closed
XML name	vs_tdConstitutionType.xml

Legal values Legal values make up a subset of the TEI specifications:[39]

Value	Description
nil	Info has not been determined yet
empty	Info is irrelevant, non-existent, or undeterminable
single	A single complete text**. Default**
frags	The text is a continuous fragment, e.g. a chapter from a novel
unknown	It is unknown whether the text is complete or fragmentary

▶ ***tdDerivationType***

Describes whether the text is translated or original.

Properties

Value set type	enumerated, closed
XML name	vs_tdDerivationType.xml

[39] http://www.tei-c.org/release/doc/tei-p5-doc/html/ref-constitution.html

Legal values Legal values follow the TEI specifications:[40]

Value	Description
nil	Info has not been determined yet
empty	Info is irrelevant, non-existent, or undeterminable
original	Original, un-translated version of the text. **Default**
translation	The text is a translation

▶ ***tdDomain***

The domain the text is associated with.

Properties

Value set type	enumerated, closed
XML name	vs_tdDomain.xml

Legal values The full set of 66 DDOC domain values is used, as experiments using it for automatic domain classification were promising, see Asmussen (2005).[41] The 66 values can be looked up in the following XML document: DDOC domain values.

▶ ***tdDomainDiscourse***

Describes whether the discourse is domain-specific or not, i.e. if the type of language used in the text can be categorized as language for general or specific purposes.

Properties

Value set type	enumerated, closed
XML name	vs_tdDomainDiscourse.xml

Legal values

Value	Description
nil	Info has not been determined yet
empty	Info is irrelevant, non-existent, or undeterminable
general	No domain-specific discourse. Language for general purposes used. **Default**
specific	Domain-specific discourse. Language for specific purposes used

[40] http://www.tei-c.org/release/doc/tei-p5-doc/html/ref-derivation.html
[41] http://korpus.dsl.dk/staff/ja/papers/cl2005_asmussen.latex.pdf

▶ ***tdFactualityType***

Tells whether a text is imaginative or non-imaginative.

Properties

Value set type	enumerated, closed
XML name	vs_tdFactualityType.xml

Legal values Values must conform with the TEI specifications[42] given in the following list:[43]

Value	Description
fiction	The text is to be regarded as entirely imaginative
fact	The text is to be regarded as entirely informative or factual
mixed	The text contains a mixture of fact and fiction
inapplicable	The fiction/fact distinction is not regarded as helpful or appropriate to this text**. Default**

▶ ***tdInteractActive***

The number of addressors having produced the text.

Properties

Value set type	enumerated, closed
XML name	vs_tdInteractActive.xml

Legal values Values conform to the suggestions made in the TEI specifications.[44]

Value	Description
nil	Info has not been determined yet
empty	Info is irrelevant, non-existent, or undeterminable
singular	A single addressor**. Default**
plural	Many addressors
corporate	A corporate addressor

[42] http://www.tei-c.org/release/doc/tei-p5-doc/html/ref-factuality.html
[43] TEI does not allow to distinguish between “unknown” and “inapplicable”.
[44] http://www.tei-c.org/release/doc/tei-p5-doc/html/ref-interaction.html

- ***tdInteractAge***

  The age group to which addressor and addressee belong.

Properties

Value set type	enumerated, closed
XML name	vs_tdInteractAge.xml

Legal values

Value	Description
nil	Info has not been determined yet
empty	Info is irrelevant, non-existent, or undeterminable
infant-infant	A person aged 0–5 addressing another infant
infant-child	A person aged 0–5 addressing a child
infant-teen	A person aged 0–5 addressing a teen
infant-adult	A person aged 0–5 addressing an adult
infant-senior	A person aged 0–5 addressing a senior
child-infant	A person aged 6–12 addressing an infant
child-child	A person aged 6–12 addressing another child
child-teen	A person aged 6–12 addressing a teen
child-adult	A person aged 6–12 addressing an adult
child-senior	A person aged 6–12 addressing a senior
teen-infant	A person aged 13–25 addressing an infant
teen-child	A person aged 13–25 addressing a child
teen-teen	A person aged 13–25 addressing another teen
teen-adult	A person aged 13–25 addressing an adult
teen-senior	A person aged 13–25 addressing a senior
adult-infant	A person aged 26–60 addressing an infant
adult-child	A person aged 26–60 addressing a child
adult-teen	A person aged 26–60 addressing a teen
adult-adult	A person aged 26–60 addressing another adult**. Default**
adult-senior	A person aged 26–60 addressing senior
senior-infant	A person aged 61 and above addressing an infant
senior-child	A person aged 61 and above addressing a child
senior-teen	A person aged 61 and above addressing a teen
senior-adult	A person aged 61 and above addressing an adult
senior-senior	A person aged 61 and above addressing another senior

▶ ***tdInteractPassive***

The number of addressees to whom a text is directed.

Properties

Value set type	enumerated, closed
XML name	vs_tdInteractPassive.xml

Legal values Values are taken from the TEI suggestions.[45]

Value	Description
nil	Info has not been determined yet
empty	Info is irrelevant, non-existent, or undeterminable
self	Text is addressed to the originator e.g. a diary
single	Text is addressed to one other person e.g. a personal letter
many	Text is addressed to a countable number of others e.g. a conversation in which all participants are identified
group	Text is addressed to an undefined but fixed number of participants e.g. a lecture
world	Text is addressed to an undefined and indeterminately large number e.g. a published book**. Default**

▶ ***tdInteractRole***

Describes the roles of addressor and addressee in terms of technical expertise concerning the topic of the text. This information is usually only interesting if *tdDomain* has a value other than its default. Otherwise *tdInteractRole* will default to “basic-basic”.

Properties

Value set type	enumerated, closed
XML name	vs_tdInteractRole.xml

[45] http://www.tei-c.org/release/doc/tei-p5-doc/html/ref-interaction.html

Legal values

Value	Description
nil	Info has not been determined yet
empty	Info is irrelevant, non-existent, or undeterminable
basic-basic	A person with basic knowledge of the topic, i.e. a layperson, addresses another person with basic knowledge. **Default**
basic-advanced	Somebody with basic knowledge addressing somebody with advanced knowledge
basic-expert	Somebody with basic knowledge addressing somebody with expert knowledge
advanced-basic	Advanced addressing basic
advanced-advanced	Advanced addressing advanced
advanced-expert	Advanced addressing expert
expert-basic	Expert addressing basic
expert-advanced	Expert addressing advanced
expert-expert	Expert addressing expert

- ***tdPrepType***

  Describes the extent to which a text may be regarded as prepared or spontaneous.

Properties

Value set type	enumerated, closed
XML name	vs_tdPrepType.xml

Legal values A subset from the TEI suggestion:[46]

Value	Description
nil	Info has not been determined yet
empty	Info is irrelevant, non-existent, or undeterminable
none	The text is spontaneous or unprepared
revised	Polished or revised before presentation. **Default**

[46] http://www.tei-c.org/release/doc/tei-p5-doc/html/ref-preparedness.html

► ***tdPurposeType***

Characterizes a single purpose or communicative function of the text, e.g. whether it is informative, expressive, etc.

Properties

Value set type	enumerated, closed
XML name	vs_tdPurposeType.xml

Legal values Following the TEI suggestions:[47]

Value	Description
nil	Info has not been determined yet
empty	Info is irrelevant, non-existent, or undeterminable
persuade	Didactic, advertising, propaganda, etc.
express	Self expression, confessional, etc.
inform	Convey information, educate, etc.. **Default**
entertain	Amuse, entertain, etc.

► ***textCreationYear***

The year in which the text was authored.

Properties

Value set type	descriptive
XML name	n/a

Legal values Four-digit date. If the year of text creation is not known, *textCreationYear* is set to the same value as *publDate*.

► ***textFileName***

Name of the source file from which this text is drawn, that is usually the name of the file the text was delivered in. The organization having collected the text is responsible for keeping a copy of its source file in an archive if it wants to enable future corrections or modifications of the CTB version of the text with regard to certain information only contained in the source file.

[47] http://www.tei-c.org/release/doc/tei-p5-doc/html/ref-purpose.html

Properties

Value set type	descriptive
XML name	n/a

Legal values Any legal (path and) filename pointing to the source file in the archive.

- ***textId***

  Unique text identifier.

Properties

Value set type	system: descriptive
	prefixes listed below: enumerated, open
XML name	system: n/a
	prefixes: vs_textId.xml

Legal values Values for *textId* of *textIdType* "ctb" (cf. below): Specified 10-digit integer. Identifiers of this type are composed as follows: The first two digits (from the left) indicate the project framework within which the texts were collected (which can be some other than DK-CLARIN). Thus, the first two digits can be viewed as a kind of prefix. The following set of prefixes of *textIdType* "ctb" is used:

Value	Description
99999999	Info has not been determined yet
0	Info is irrelevant, non-existent, or undeterminable
10	Korpus 2000 material from 'Politiken', 'Jyllands-Posten' and 'fyldepennen.dk'
11	Other Korpus 2000 material
120	PAROLE (OBS! PAROLE comprises some material from DDOC)
121	Material from the Corpus of The Danish Dictionary (DDOC)
122	Berling CD-ROM material 1995-2000
13	Material collected by DSL's ordnet.dk project
139	Manually prepared material collected by DSL's ordnet.dk project
14	Infomedia material collected by DSL's ordnet.dk project
20	Infomedia material collected by DK-CLARIN WP2.1, LGP Corpus
2009	Infomedia magazines 2010-11 collected by DK-CLARIN WP2.1, LGP Corpus
21	Material collected by DK-CLARIN WP2.1, LGP Corpus
22	Material collected by DK-CLARIN WP2.2, LSP Corpus
23	Material collected by DK-CLARIN WP2.3, Renaissance Corpus
24	Material collected by DK-CLARIN WP2.4, ADL/JVJ
25	Material collected by DK-CLARIN WP2.5, Nationalmuseet
26	Material collected by DK-CLARIN WP2.6, Parallel Corpus
8	sdewac - German Web Corpus
90000	DiaKo - optegnelser af dialekter, NFI/ØMO

However, depending on the actual id system (see *textIdType* below), strings are acceptable as well.

- ***textIdType***

  Identifies the type of *textId* given.

Properties

Value set type	enumerated, open
XML name	vs_textIdType.xml

Legal values Default type is “ctb”, but other project- or institution-internal types can be added.

Value	Description
nil	Info has not been determined yet
empty	Info is irrelevant, non-existent, or undeterminable
ctb	Text id according to the id system specified for the Clarin Text Bank. **Default**
ddo	Text id according to the id system specified for the Corpus of The Danish Dictionary
berling	Text id according to the id system in the Berlingske Corpus, 1995-2000
k2000	Text id according to the id system specified for Korpus 2000
dsst	Text id according to the id system of Dansk Sprog- og Stilhistorisk Tekstbase (WP2.3)
im	Text id according to the id system used by Infomedia (WP2.1)
wiki	Wikipedia ID found in Wikipedia export documents at /mediawiki/page/id/text()
extUri	External URI/URL of the text resource

- ***textTitle***

Title of the text from which the sample is taken.

Properties

Value set type	descriptive
XML name	n/a

Legal values Any string denoting a text title.

- ***textUri***

Resource identifier locating the text source.

Properties

Value set type	descriptive
XML name	n/a

Legal values Any valid URI pointing at a source instance of the text.

- ***theirClassification***

URL of an official text classification scheme.

Properties

Value set type	enumerated, open
XML name	vs_theirClassification.xml

Legal values Any valid URL pointing to a classification scheme. Currently, the following official classification scheme URLs are defined:

Value	Description
nil	Info has not been determined yet
empty	Info is irrelevant, non-existent, or undeterminable
http://ctb.dsl.dk/class/classCode/CLARIN/demo.xml	Classification containing some demo values

- ***theirValue***

Value given in an official text classification system.

Properties

Value set type	n/a
XML name	n/a

Legal values Legal values according to official classification.

- ***titleLevel***

Indicates the level of the title within a publication, whether the title is on analytic level, i.e. the text is part of a collection (e.g. a newspaper), or whether it is on the monographic level, i.e. a stand-alone publication (e.g. a novel).

Properties

Value set type	enumerated, closed
XML name	vs_titleLevel.xml

Legal values

Value	Description
empty	No title, hence no title level. **Default**
m	Monographic title
a	Analytic title

3.3 Additional value sets for text classification

Text classification outside the scope of standard TEI header semantics is achieved by using a number of `<catRef>` schemes inside the `<textClass>` element. This special information is needed to enable older corpus material like the DDOC and KORPUS 2000 to be easily integrated in the new structure. The following types of information are inherited from these two corpora, the general structure for the `<catRef>` element being

```
<catRef
  scheme="http://ctb.dsl.dk/class/catRef/textGroup/scheme"
  target="#target"/>
```

where the *schemes* are in use can be seen under *myClassification*, see 3.2.1 on page 42.

In CTB, there is no `<catRef>` scheme for genre information. Instead, the `<factuality>` element under `<textDesc>` is used. DDOC and KORPUS 2000 genre values (as well as other obsolete values in an CTB context) should be mapped to the CTB header, see Asmussen (2009).

4 Document history

The most recent version may have added new values to the value sets listed in Section 3.2.1 and contain some minor fixes. This is not always explicitly listed in the history. The most recent version can be downloaded from here:
http://korpus.dsl.dk/clarin/corpus-doc/text-header.pdf

A detailed document history is no longer maintained.

5 References

Andersen, M. S., Asmussen, H., and Asmussen, J. (2002). The project of Korpus 2000 Going Public. In Braasch, A. and Povlsen, C., editors, *Proceedings of the 10th EURALEX International Congress*, volume 1, pages 291–299, Copenhagen. Euralex.

Asmussen, J. (2005). Automatic detection of new domain-specific words, using document classification and frequency profiling. In *Proceedings of the Corpus Linguistics 2005 conference*, volume 1, Birmingham.

Asmussen, J. (2009). Converting existing corpora to CTB-TEI. Technical report, Det Danske Sprog- og Litteraturselskab, korpus.dsl.dk/clarin/corpus-doc/converting_corpora.pdf.

Asmussen, J. (2013a). Aim and concepts. Technical report, DK-CLARIN, korpus.dsl.dk/clarin/corpus-doc/concepts.pdf.

Asmussen, J. (2013b). Text formatting. Technical report, DK-CLARIN, korpus.dsl.dk/clarin/corpus-doc/text-format.pdf.

Burnard, L. (2007). Reference Guide for the British National Corpus (XML Edition). Technical report, Research Technologies Service at Oxford University Computing Services, www.natcorp.ox.ac.uk/XMLedition/URG/index.html.

Keson, B. K. (1998a). Documentation of The Danish Morphosyntactically Tagged PAROLE Corpus. Technical report, DSL, korpus.dsl.dk/e-resurser/paroledoc_en.pdf.

Keson, B. K. (1998b). Vejledning til det danske morfosyntaktisk taggede PAROLE-korpus. Technical report, DSL, korpus.dsl.dk/e-resurser/paroledoc_dk.pdf.

Norling-Christensen, O. and Asmussen, J. (1998). The Corpus of The Danish Dictionary. *Lexikos. Afrilex Series*, 8:223–242.