

Harald Lungen and Ines Pisetta

Conversion into the archival format I5

Abstract: The IDS repository aims at long-term archival of linguistic resources and tools in the field of German studies. This chapter introduces I5, the archival format for textual data in the repository. I5 is characterised as a TEI customisation, featuring a tripartite corpus macrostructure and some renamings and restructurings of TEI elements to fit the corpus holdings compiled since 1964 at the IDS. After a brief introduction to I5, the contribution shows how the conversion to I5 is achieved for various input formats by using the examples of features of concrete corpora in the archive. The two cases covered are instances of the KED corpus of simple German which came as a CWB-based XML called VRT, and Wikipedia corpora which came in Wiki markup, both of which were converted to I5 using different strategies. We further show how I5 data are ingested into the archive and what happens if someone wishes to extract such data from the archive for their own subsequent use. Finally we mention some alternative archival formats used in other archives such as TEI proper or DTA-Bf.


Keywords: TEI, I5, XML, verticalised text format (VRT), Component Metadata Infrastructure (CMDI), InvenioRDM, repository

1 Introduction: I5 as an archival format

The IDS Repository aims at long-term archival of linguistic resources and tools in the field of German studies. It archives a most of the corpus holdings of the Leibniz Institute for the German Language (IDS), but also other, IDS-external corpus resources that are relevant in the field of German linguistics, provided they are (re-)encoded in the archival format I5.

Acknowledgement: Section 4 of this chapter by author Ines Pisetta is based on work within Text+, funded by the Deutsche Forschungsgemeinschaft (DFG, project number 460033370) as part of the German National Research Data Infrastructure (NFDI e. V.). Furthermore the author Ines Pisetta acknowledges the use of Large Language Models (LLMs) as writing aids in phrasing Section 4, based on the author's notes, ideas and concepts. The author retains full responsibility for the content.

Harald Lungen, Ines Pisetta, Digitale Sprachwissenschaft, Leibniz-Institut für Deutsche Sprache, R5 6-13, D-68161 Mannheim, Baden-Württemberg, Germany, e-mails: luengen@ids-mannheim.de, pisetta@ids-mannheim.de

Open Access. © 2025 the author(s), published by Walter de Gruyter GmbH, Berlin/Boston  This work is licensed under the Creative Commons Attribution 4.0 International License.
<https://doi.org/10.1515/9783112208212-010>

I5 is short for IDS-TEI P5. It is defined as a TEI customisation (a formal derivation of the official TEI format) that features a tripartite corpus macrostructure and some renamings and restructurings of TEI elements to fit all written corpus holdings compiled since 1964 at the IDS. It is used for the more than 60 billion tokens of text of DEREKO, the German Reference Corpus (Kupietz et al. 2022), i. e. for the IDS corpora that can be analysed online using the IDS corpus research software COSMAS II (older system, Bodmer Mory 2014) or KorAP (new system, Kupietz et al. 2022), which are both based on I5 as well. As a consequence, I5 is also used as an archival format in the IDS long-term data repository. This chapter describes the main features of I5 and gives an idea of what it takes to convert an annotated corpus to I5. The terms repository and archive will be used interchangeably.

IDS has been building written corpora since 1964, and they have not always been marked up as I5. Before I5, IDS corpora were formatted according to the XCES (Ide et al. 2000) standard. In fact, a large part of the definition of I5 consisted in customising the definition of TEI elements so that they matched the definitions and names of elements in XCES. For example, XCES had @id as a global attribute while TEI P5 has @xml:id for the same purpose, i. e. to indicate an identifier. Consequently, I5 uses @id, i. e. the formal customisation of I5 contains a renaming of @xml:id to @id. Other changes concern the introduction of IDS-specific elements such as <korpusSigle>, <dokumentSigle>, and <textSigle> for specific kinds of IDs relating to the three parts that make up the macrostructure of a corpus in I5. These three parts are the corpus level (the corpus file as a whole, characterised by the root element <idsCorpus>), the document level (a certain grouping of texts, characterised by the element <idsDoc>), and the text level (<idsText>). Each corpus has these three levels, and one <idsCorpus> must consist of at least one <idsDoc>, and one <idsDoc> must in turn consist of at least one <idsText> as depicted in Listing 1.

```

<idsCorpus>
  <idsHeader>...</idsHeader>
  <idsDoc>
    <idsHeader>...</idsHeader>
    <idsText>
      <idsHeader>...</idsHeader>
      [...]
    </idsText>
    <idsText>...</idsText>
    [... more idsTexts]
  </idsDoc>
  [... more idsDocs]
</idsCorpus>

```

Listing 1: Corpus macrostructure in I5.

For the record, this means that `<idsCorpus>` corresponds to the element `<teiCorpus>` in TEI P5, and `<idsText>` corresponds to the element `<TEI>` in TEI P5. It also means that for each corpus that is to be converted to I5 it has to be decided how to map the original corpus structure onto the tripartite macrostructure of an I5 file. For example, in the case of the press corpora in DeReKo, it was decided that each year of a newspaper source corresponds to an `<idsCorpus>`, each month corresponds to an `<idsDoc>`, and each single newspaper article corresponds to an `<idsText>` (sic: one day i. e. a single edition of the paper has no corresponding element in the I5 macrostructure). In the case of a corpus of novels, each `<idsDoc>` contains exactly one `<idsText>` which is one novel (book) as a whole.

Each macrostructure level will also contain an `<idsHeader>` element containing metadata pertaining to the specific level (cf. Listing 1). For example the `<idsHeader>` under `<idsCorpus>` will contain metadata on the corpus level, such as the name of the corpus and of the project and/or people responsible for compiling the corpus, information about the source of the corpus data, such as its bibliographical specification or an URL, about licensing the corpus, and also some textual and linguistic characteristics that pertain to the whole corpus such as its language(s) or text genre(s). The `<idsHeader>` under the `<idsDoc>` element will specify information about the particular grouping that `<idsDoc>` stands for, sometimes only scarce when the grouping is fairly arbitrary or only formal (such as the `<idsDoc>` elements representing initial letters within the IDS Wikipedia corpora). And finally, the `<idsHeader>` under `<idsText>` will contain metadata about the particular text, for example in case of a newspaper article the issue and the date when it was published, the page number, the column heading, its topic domain, its language, and its title and authors if available.

The history and setup of I5 is laid out in an article by Lungen and Sperberg-McQueen (2012). When I5 was introduced at the IDS in 2012, the actual DeReKo corpora and their markup were not changed. They were merely no longer validated against the XCES DTD but instead validated against the new I5 DTD. Since then, the I5 schema has sometimes been extended e. g. to include new CMC-specific elements such as `<posting>`. The I5 ODD (TEI customisation file) and the derived DTD and other documentation can be found online.¹

¹ <https://www.ids-mannheim.de/en/digspra/corpus-linguistics/projects/corpus-development/ids-text-model/>

2 Conversion of features of specific sources into the archival format

2.1 Conversion of annotated verticalised text

Many corpora that have been built in corpus linguistic projects come in a one-word-per-line format called verticalised text format (VRT). Each token appears in one line of text, and its token-based annotations are added as columns to the same line. Such a vertical format is for example required by the analysis tools Corpus Workbench (Evert and CWB Development Team 2022) and Sketch Engine (Kilgarriff et al. 2014). Structural XML markup such as `<text>`, `<paragraph>` (or `<p>`) and `<sentence>` (or `<s>`) elements can also be included in VRT when each such XML element tag is placed on a single line.

Listing 2 shows a VRT-like encoded sentence from the source encoding of the corpus KED (Korpus einfaches Deutsch, Jach and Dietz 2024). The first column of each line representing a token contains the orthographic form of the token as it occurred in the primary text. The second column contains its base form or lemma, i. e. the nominative singular form for nouns or the infinitive in case of verbs (see Ljubešić and Erjavec 2025). The third column contains a part-of-speech (POS) tag according to the STTS tagset (Schiller et al. 1999) which is a de-facto standard for German in terms of a tagset for POS tagging and used by most German language taggers.² In fact, the triple consisting of orthographic form, base (lemma) form and POS tag corresponds to the typical output of a POS tagger such as the one integrated in the spacy NLP library for python (Honnibal et al. 2020) which was used for KED.

When a tagging structure like Listing 2 is converted to I5 or TEL, each token form should be represented by a `<w>` element, where the original form of the token (VRT column 1) appears as its immediate text content, and its base form (VRT column 2) and POS information (VRT column 3) should go into the attributes `@lemma` and `@pos`.³ Note that while in VRT the order of the columns is crucial, in I5 (and generally in XML) the order of attributes is irrelevant. Similarly, while the newline character is used as a token delimiter in VRT, in I5 is it any amount of whitespace. That way, the VRT code in Listing 2 will be converted to the I5 snippet in Listing 3.

² The STTS tags shown and their meanings are: ADJA: attributive adjective, APPR: preposition, ART: definite or indefinite article, CARD: cardinal number, NN: common noun, PIAT: attributing indefinite pronoun, PIS: substituting indefinite pronoun, PRELS: substituting relative pronoun, VAFIN: finite auxiliary verb, VVPP: present participle of main verb, VVFIN: finite main verb, \$,: comma, \$.: sentence terminating punctuation.

³ A prefixed '@' marks an attribute name in XML parlance.

```

<paragraph>
<sentence>
Neandertaler      Neandertaler      NN
nennt            nennen            VVFIN
man              man               PIS
eine             ein               ART
bestimmte        bestimmt          ADJA
Art              Art               NN
von              von               APPR
Menschen         Mensch            NN
,                --                $,
die              der               PRELS
vor              vor               APPR
vielen           vieler            PIAT
tausend          tausend           CARD
Jahren           Jahr              NN
gelebt           leben             VVPP
haben            haben             VAFIN
.                --                $.
</sentence>
[... ]
</paragraph>

```

Listing 2: Tagged sentence from the corpus KED (Korpus Einfaches Deutsch/Corpus Simple German, Jach and Dietz 2024) in a verticalised format (corpus text KED_#05999_klexikon.xml, retrieved from https://klexikon.zum.de/wiki/Neandertaler_2023-03-09). Translation: ‘Neanderthals are a specific type of human that lived many thousands of years ago’. Klexikon is a German online lexicon for children. The POS tags are from the tagset Schiller et al. (1999).

Within VRT, even more token-related annotations may be represented in additional columns, or positional attributes. The corpora of the Finnish Kielipankki corpus archive,⁴ for example, make extensive use of this. Consider the token-based annotations contained in the first eight columns of a sentence in the Suomi 24 corpus (Suomi24 2021) in Listing 4. Here, the output of the TurkuNLP parser-tagger⁵ in the form of the CoNLL-U standard⁶ has been integrated into VRT (in the original, there are even more columns representing further analyses). For the sentence “Nukuin todella hyvin ja tosi pitkään” (engl. “I slept really well and very long”), the VRT column 1 contains the original token, column 2 the running number of the token in the sentence, column 3: the lemma (base form), column 4: lemmacomp i. e. lemma with compound segmentation information (nothing segmented in the cur-

4 <https://www.kielipankki.fi/>

5 <https://turkunlp.org/>

6 <http://universaldependencies.org/docs/format.html>

```

<p>
<s>
<w lemma="Neandertaler" pos="NN">Neandertaler</w>
<w lemma="nennen" pos="VFIN">nennt</w>
<w lemma="man" pos="PIS">man</w>
<w lemma="ein" pos="ART">eine</w>
<w lemma="bestimmt" pos="ADJA">bestimmte</w>
<w lemma="Art" pos="NN">Art</w>
<w lemma="von" pos="APPR">von</w>
<w lemma="Mensch" pos="NN">Menschen</w>
<w lemma="--" pos=",$,>,</w>
<w lemma="der" pos="PRELS">die</w>
<w lemma="vor" pos="APPR">vor</w>
<w lemma="vieler" pos="PIAT">vielen</w>
<w lemma="tausend" pos="CARD">tausend</w>
<w lemma="Jahr" pos="NN">Jahren</w>
<w lemma="leben" pos="VPP">gelebt</w>
<w lemma="haben" pos="VAFIN">haben</w>
<w lemma="--" pos=",$.>.</w>
</s>
...
</p>

```

Listing 3: VRT of Listing 2 converted to I5.

rent example), column 5 the POS according to the Finnish Universal Dependencies (Pyysalo et al. 2015),⁷ column 6: a morphosyntactic description of the form (‘_’ in the case of function words), column 7: a pointer to the index of the token to which the current one stands in a dependency relation, and column 8: the name of the dependency relation.

Listing 5 shows the same sentence and its annotations converted to I5. The @head and depre1 attributes were specifically introduced in I5 to represent the CoNLL-U columns for dependency relations. The pointer in @head naturally points to a <w> with a corresponding running number in @n, and marks it as the head of the current token. Hence, *pitkään* (“long”), for example, is the head of *tosi* (“very”), which is an adverbial modifier (advmod) of *pitkään* as annotated in @depre1. Note that @head and @depre1 are not (yet) part of the official TEI P5. The attribute @msd contains the morphosyntactic description. For converting further VRT columns in I5, the best way would be to customise additional attributes for <w>, including one for the “lemma-

⁷ The POS tags shown and their meanings are: Adv: adverb, C[SUBCAT=CC]: coordinating conjunction, Punct: punctuation.

```

<sentence id="34" lang="fin" polarity="pos">
Nukuin 1 nukkua nukkua V PRS_Sg1|VOICE_Act|TENSE_Prt|MOOD_Ind|CASECHANGE_Up
↳ 0 ROOT
todella 2 todella todella Adv _ 3 advmod
hyvin 3 hyvin hyvin Adv _ 1 advmod
ja 4 ja ja C SUBCAT_CC 3 cc
tosi 5 tosi tosi Adv _ 6 advmod
pitkään 6 pitkään pitkään Adv _ 3 conj
. 7 . Punct _ 1 punct
</sentence>

```

Listing 4: A sentence from the corpus Suomi 24 in VRT (from text 15105561:92340676). The original contains several more columns with further annotations.

```

<s id="s35" xml:lang="fin">
  <w n="1" lemma="nukkua" pos="V"
    ↳ msd="PRS_Sg1|VOICE_Act|TENSE_Prt|MOOD_Ind|CASECHANGE_Up" head="0"
    ↳ deprel="ROOT">Nukuin</w>
  <w n="2" lemma="todella" pos="Adv" msd="_" head="3" deprel="advmod">todella</w>
  <w n="3" lemma="hyvin" pos="Adv" msd="_" head="1" deprel="advmod">hyvin</w>
  <w n="4" lemma="ja" pos="C" msd="SUBCAT_CC" head="3" deprel="cc">ja</w>
  <w n="5" lemma="tosi" pos="Adv" msd="_" head="6" deprel="advmod">tosi</w>
  <w n="6" lemma="pitkään" pos="Adv" msd="_" head="3" deprel="conj">pitkään</w>
  <w n="7" lemma="." pos="Punct" msd="_" head="1" deprel="punct">.</w>
</s>

```

Listing 5: Conversion in I5 of the sentence tokens and tags in Listing 4.

comp" information in VRT column 4, which so far is not contained in the I5 conversion.

2.2 Conversion of corpora of computer-mediated communication: Wikipedia talk pages

In the case of corpora based on the online encyclopedia Wikipedia, the source data come as wiki markup (also known as wikitext or wikiCode), contained in a Wikipedia archive, i.e. the content of a Wikipedia at one specific point in time, deployed as a database dump by the Wikimedia foundation.⁸ Wiki markup is a specific kind of light markup used on wiki platforms that are driven by the MediaWiki technology, such as Wikipedia.

⁸ at <https://dumps.wikimedia.org/>

Besides the well-known encyclopedic articles, Wikipedia also contains ‘talk pages’, which are a space where the Wikipedia authors discuss the composition of articles in the collaborative writing process, for example when they make changes, add images, or put forward criticism of what somebody else has written in the article. In fact, there is a talk page associated with each article in Wikipedia that can be reached via the “talk” link (“Diskussion” in German) under the title of the article page. These talk pages form very large and linguistically interesting archives of Computer-mediated communication (CMC), i. e. dialogic communication “mediated by digital technologies (such as text on web pages, written exchanges in chats and forums, interactions with artificial intelligence systems, [or] the spoken conversations in internet video meetings)” (TEI Consortium 2025). The IDS has built several corpora of Wikipedia articles and talk pages of different languages and from different Wikipedia snapshots (version at a specific point in time). For composing a contribution to a talk page on Wikipedia, authors use the wiki markup just as they use it for articles, while the Wikipedia Guidelines⁹ encourage them to use indentations to mark their contribution on the talk page. The Media Wiki software generates a web page that looks for example like the one depicted in Figure 1.

Alternative Begriffe [Quelltext bearbeiten]

Letzter Kommentar: [vor 2 Jahren](#) | [2 Kommentare](#) | [2 Personen sind an der Diskussion beteiligt](#)

Ich meine im Alltag schon oft die Begriffe 'Wahlschwelle' und 'Wahlhürde' gehört zu haben. Was ist der Status dieser Ausdrücke? Beim Googeln ergibt sich, dass sie auch in Medien vorkommen, obwohl nicht oft, z. B. [Wahlschwelle](#), [Wahlhürde](#). Sollten sie als umgangssprachliche Alternativen im Artikel erwähnt werden? —[Caoimhin ceallach \(Diskussion\)](#) 13:10, 23. Mai 2022 (CEST) [Beantworten](#)

In der Schweiz wird häufig auch der Begriff *Quorum* verwendet, besonders von offizieller Seite. [Wahlgesetz AG](#), [Wahlanleitung GR](#) —[Gbuvn \(Diskussion\)](#) 13:53, 23. Mai 2022 (CEST) [Beantworten](#)

Figure 1: Part of the Wikipedia talk page for the article on Sperrklausel (“electoral threshold”).

When comparing the wiki markup shown in Listing 6 with the screenshot of the talk page clip in Figure 1, we can see that in the wiki markup for example “==” is used to mark headings, single square brackets (‘[]’) are used to mark external links, double square brackets (‘[[]]’) are used to mark internal links (“Wikilinks”), “:” is used to mark indentations (marking user contributions), and “—” marks the insertion of a user signature and a timestamp, and the reply template (“Beantworten”) which indicates the end of the contribution.

Listing 7 shows how the structure of this piece of dialogue in wiki markup is rendered in I5 using the elements `<div>`(ision), here used to mark a thread, and `<post ing>`

⁹ <https://en.wikipedia.org/wiki/Wikipedia:Indentation>

== Alternative Begriffe ==

Ich meine im Alltag schon oft die Begriffe 'Wahlschwelle' und 'Wahlhürde' gehört
 ↳ zu haben. Was ist der Status dieser Ausdrücke? Beim Googeln ergibt sich, dass
 ↳ sie auch in Medien vorkommen, obwohl nicht oft, z. B.
 ↳ [https://www.tagblatt.ch/ostschweiz/
 ↳ rheintal/wir-liegen-offenbar-richtig-ld.217751 Wahlschwelle],
 ↳ [https://www.spiegel.de/politik/europaeische-union-eu-parlament-
 ↳ stimmt-fuer-sperrklausel-bei-europawahlen-a-c5de1b8a-f327-4bbe-
 ↳ 9732-242acb392fe5 Wahlhürde]. Sollten sie als umgangssprachliche Alternativen
 ↳ im Artikel erwähnt werden? --[[Benutzer:Caoimhin ceallach|Caoimhin ceallach]]
 ↳ ([[Benutzer Diskussion:Caoimhin ceallach|Diskussion]]) 13:10, 23. Mai 2022
 ↳ (CEST)
 :In der Schweiz wird häufig auch der Begriff ''Quorum'' verwendet, besonders von
 ↳ offizieller Seite.
 ↳ [https://gesetzessammlungen.ag.ch/app/de/texts_of_law/152.100/ versions/1270
 ↳ Wahlgesetz AG], [https://www.gr.ch/DE/publikationen/abstimmungenwahlen/
 ↳ Grossratswahlen-2022/faq/vier/Seiten/01_Doppelter_Pukelsheim.aspx
 ↳ Wahlanleitung GR] --[[Benutzer:Gbuvn|Gbuvn]] ([[Benutzer
 ↳ Diskussion:Gbuvn|Diskussion]]) 13:53, 23. Mai 2022 (CEST)

Listing 6: Wiki markup underlying the talk page clip in Figure 1.

which is used to mark a post in any CMC genre represented in I5. The original indentation is not reflected in the I5/XML structure but encoded in the @indentLevel attribute at <posting> (level "0" for the top level, "1" for the first indentation, and so forth). Links (internal and external) are encoded using the <ref> element, while <signed>, <name> and <date> are used to encode the signature, the user name, and the timestamp, respectively. There is also a list of users contained in the I5 header of the text, and the @who attribute at <posting> points to the respective <listPerson> element that represents information about the author of the post, such as their nickname and user page (not shown).

Wiki markup is difficult to parse which makes it a challenge to create corpora in I5 or TEI from the Wikipedia archives. The IDS approach based on the Sweble parser (Dohrn and Riehle 2011) and XSLT is described in Margaretha and Lungen (2014).¹⁰ An approach to convert wiki markup sources to the CMC-core TEI customisation is described in Ho-Dac (2024). Note that the official TEI Guidelines contain markup for CMC as of 2024 (TEI Consortium 2025). However, the I5 markup for CMC is based on the 2012 DeRiK proposal (Beißwenger et al. 2012) (with e.g. <posting> instead of <post>).

¹⁰ see also <https://github.com/IDS-Mannheim/Wikipedia-Corpus-Builder>

```

<div n="2" type="thread">
  <head> Alternative Begriffe </head>
  <posting id="i.1513683_1_1" indentLevel="0" who="WU00285477"
    ↪ when-iso="2022-05-23T13:10+02">
    <p>Ich meine im Alltag schon oft die Begriffe 'Wahlschwelle' und 'Wahlhürde'
    ↪ gehört zu haben. Was ist der Status dieser Ausdrücke? Beim Googeln ergibt
    ↪ sich, dass sie auch in Medien vorkommen, obwohl nicht oft, z. B. <ref
    ↪ target="https://www.tagblatt.ch/ostschweiz/rheintal/wir-liegen-offenbar-
    ↪ richtig-ld.217751">Wahlschwelle</ref>, <ref target="https://
    ↪ www.spiegel.de/politik/europaeische-union-eu-parlament-
    ↪ stimmt-fuer-sperrklausel-bei-europawahlen-a-c5de1b8a-f327-
    ↪ 4bbe-9732-242acb392fe5">Wahlhürde</ref>. Sollten sie als
    ↪ umgangssprachliche Alternativen im Artikel erwähnt werden? &#x2014;<signed
    ↪ type="signed"><ref target="https://de.wiki-
    ↪ pedia.org/wiki/Benutzer:Caoimhin_ceallach"> <name>Caoimhin
    ↪ ceallach</name></ref><date>13:10, 23. Mai 2022 (CEST)</date> </signed>
    </p>
  </posting>
  <posting id="i.1513683_1_2" indentLevel="1" who="WU00075069"
    ↪ when-iso="2022-05-23T13:53+02">
    <p>In der Schweiz wird häufig auch der Begriff <hi rend="it">Quorum</hi>
    ↪ verwendet, besonders von offizieller Seite. <ref
    ↪ target="https://gesetzessammlungen.ag.ch/app/de/texts_of_
    ↪ law/152.100/versions/1270">Wahlgesetz AG</ref>, <ref target=
    ↪ "https://www.gr.ch/DE/publikationen/abstimmungenwahlen/
    ↪ Grossratswahlen-2022/faq/vier/Seiten/01_Doppelter_Pukelsheim
    ↪ .aspx">Wahlanleitung GR</ref> --<signed type="signed"><ref
    ↪ target="https://de.wikipedia.org/wiki/Benutzer:Gbuvn">
    ↪ <name>Gbuvn</name></ref><date>13:53, 23. Mai 2022 (CEST) </date></signed>
    </p>
  </posting>
</div>

```

Listing 7: Conversion to I5 of the thread shown in wiki markup in Listing 6.

3 Alternative formats

Why is I5 a good format for archiving textual data? First of all, it is based on the established XML standard which has been published in 1998 and which is a non-proprietary, well-defined, well-understood data format, used and accepted widely in a variety of fields even outside text corpora, such as databases or web technologies. It comes with a wealth of tools and systems, even programming languages and programming libraries that are specifically designed to handle it. In short, the chances are quite high that XML will be used and understood for some time into

the future. Even if it should get out of use, the semantics of data and data structures in XML can always be recovered to some degree by looking at the named elements and attributes and the document grammar (schema) that comes with valid XML files. This stands in contrast with e. g. tabular data formats like CoNLL-U or untyped formats like JSON, where the meaning of the lines and columns or hierarchy of the data structure must be encoded elsewhere, outside the structure itself, and can get lost or is sometimes never even recorded.

Applications of XML like the TEI that have the status of a community standard provide an even more elaborated level of semantics for data categories i. e. XML elements and attributes. The TEI Guidelines constitute a comprehensive semantics written in prose of the XML application that is TEI, which has been developed over the years by the international Digital Humanities community. This is of course also a plus for the sustainability of data and data categories. We think that all customisations derived from the TEI via the formal ODD mechanism are sustainable in the same way as any TEI. Such a customisation is I5, but also for example the DTA-Bf format developed for the German Text Archive (Haaf et al. 2014), which would consequently be a likewise suitable format for archiving textual data. In comparison with I5, DTA-Bf is more geared towards historical documents. One could argue that corpora marked up as VRT, which constitutes a kind of unofficial standard for corpus data as well, is already sufficient for long-term archiving and ingestion into a repository. But first of all, VRT, although it looks like XML, is not proper XML, hence it can frequently not be parsed using XML technology but needs to be manipulated by specific VRT corpus tools such as the CorpusWorkBench CWB. Second, there is no standard set of categories to be used in VRT and no established mechanism to encode the semantics of VRT structures, so that for example, the semantics of the positional attributes used for a corpus would be explained in a ReadMe file.

4 Ingestion into the archive: a detailed workflow for I5 data

After presenting the advantages of I5 as an archival format for textual corpus data and detailing how a conversion into I5 can be achieved, this section will show how such data is ingested into a long-term research data repository. The terms repository and archive will be used interchangeably. Furthermore, a pipeline graphic is provided as a supplementary material in the appendix (Figure A1) to guide the reader along the process in the most basic steps and is intended to serve as a general overview. The text will provide detailed information on the most important steps in the ingest process.

In the process of ingesting I5 data into the archive, several measures are employed to ensure the integrity, accessibility, and long-term preservation of the data. The steps range from the creation of the Submission Information Package (SIP) to the final publication of the data in the archive. The workflow adheres to the Open Archival Information System (OAIS) reference model, ensuring that the data is properly managed and preserved according to international standards. To achieve semantic preservation, descriptive metadata is included and the logical structure of the data is mirrored by arranging the digital objects in the archive in a strict hierarchy (see Pisetta and Trippel 2025).

4.1 Creating the metadata

The first step in the ingestion process is the generation of Component Metadata Infrastructure (CMDI) metadata for each I5 file. CMDI is a metadata standard within the Common Language Resources and Technology Infrastructure (CLARIN).¹¹ It is highly flexible through the use of components and profiles that can be recombined, reused or created from scratch and is usually specifically used for language data. For each I5 file, a TextCorpusProfile metadata file is created. This is done by parsing the `<idsHeader>` of the `<idsCorpus>` element of the I5 file (Section 1) and retrieving information about the content, structure, context, language, text type, and any annotations or markup that may be present. Subsequently this metadata information is then stored in the TextCorpusProfile CMDI file.

The TextCorpusProfile metadata uses the CMDI schema `clarin.eu:cr1:p_1696338267545`,¹² which is specifically designed for describing text corpora. This schema ensures that all relevant metadata is captured in a standardised format, making it easier for users to search and retrieve the data in the future.

In addition to the corpus-level metadata, CMDI metadata is also generated at the collection level. As an example, assume that all newspaper articles from a specific outlet from one year will form the corpus. Assembling all the corpora for each year will form the collection. The collection-level metadata provide an overview of the entire database and links to the different component corpora. These metadata files are created using the CMDI schema `clarin.eu:cr1:p_1659015263839`,¹³ which is tailored for describing collections of linguistic resources.

¹¹ <https://www.clarin.eu/>

¹² TextCorpusProfile: `clarin.eu:cr1:p_1696338267545`

¹³ CollectionProfile: `clarin.eu:cr1:p_1659015263839`

4.2 Assembling a Submission Information Package (SIP)

After creating corpus- and collection-level metadata, the next step is to assemble the Submission Information Package (SIP), which is what is going to be ingested into the archive. A SIP can take multiple forms. As an example, we discuss a SIP that is realised as BagIt package (see Pisetta and Trippel 2025). A detailed overview can be found in Listing 8.

```
sip/
|-- data
|   |-- Metadata {1}
|       |-- CmdICollection {1}
|           \-- Collection-CMDI-File {1,}
|   | \-- I5 {1}
|       | \-- TextCorpus-CMDI-File {1,}
|   \-- Content {1}
|       | \-- I5 {1}
|           \-- I5 content file {1,}
|-- bagit.txt {1}
|-- manifest-sha512.txt {1}
|-- package-info.txt {1}
|-- recordmap.xml {1}
\-- tagmanifest-sha512.txt {1}
```

Listing 8: SIP example.

- *data/*: This directory contains the actual data and metadata that will be ingested into the archive. It is divided into two subdirectories: *Metadata* and *Content*.
 - *Metadata/*: This subdirectory contains the CMDI metadata files. It is further divided into *CmdiCollection*, which holds the collection-level metadata, and *I5*, which holds the corpus-level metadata for each I5 file.
 - *Content/*: This subdirectory contains the actual I5 content files. Each I5 file is stored in the *I5* subdirectory, with a potential to also have other subdirectories besides *I5*.
- *bagit.txt*: This file contains basic information about the BagIt package, including the version of the BagIt specification used and the character encoding of the tag files.
- *manifest-sha512.txt*: This file contains a list of all the files in the *data/* directory, along with their SHA-512 checksums. This ensures the integrity of the data during transfer and storage.
- *package-info.txt*: This file contains information on the amount of files and total byte size in the *data/* directory.

- *recordmap.xml*: This file is a custom addition to the BagIt package and is required for the ingest process. It contains a mapping of the metadata and content files in the SIP to the corresponding records in the archive. This mapping is used during the ingest process to create the necessary records in the archive (Section 4.3.3).
- *tagmanifest-sha512.txt*: This file contains a list of all the aforementioned files, along with their SHA-512 checksums. This ensures the integrity of the entire package, including the metadata and manifest files.

4.3 From SIP to AIP

Following the OAIS model, the next step is to create the Archival Information Package (AIP) (see Pisetta and Trippel 2025), the form in which the data is stored within the archive. Converting the SIP into the AIP involves several procedures which will be laid out next.

4.3.1 Validation & upload

Before the SIP can be uploaded it needs to be validated first. The validation process checks that all the necessary files are present, that the checksums match and that the structure of the SIP is correct. In the subsequent upload process, the SIP is transferred to the archive's storage system. During this process, the archive system performs a series of checks to ensure that the data is complete and that there are no errors or inconsistencies.

4.3.2 InvenioRDM – a digital archive solution

To understand what happens next, it is important to go a little into detail about the archival system. There are several digital archive software solutions readily available, such as Dspace,¹⁴ Fedora Commons¹⁵ or Invenio.¹⁶ As the goal is to store language data for research and preservation purposes, this chapter will focus specifically on InvenioRDM (Invenio Research Data Management). InvenioRDM organises

14 <https://dspace.org/>

15 <https://fedorarepository.org/>

16 <https://invenio-software.org/>

data in *records*. The most important parts of a record are the access restriction, the metadata and the actual data files. The access parameter decides whether the record is publicly visible or restricted to specific groups of users. It is also possible to only restrict file access and keep the metadata visible. InvenioRDM records have several metadata fields, most of which are optional. Beyond mandatory information like record title, creators, publication date and resource type, it is also possible to provide details about rights, subjects, languages, related work, alternate identifiers, size and many more. Apart from that it is even possible to create custom metadata fields.

4.3.3 Organizing records in a logical structure

The creation of records is guided by the `recordmap.xml` file, which defines the hierarchy, metadata and data of the records. An example recordmap for the DeReKo-2014-I corpus can be found in Listing 9.

The process begins with the creation of the root record, which represents the entire DeReKo release and is the top-level collection. Following that are lower-level collections like ZGE (the magazine *ZEIT Geschichte*). The necessary metadata information for an InvenioRDM record is extracted from the CMDI file and the CMDI metadata itself is stored in a custom CMDI metadata field. These collection-level records do not contain data in files and instead are “metadata-only” records. After creating all the records for the collections, the records for the I5 files will be set up next. These are created using the corpus-level CMDI metadata and are linked to the collection records they belong to via the InvenioRDM metadata field “related works” and “isPartOf”/“HasPart” relations. This ensures that the hierarchy and logical structure of the data is preserved. As with the collection-level records, the InvenioRDM metadata is retrieved from the CMDI file and the CMDI metadata itself is stored in the custom CMDI field. The I5 file is being uploaded and stored as a file with restricted access.

Deciding what should be a separate record requires knowledge about the data and its logical structure as well as careful consideration of granularity and cost (see Pisetta and Trippel 2025).

After the initial record creation, the records still remain in a draft state and are not yet published, since there are still some changes to be made.

4.3.4 Ensuring persistence & findability

A cornerstone of long-term digital archiving is to ensure that the data remains intact and findable over several years or even decades. To ensure lasting findability, the

```

<rootRecord title="DeReKo-2014-I">
  <metadata>
    data/Metadata/CmdiCollection/dereko-collection.cmdi
  </metadata>
  <records>
    <record title="zge-collection">
      <metadata>
        data/Metadata/CmdiCollection/zge-collection.cmdi
      </metadata>
      <records>
        <record title="zge12">
          <metadata>
            data/Metadata/I5/zge12.cmdi
          </metadata>
          <records/>
          <files>
            <file public="false">
              data/Content/I5/zge12.i5.xml
            </file>
          </files>
        </record>
        <record title="zge13">
          <metadata>
            data/Metadata/I5/zge13.cmdi
          </metadata>
          <records/>
          <files>
            <file public="false">
              data/Content/I5/zge13.i5.xml
            </file>
          </files>
        </record>
      </records>
    </record>
  </records>
</rootRecord>

```

Listing 9: Example of a recordmap.

use of Persistent Identifiers (PIDs) (see Pisetta and Trippel 2025) is paramount. In this case, Digital Object Identifiers (DOIs)¹⁷ will be created for each record. DOIs are persistent identifiers that provide a stable link to the records in the archive, ensuring that they can be easily cited and accessed over time.

The system connects to the DataCite API and generates a DOI for each record. This is called *minting*. As with the InvenioRDM records, any minted DOI will initially be in draft state. Draft DOIs are not yet registered in the global handle system and as such can still be deleted.

¹⁷ <https://www.doi.org/>

Once the DOIs have been minted, the next step is to edit the InvenioRDM records and CMDI metadata to include the minted DOIs that serve as persistent identifiers for these records.

With the information in the records now being complete, DOI and records are being published.

Once the records and DOIs have been published, the data is considered to be part of the repository's Archival Information Package (AIP). The AIP includes the Invenio record metadata, the CMDI metadata (as part of the Invenio metadata in the form of a custom metadata field) and the I5 content files (in the case of text corpora). The AIP is stored in the archive's long-term preservation system, where it will be managed and preserved according to the OAIS reference model.

4.3.5 Extracting data from the archive

Following the OAIS model it also needs to be possible to extract the data or information about it from the archive in the form of a Dissemination Information Package (DIP) (see Pisetta and Trippel 2025). With InvenioRDM there are several ways to extract information from the archive that can also be expanded upon and customised.

By default, InvenioRDM supports exporting records as JSON, DataCite,¹⁸ Citation Style Language (CSL),¹⁹ Dublin Core²⁰ and many more.²¹ The uploaded files can only be accessed if they are set to public.

Beyond that, InvenioRDM also provides the option to harvest metadata via the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).²² The standard formats here are `oai_dc` (Dublin Core), `oai_datacite` and `datacite` (not OAI-PMH v2 compliant). Again, because of the highly customisable nature, it is possible to deliver other metadata formats than the standard ones. After setting up a custom metadata field for CMDI metadata as explained previously, it is possible to adjust the OAI-PMH service so that it can also provide CMDI metadata by retrieving the information from that custom field. The CMDI metadata can then be harvested. These harvesting procedures are used, among other things, by services that provide infor-

¹⁸ <https://datacite.org/>

¹⁹ <https://citationstyles.org/>

²⁰ <https://www.dublincore.org/>

²¹ https://inveniordm.docs.cern.ch/reference/export_formats/

²² https://inveniordm.docs.cern.ch/reference/oai_pmh/

mation about resources archived at different places, such as the Text+ Registry²³ or the Virtual Language Observatory (VLO).²⁴

5 Conclusion

This chapter has provided a comprehensive overview of the I5 format, its role as an archival format for textual data within the German Reference Corpus (DeReKo), and the processes involved in converting various input formats into I5. The I5 format, as a TEI customisation, offers a structured and sustainable approach to archiving corpus data, ensuring long-term preservation and accessibility. By detailing the conversion of VRT tokens and their annotations, as well as Wiki markup from Wikipedia corpora, the chapter has demonstrated the flexibility and robustness of I5 in handling diverse data sources.

The ingestion process into the IDS long-term data repository, adhering to the OAIS reference model, further underscores the importance of metadata and persistent identifiers in maintaining the findability of archived data. The use of CMDI metadata ensures that the data is well-documented, while the integration with InvenioRDM facilitates efficient management, preservation and dissemination of the archived corpora.

In conclusion, the I5 format, with its strong foundation in XML and TEI standards, provides a reliable and future-proof solution for archiving textual data. The detailed workflows and examples presented in this chapter highlight the practical steps involved in converting, ingesting, and managing corpus data, ensuring that it remains accessible and usable for future research. As the field of corpus linguistics continues to evolve, the I5 format and the associated archival practices will play a crucial role in preserving and leveraging linguistic resources for years to come.

23 <https://registry.text-plus.org>

24 <https://www.clarin.eu/content/virtual-language-observatory-vlo>

Appendix

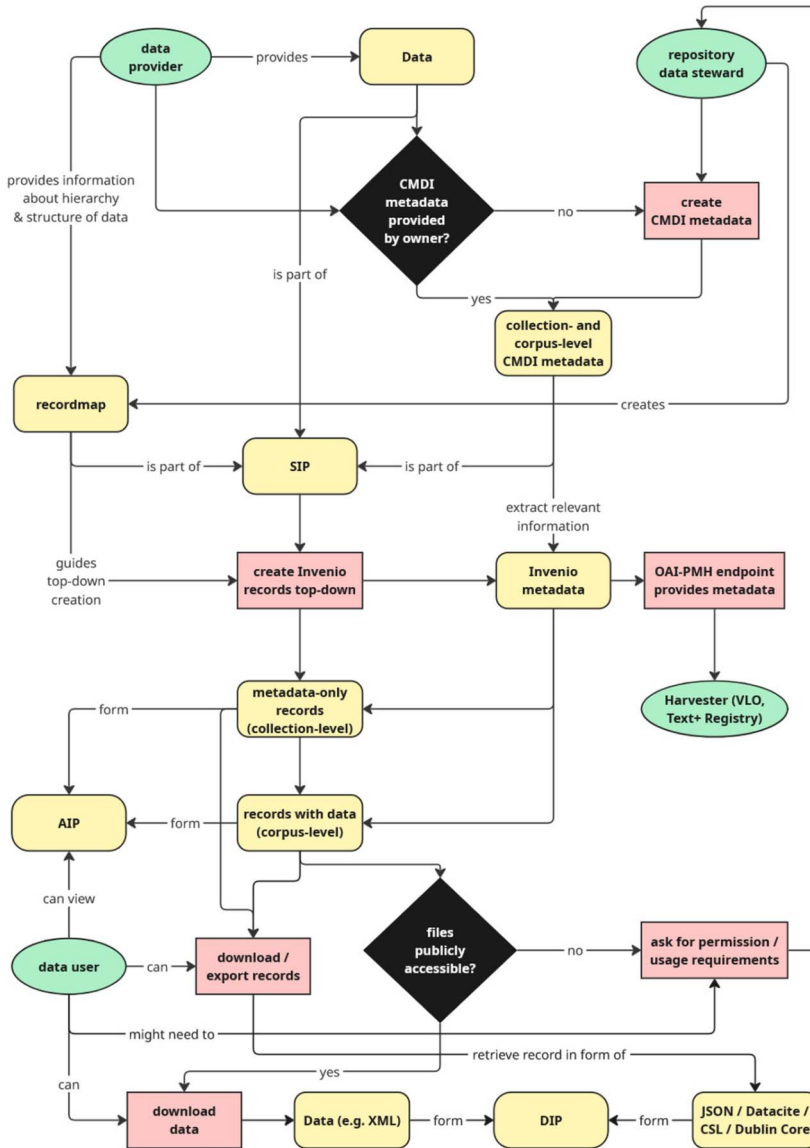


Figure A1: Simplified breakdown of ingest (SIP), preservation (AIP) and retrieval (DIP) of data in the repository. It also shows the involvement of key roles like the data provider, the data steward and the data user (see Pisetta and Trippel 2025).

Bibliography

- Beißwenger, Michael, Maria Ermakova, Alexander Geyken, Lothar Lemnitzer & Angelika Storrer. 2012. A TEI schema for the representation of computer-mediated communication. *Journal of the Text Encoding Initiative* 3.
- Bodmer Mory, Franck. 2014. Mit COSMAS II »in den Weiten der IDS-Korpora unterwegs«. In Melanie Steinle & Franz Josef Berens (eds.), *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache*, 376–385. Institut für Deutsche Sprache..
- Dohrn, Hannes & Dirk Riehle. 2011. Design and implementation of the sweble wikitext parser: Unlocking the structured data of Wikipedia. In *Proceedings of the 7th international symposium on Wikis and open collaboration*, 72–81.
- Evert, Stephanie & the CWB Development Team. 2022. The IMS Open Corpus Workbench (CWB). Corpus Encoding and Management Manual. CWB Version 3.5. https://cwb.sourceforge.io/files/CWB_Encoding_Tutorial.pdf.
- Haaf, Susanne, Alexander Geyken & Frank Wiegand. 2014. The DTA “base format”: A TEI subset for the compilation of a large reference corpus of printed text from multiple sources. *Journal of the Text Encoding Initiative* 8.
- Ho-Dac, Lydia-Mai. 2024. Building a comparable corpus of online discussions on Wikipedia: The EFG WikiCorpus. In Céline Poudat, Harald Lungen & Laura Herzberg (eds.), *Investigating Wikipedia*, 12–44. John Benjamins Publishing Company. <https://doi.org/10.1075/scl.121.01hod>.
- Honnibal, Matthew, Ines Montani, Sofie Van Landeghem & Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in Python. *Zenodo*. <https://doi.org/10.5281/zenodo.1212303>.
- Ide, Nancy, Patrice Bonhomme & Laurent Romary. 2000. XCES: An XML-based encoding standard for linguistic corpora. In *Proceedings of the second international conference on language resources and evaluation (LREC'00)*.
- Jach, Daniel & Gunther Dietz. 2024. Korpus Einfaches Deutsch (KED). *Korpora Deutsch als Fremdsprache* 4(1). 123–130.
- Kilgarriř, Adam, Vít Baisa, Jan Buřta, Miloř Jakubiřek, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý & Vít Suchomel. 2014. The Sketch Engine: ten years on. *Lexicography* 1(1). 7–36.
- Kupietz, Marc, Nils Diewald & Elisa Margaretha. 2022. Building paths to corpus data. In Darja Fiřer & Andreas Witt (eds.), *CLARIN. The infrastructure for language resources*, Vol. 1, 163–189. Digital Linguistics. de Gruyter.
- Ljubešić, Nikola & Tomař Erjavec. 2025. Part-of-speech tagging and related annotation. In Piotr Bański, Ulrich Heid & Laura Herzberg (eds.), *Harmonising language data: Standards for linguistic resources*. de Gruyter.
- Lungen, Harald & Michael C. M. Sperberg-McQueen. 2012. A TEI P5 document grammar for the IDS text model. *Journal of the Text Encoding Initiative (JTEI)* 3. <https://doi.org/10.4000/jtei.508>. <http://journals.openedition.org/jtei/508>.
- Margaretha, Eliza & Harald Lungen. 2014. Building linguistic corpora from Wikipedia articles and discussions. *Journal for Language Technology and Computational Linguistics* 29(2). 59–82.
- Pisetta, Ines & Thorsten Trippel. 2025. Standards and practices for long-term digital archiving. In Piotr Bański, Ulrich Heid & Laura Herzberg (eds.), *Harmonising language data: Standards for linguistic resources*. de Gruyter.
- Pyysalo, Sampo, Jenna Kanerva, Anna Missilä, Veronika Laippala & Filip Ginter. 2015. Universal dependencies for Finnish. In *Proceedings of the 20th nordic conference of computational linguistics (NODALIDA 2015)*, 163–172.

- Schiller, Anne, Simone Teufel, Christine Stöckert & Christine Thielen. 1999. *Guidelines für das Tagging deutsche Textkorpora mit STTS (Kleines und großes Tagset)*, Stuttgart: Universität Stuttgart.
- Suomi24. 2021. The Suomi24 Corpus 2018-2020, VRT version. Accessed: 2025-02-10. <http://urn.fi/urn:nbn:fi:lb-2021101524> (visited on 02/10/2025).
- TEI Consortium (ed.). 2025. *TEI P5: Guidelines for electronic text encoding and interchange*. Last updated on 24th January 2025, Accessed: 2025-02-10. TEI Consortium. Chap. 9. Computer-mediated Communication. <https://tei-c.org/release/doc/tei-p5-doc/en/html/CMC.html> (visited on 02/10/2025).