Published in: Navarretta, Costanza/Eskevich, Maria (Eds.): Proceedings of CLARIN Annual Conference 2020.
05 – 07 October 2020, Online Edition. - Utrecht: CLARIN, 2020. Pp. 146-150.
https://office.clarin.eu/v/CE-2020-1738-CLARIN2020_ConferenceProceedings.pdf

Signposts for CLARIN

Denis Arnold
Leibniz-Institut für Deutsche Sprache
Mannheim, Germany
arnold@ids-mannheim.de

Bernhard Fisseni
Leibniz-Institut für Deutsche Sprache
Mannheim, Germany
fisseni@ids-mannheim.de

Thorsten Trippel
Eberhard Karls Universität
Tübingen, Germany
thorsten.trippel@uni-tuebingen.de

Abstract

An implementation of CMDI-based signposts and its use is presented in this paper. Arnold et al. 2020 present Signposts as a solution to challenges in long-term preservation of corpora, especially corpora that are continuously extended and subject to modification, e.g., due to legal injunctions, but also may overlap with respect to constituents, and may be subject to migrations to new data formats. We describe the contribution Signposts can make to the CLARIN infrastructure and document the design for the CMDI profile.

1 Introduction

The current paper presents an implementation of the concept of *signposts* (Arnold et al. 2020) which is based on the Component Metadata Infrastructure (CMDI, see Broeder et al. 2012), and explains how signposts can contribute to the overall CLARIN infrastructure. The contribution concerns the use of persistent identifiers (PIDs) for resources, and the handling of data removal, data migrations and versioning as well as deduplication.

A **signpost** is a metadata file for a leaf on the tree of resources, for instance a single text or an audio recording. Using terminology from the area of long-term archival, we distinguish *conceptual object* (CO) from *logical object* (LO) (see chapter 9 by Stefan Funk in Neuroth et al. 2009).[1] A CO can be realized in different LOs, for instance an audio recording (CO) can be realized in files of different audio formats (LO). **A signpost represents a conceptual object (CO), and also refers to logical objects (LOs,** typically at least one**) belonging to it.**

The most important point about signposts is that they change the idea what a PID refers to when providing data: While traditionally, PIDs may point to directly to data files (LOs),[2] it is suggested here that PIDs only refer to signposts (COs), and to leave it to signposts to point to files. The reason for this change is that LOs may be volatile, even if the represented information stays the same. By adding signposts as a layer of indirection, we can achieve an acceptable trade-off between the necessity of modifying data and the demands of long-term archival on the one hand and Open Science as well as reproducibility on the other.

Signposts are motivated with respect to the area of *growing corpora*, i.e. large corpora that are constantly extended and contain material where the conglomerate of commercial interests, intellectual property rights and privacy rights constitutes a non-trivial problem. However, all aspects signposts address are relevant to other kinds of corpora as well, generally to a different degree. In case of small and 'legally' permanent corpora, signpost information may be included in the corpus metadata. Signposts replace the concept of tombstones, which are less flexible than signposts (see Arnold et al. 2020).

2 Motivating Signposts

The motivation for signposts comes from the impermanence of logical objects, specifically three aspects: the necessity of *deletions* due to legal actions, (conceptual) *deduplication* and data *migration*.

Long-term Preservation vs. Legal Necessity Of Deletions. In the realm of long-term preservation, we assume that original data, in the sense of COs, will always be retained. However, e.g. when building corpora from newspapers, it may become necessary to remove data from COs due to injunctions or revocation of licenses.

[1] Funk (chapter 9 in Neuroth et al. 2009) also distinguish the level of *physical object* which, however, is not immediately relevant for our current discussion.

[2] PIDs are also used to refer to datasets (see, e.g., De Smedt, Koureas, and Wittenburg 2020 for a suggestion on how to structure datasets). However, we focus on data here.

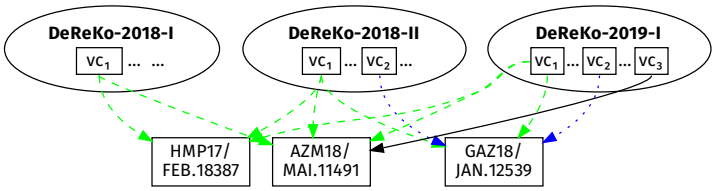

Figure 1: Relationship between DeReKo releases, virtual sub-corpora and texts (from Arnold et al. 2020). Texts may be part not of one, but many (virtual) corpora, and may belong to different versions of corpora.

Migration. File formats may fall out of use, so that data must be converted to new formats, which in the OAIS model is called *migration*. Anecdotally consider the German Reference Corpus (DeReKo, see, e.g. Kupietz et al. 2010) compiled at the Leibniz Institute for the German Language (IDS). Between 1999 and 2005, SGML (ISO8879:1986 1986) / CES were used as its data format, then DeReKo was converted to XML (for the history and the decisions involved, see Lüngen and Sperberg-McQueen 2012), based on the TEI's P3 and later P5 recommendations (Sperberg-McQueen and Burnard 1999; Burnard and Bauman 2020). Similar conversions occurred in the IDS' oral corpora. Even if we assume that we retain the original LOs, which goes beyond the OAIS model, we would want to add new ones as time progresses. For instance, we want to provide XML files conforming to P5 today rather than P3. It may then be a good idea to retire the intermediate versions to avoid storage cost. With the traditional approach to metadata, these changes mean that we have to change the metadata in each of these steps. With signposts, we only change the signpost.

Complex Corpus Structures. Especially growing corpora may have intricate structures, e.g. overlapping with respect to COs. If information were recorded in the metadata of the parent structures of the leaf COs, the metadata records would have to be changed for several corpora, while with signpost only the latter must be adapted. Figure 1 shows the relationships between the DeReKo corpus releases and virtual corpora $vc_{1,\ldots,3}$, and three texts. Based on release DeReKo-2018-I, vc_1 was defined,[3] already containing the texts HMP17/FEB.18387 and AZM18/MAI.11491. DeReKo-2018-II added GAZ18/JAN.12539 to vc_1. Based on DeReKo-2018-II, vc_2 was defined, containing the text GAZ18/JAN.12539. vc_3 was defined initially on DeReKo-2019-I, also containing AZM18/MAI.1149. This shows that texts in DeReKo may belong to many different corpora. In this case, removal becomes a complex matter.

In the next releases of both corpora in the IDS repository, we plan to implement signposts to avoid manually editing thousands of files in cases of conversion and legal issues.

3 A CMDI Profile for Signposts

Reusing existing CMDI components, we developed a metadata profile for signposts, the signpost profile has the identifier `clarin.eu:cr1:p_1587363818266`[4] in the component registry.

The CMDI profile reuses existing components and intends to include technical information that can be automatically extracted based on a file. This information can be gained by the File Information Tool Set (FITS)[5] or other readily available tools to facilitate processing. The collected information includes the original filename, media type, file size in bytes, various checksums and cryptographic hashes. Besides this basic technical information on a LO, the signpost should also include information on the status of each LO, i.e. whether this is the currently maintained version, whether the use of the LO is deprecated (for example in the case of a migration to other data formats) or whether a file is no longer available. Reasons can be given in the provenance information (see next subsection). The schema also allows referring to a LO by a specific name that is not its filename, which is occasionally necessary.

The nature of the signpost profile is different from other profiles, in the sense that it is not intended to provide a meaningful description of a resource and does not foster findability by search engines such as the VLO. Hence

[3] For the importance of virtual corpora in DeReKo's *primordial sample* design and extensionally or intensionally defined virtual corpora see Kupietz et al. (2010).

[4] see `https://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/1.1/profiles/clarin.eu:cr1:p_1587363818266/1.2/xsd`, which is still in the development state, but accessible

[5] See `http://fitstool.org`

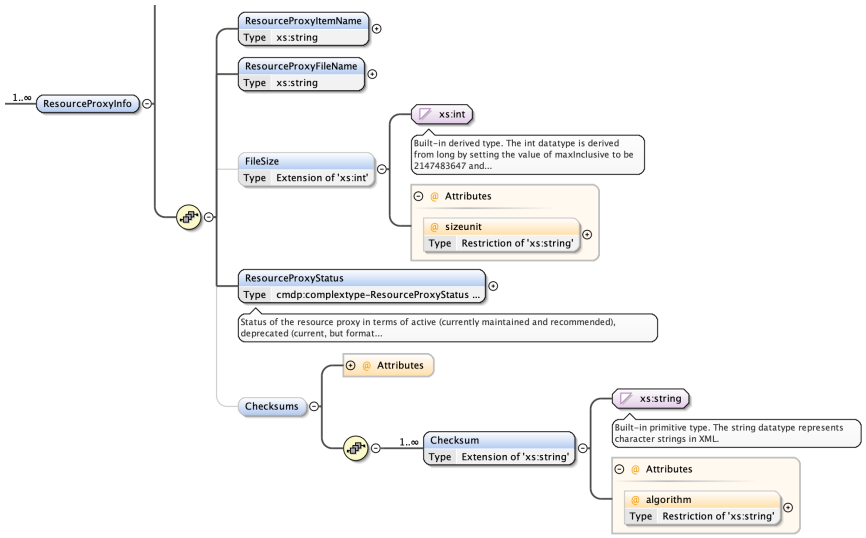

Figure 2: Visualisation of the structure for the metadata provided for each resource proxy in the CMDI Profile *signpost* (`clarin.eu:cr1:p_1587363818266`)

it does – by design – neither cater for the VLO facets, nor respect quality criteria which are automatically evaluated by the Curation Module[6]. As the media type is already provided in the resource proxy list of the CMDI file, there is no component including it. For each LO in the CMDI's resource proxy list, the profile allows the provision of various and multiple checksums, each specifying the algorithm in an attribute.

3.1 Provenance

The CMDI 1.2 specification (CMDI Taskforce 2016) implies that provenance information is not to be included in the CMDI file, but instead in one ore more separate file(s) called journal files. One reasoning behind this is that provenance information is not (necessarily) to be machine-interpretable and should be directed to human readers. We implement the journal file in HTML with microformats, as this approach allows formatting for human consumption by means of a web browser and aims at semantic interoperability. We include information in the journal file which is useful both for ensuring reproducibility of research and for keeping track of the development of a resource.

We include a log of every modification to the CO described by the signpost. These **changes** contain the following: creation, ingest, injunction, and migration. Moreover, all changes are dated with a **time stamp** and include a short human-readable **log message**. We suggest to also include modifications of LOs, i.e. **object changes**. Changes of the LO are marked as a **addition**, a **replacement**, or a **removal**. An `xml:id` (Marsh, Veillard, and Walsh 2005) attribute can be used in the signpost to identify LOs. This way, the log allows to determine the lifespan of a LO in a machine-readable way.

```
<h1>Log for <a href="http://PID-1">Conceptual Object
<code>http://PID-1</code></a></h1>
<ul class="sign_post_log">
  <li class="creation_entry">
    <span class="timestamp">2021-05-15T02:00:00+02:00</span> <span class="log_message">object created</span></li>
  <li class="ingest_entry">
    <span class="timestamp">2021-07-07T02:00:00+02:00</span> <span class="log_message">File ingested into IDS LTA</span>
    <ul class="object_changes">
      <li class="change_addition"><a href="http://PID-1#lo_1">Element</a> added</li></ul></li>
...
```

[6]https://curate.acdh.oeaw.ac.at/

3.2 PIDs

The usage of persistent identifiers differs significantly from the current usage in repositories. Traditionally, care is taken to assure that links to logical objects remain available and persistent; COs are not necessarily represented. We reverse this: Not the LO (file) but the CO (signpost) is primary. This means, only the signpost is granted a persistent identifier, and the access to logical objects is through URLs for which the archive does not give any guarantees. As this is currently unconventional, we must alert the user to the impermanence of LO URLs. We have considered the following strategies:

We can delegate implement a notice at the **presentation layer** of the repository: e.g. using a link text line *temporary download link*. This would inform human users, but is of no consequence for machine-processing of CMDI records. We suggest this is not a grave problem, for two reasons: First, as long as URLs for logical objects are not reused, tools relying on the `ResourceProxyList` and even caches will have no problems. This means that for a tool like the CMDI Explorer currently developed by CLARIN-D and the CLARIN ERIC, and which will recursively process chains of CMDI records, nothing changes, except there is one link more in the chain for each conceptual object. Secondly, we assume that by the time signposts are in widespread use, tool authors will have been made aware of the concept, and will take care not to download LOs blindly, but rely on signposts. It may be advantageous to integrate licensing information per LO, potentially in tandem with access control lists. Crawling of resources rather than metadata (including signposts), etc., can be prohibited in the `robots.txt`.

Alternatively or additionally, one could take care to generate temporary links to logical objects and hence force users to not rely on their URLs. We assume that this strategy generally wastes resources and should only be the last resort.

4 Conclusion

We proposed the notion of signpost for addressing data removal, migration and deduplication in long-term archival of resources, with a specific focus on growing corpora. We also presented an implementation of the concept in the CLARIN infrastructure. We welcome feedback on the concept and on the implementation. Future work will concern the adaptation of the format, and the integrations with tools, as outlined above.

Acknowledgements

The work reported here was funded by the German Federal Ministry of Education and Research (BMBF), the Ministry of Science, Research and Art of the Federal State of Baden-Württemberg (MWK), Project Management Agency German Aerospace Centre (DLR), and CLARIN-D.

We thank the anonymous reviewers for helpful comments that have allowed us to sharpen the text.

References

Arnold, Denis, Bernhard Fisseni, Paweł Kamocki, Oliver Schonefeld, Marc Kupietz, and Thomas Schmidt (2020). 'Addressing Cha(lle)nges in Long-Term Archiving of Large Corpora'. In: *Proceedings of the LREC 2020 Workshop 'Challenges in the Management of Large Corpora' (CMLC-8)*. Marseille, France.

Marsh, Jonathan, Daniel Veillard, and Norman Walsh (Sept. 2005). *xml:id Version 1.0*. W3C Recommendation TR xml-id. The World Wide Web Consortium. URL: https://www.w3.org/TR/xml-id/.

Broeder, Daan, Menzo Windhouwer, Dieter Van Uytvanck, Twan Goosen, and Thorsten Trippel (2012). 'CMDI: a component metadata infrastructure'. In: *Describing LRs with metadata: towards flexibility and interoperability in the documentation of LR workshop programme*. Vol. 1.

Burnard, Lou and Syd Bauman, eds. (2020). *Guidelines for Electronic Text Encoding and Interchange. TEI P5*. version 1.0.0 2007; latest release 4.0.0 on 2020-02-13. Chicago, New York: Text Encoding Initiative.

CMDI Taskforce (2016). *Component Metadata Infrastructure (CMDI): Component Metadata Specification. version 1.2*. Tech. rep. CLARIN ERIC. URL: https://office.clarin.eu/v/CE-2016-0880-CMDI_12_specification.pdf.

De Smedt, Koenraad, Dimitris Koureas, and Peter Wittenburg (2020). 'FAIR Digital Objects for Science: From Data Pieces to Actionable Knowledge Units'. In: *Publications* 8.2.

ISO8879:1986 (1986). *Information processing – Text and Office Systems – Standard Generalized Markup Language (SGML)*. Standard No. ISO 8879:1986. International Organization for Standardization.

Kupietz, Marc, Cyril Belica, Holger Keibel, and Andreas Witt (2010). 'The German Reference Corpus DeReKo: A Primordial Sample for Linguistic Research'. In: *Proceedings LREC'10*. Ed. by Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, Mike Rosner, and Daniel Tapias. Valletta/Paris: European Language Resources Association (ELRA), pp. 1848–1854.

Lüngen, Harald and Christopher Michael Sperberg-McQueen (2012). 'A TEI P5 Document Grammar for the IDS Text Model'. In: *Journal of the Text Encoding Initiative* 3, pp. 1–18. URL: http://jtei.revues.org/508.

Neuroth, Heike, Achim Oßwald, Regine Scheffel, Stefan Strathmann, and Mathias Jehn, eds. (2009). *nestor Handbuch. eine kleine Enzyklopädie der digitalen Langzeitarchivierung*. Version 2.0 [3/2010]. nestor.

Sperberg-McQueen, Christopher Michael and Lou Burnard, eds. (1999). *Guidelines for Electronic Text Encoding and Interchange. TEI P3*. initial release 1994-05-16; last version dated May 1999. Chicago, New York: Text Encoding Initiative.