Addressing Cha(lle)nges in Long-Term Archiving of Large Corpora

Denis Arnold, Bernhard Fisseni, Paweł Kamocki, Oliver Schonefeld,
Marc Kupietz, Thomas Schmidt
Leibniz-Institut für Deutsche Sprache
R5 6–13, 68161 Mannheim, Germany
{arnold | fisseni | kamocki | schonefeld | kupietz | thomas.schmidt} @ ids-mannheim.de

Abstract

This paper addresses long-term archival for large corpora. Three aspects specific to language resources are focused, namely (1) the removal of resources for legal reasons, (2) versioning of (unchanged) objects in constantly growing resources, especially where objects can be part of multiple releases but also part of different collections, and (3) the conversion of data to new formats for digital preservation. It is motivated why language resources may have to be changed, and why formats may need to be converted. As a solution, the use of an intermediate proxy object called a *signpost* is suggested. The approach will be exemplified with respect to the corpora of the Leibniz Institute for the German Language in Mannheim, namely the German Reference Corpus (DeReKo) and the Archive for Spoken German (AGD).

Keywords: long-term archival, legal issues, metadata, format migration

1. Introduction: Three Challenges

The current paper investigates long-term archival (LTA) for large corpora, specifically 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; we call them **growing corpora**. We focus on three aspects of archiving growing corpora which are related to changing resources, and in our opinion, can be approached by using tombstones or, preferably, signposts.

While we restrict our attention to growing corpora, all aspects apply to other kinds of corpora as well, but generally to a different degree. In the interest of space, we leave it to the attentive reader to judge the applicability.

Caron et al. (2017) discuss their solution in the context of the Open Archival Information System model (generally abbreviated: OAIS, see CCSDS, 2012; for an overview, see also the 4th chapter by Oßwald in Neuroth et al., 2009), and specifically the aspect of dealing with the ingest of a submission information package into the archive. In the OAIS, an *edition* is characterized by change in the content (e.g., Oßwald only speaks of additions), while a *version* is the result of a migration (cf. CCSDS, 2012, p. 1-9; § 5.1, esp. §5.1.3.4 Transformation). In this sense, while initially motivated by transparently dealing with editions, the current approach tries to integrate versions and editions into a common model.

Reproducibility in science in general and reuse of data in particular have been recognised as important goals over the last years, also connected to the adoption of the term and concept of *open science* (cf. for an overview Fecher and Friesike, 2014) and the publication and wide-spread adoption of the FAIR principles (Wilkinson et al., 2016). To maintain reproducibility of a documented scientific procedure, it also is necessary to maintain the access to some form of the data. This is in immediate conflict with the removal of data. Surprisingly, while the change of data regarding formats is generally considered in the context of preservation (cf. an example for research data Conway et al., 2011), we have found no example that considers the change of data sets due to legal issues.

One aspect that distinguishes growing corpora from others is that for legal reasons, it may be necessary to remove or modify part of the data. In long-term archival (see, e.g., Digital Preservation Coalition, 2015; Neuroth et al., 2009), this case is generally neglected, as it is normally assumed that data will stay unchanged. To remain as close to the spirit of LTA as possible, one will still want to deliver some useful information when someone tries to retrieve the removed objects. This is especially important since it is reasonable to assume that a researcher will not expect that data referenced with some form of persistent identifier has been altered. To our knowledge, only Caron et al. (2017) have considered the deletion of objects in LTA. They describe a ‘tombstone’ which steps in for objects that needed to be deleted for legal reasons. Systems like Fedora Commons or DSpace allow for removal of resources and provide tombstone objects.[1] We will discuss the differences between these and our approaches below.

Another aspect is parsimonious representation of data in corpora with many releases: Objects may be referenced in different releases and resources. Corpora that are curated in projects where the corpora are constantly extended and published in frequent releases will have many (unchanged) objects in common across different releases. Furthermore, an object may belong to different collections. Generally, a digital long-term archive will avoid storing the same digital objects multiple times. Keeping only one copy per object ensures that there is no confusion about the state of an object and storage space is not wasted, especially when objects are considerably large.

The third and last aspect is specific to long-time preservation: It is unforeseeable if and when a given file format may become deprecated. But once this is the case, the archive will have to migrate the respective files to the new format and make them accessible along with the original files. This can be seen as a departure from the original model, which states “[…] that the new archival implementation of the information is a replacement for the old” (CCSDS, 2012, p. 1-11).

To explain the proposal, we distinguish the notions of *conceptual object* (CO) and *logical object* (LO) (see chapter 9 by

[1] https://wiki.lyrasis.org/display/FEDORA4x/RESTful+HTTP+API; https://wiki.lyrasis.org/display/DSDOC6x/Functional+Overview

Proceedings of he 8th Workshop on Challenges in the Management of Large Corpora, pages 1–9
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Stefan Funk in Neuroth et al., 2009).[2] A CO can be realized in different LOs, for instance an audio recording (CO) can be realized in files of different audio formats (LO).

The first two cases, i.e. removal and and versioning primarily concern changes of conceptual objects – although the changes will be mirrored in LOs –, while the case of format conversion only concerns logical objects. This observation will form the basis of our technical proposal.

1.1. Background

The Leibniz Institute for the German Language (IDS) is building up a long-term archiving repository for linguistic data. Current work is focusing on the development of appropriate ways of ingesting the IDS's own corpora of written and spoken language. Both can be viewed as exemplars of large and growing corpora: The German Reference Corpus DeReKo (Kupietz et al., 2010, 2018) has been built at IDS since its foundation in the mid-60s (Teubert and Belica, 2014). It currently contains 46.9 billion tokens (Leibniz-Institut für Deutsche Sprache, 2020) corresponding to 56 GB disk space (without automatic annotations) and is used by more than 40 000 German linguists world-wide, primarily via specialized analysis platforms, such as COSMAS II (Bodmer, 2005) and KorAP (Bański et al., 2013). The Archive for Spoken German (AGD; Schmidt, 2017) hosts around 80 corpora of spoken language. The digital available corpora are published via the *Datenbank für Gesprochenes Deutsch* (Schmidt and Gasch, 2019). While much smaller than their written counterparts in terms of number of documents or tokens of (transcribed) text, they can also be viewed as large corpora because they comprise large digital audio or video files. For example, the 1 113 recordings of the BOLSA study (Lehr and Thomae, 1987) make up for altogether 2 833 hours of audio. Stored as mono WAV files with a sampling rate of 48 kHz, they occupy around 1 TB of disk space. For the latest version of the FOLK corpus (Schmidt, 2016), the textual data amounts to around 2.5 million transcribed tokens (less than 0.5 GB), whereas the archived media data (stereo WAV, 48 kHz for audio, MPEG-4 in a resolution of 1 980 × 1 080 for video) is also around 1 TB.

1.2. Legal Aspects

There are several possible scenarios where parts of large corpora intended for long-term archiving have to be deleted for legal reasons. Three legal frameworks seem to be of particular relevance here: intellectual property (IP), data protection and criminal law.

Intellectual Property

Firstly, concerning IP, it is important to note that language data are, for the most part, protected by copyright. As such, their use (i.e. reproduction and communication to the public) is lawful only in one of two cases: (1) a permission (license) has been obtained from the right holder or (2) the use is covered by a statutory exception or limitation (e.g. for teaching and research). In both cases, long-term archiving may be impacted.

A license can be granted for limited duration only, and once it comes to its term, the work can no longer be lawfully used. Technically speaking, it does not have to be deleted, but any further copying (even in a computer's memory) would amount to copyright infringement. Although from the user's point of view it is advantageous to use licenses that are not limited in time (or, rather, are granted for the whole duration of copyright), such as Creative Commons licenses, right holders cannot be forced to grant them. In practice, it is usually the case that the longer the licence period, the less likely it is that the licence will be granted, or higher fees may be charged, or both (see Kupietz et al., 2014, for a more detailed discussion of the trade-offs). In the case of DeReKo, the typical duration of commercial licenses is one year, while donated licenses are almost always unlimited in time.

A license can also be revoked by the licensor – usually because the right of revocation is specifically provided for in the license itself, in which case, of course, the stipulated modalities (e.g. prior notice) would have to be respected. Creative Commons licenses, for example, terminate automatically upon any breach of the license's terms. It is also possible, albeit in very limited cases, that the statute grants the right holder the right of revocation (i.e. the license can be revoked even if it does not stipulate so). Under German law, for example, an exclusive license can be revoked if the work is not used by the licensee (§41 UrhG[3]). More interestingly, still under German law, the author has the right to revoke a license 'for changed conviction' (§42 UrhG), i.e. when he decides that the work no longer reflects his conviction. In this case, the author has to adequately compensate the licensee for the revocation, and if in the future he decides to use the work again, he shall grant a new license to the licensee 'on reasonable conditions'. Exceptionally, the right of revocation for changed conviction can also be exercised by the author's heirs; then, however, it would require a proof that the author would have exercised exercised the right prior to his death. Upon revocation, whether on contractual or statutory grounds, the licensed data can no longer be lawfully used.

In the case of DeReKo, many licenses are explicitly revocable at any time with a period of a few weeks. Since the 2000s at the latest, a corresponding addition to the license conditions has often proved necessary in order to be able to conclude license agreements at all and in a reasonable time. So far, however, no licensor has made use of his right of revocation.

Statutory exceptions may seem to provide for a more stable ground for long-term archiving, but it is not always the case. It should be kept in mind that the exception may simply not allow for long-term archiving (such as the current data mining exception in German law – §60d UrhG), and even if it does, it may simply cease to apply at some point, or be replaced by a different, stricter norm (even though in the past decade or two the trend seems to be towards broadening the scope of statutory exceptions). Moreover, albeit very rarely, an exception may come with an 'expiry date' – this is the case of exceptions introduced in German law by the UrhWissG[4] (covering such uses as teaching, research, data mining and uses made by libraries), which will cease to apply at the end of February 2023 (although they are expected to be either maintained by the legislator, or replaced by other similar exceptions. Therefore, when long-term

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

[3] Act on Copyright and Related Rights (Gesetz über Urheberrecht und verwandte Schutzrechte (Urheberrechtsgesetz, UrhG)

[4] Gesetz zur Angleichung des Urheberrechts an die aktuellen Erfordernisse der Wissensgesellschaft

archiving is based on statutory exceptions, it is of the essence to stay informed about the developments in the legal framework, and adjust the archiving policy accordingly.

Data Protection

Another legal framework that crucially impacts long-term archiving is data protection; the most important source of data protection law is the famous General Data Protection Regulation (GDPR) which entered into application in the European Union on 25 May 2018.

First of all, it should be mentioned that when the corpus contains personal data (which is not unlikely to happen, taking into account the large scope of the notion), archiving for an undefined (and potentially unlimited) duration may simply not be an option. Storage limitation (the principle according to which personal data can be kept for longer than necessary to achieve the purposes of the processing – Article 5(1)(e) of the GDPR) is one of the fundamental principles governing the processing of personal data under the GDPR. However, the GDPR allows for derogations from this principle when the processing is carried out solely 'for archiving purposes in the public interest', or for research or statistical purposes (Article 89 of the GDPR). To be able to qualify for the derogation, however, the processing has to be subject to 'appropriate safeguards' such as e.g. pseudonymization. The rules regarding these purposes of processing remain largely country-specific – in Germany, at the federal level, processing for research purposes is governed by §27 of the BDSG[5], archiving in the public interest by §28, and 'appropriate safeguards' are listed in §22 of the same Act.

Even if the storage limitation principle with its derogations is observed, some data may still have to be deleted on the grounds of data protection. When the processing is based on consent of the data subject (which, alongside 'legitimate interest' is probably the most common ground for the processing of personal data in language corpora), the consent can be withdrawn at any time (Article 7(3) of the GDPR). The withdrawal has no retroactive effect (i.e. the processing based on consent prior to its withdrawal does not 'become unlawful'), but any further processing should stop (although it is possible to resume processing on a different ground, e.g. based on legitimate interest).

However, if the data is processed on the ground of legitimate interest, the data subject may still exercise the right to object (Article 21 of the GDPR), in which case the processing should stop, unless the controller (the person or entity who defines the means and purposes of the processing) demonstrates 'compelling legitimate grounds' for the processing which override the interests, rights and freedoms of the data subject. The right to object does not apply to processing for research purposes, or for the purpose of 'archiving in the public interest' (Article 89 of the GDPR, §§ 27–28 of the BDSG).

Finally, the data subject may also exercise his right of erasure (Article 17 of the GDPR), commonly referred to as 'the right to be forgotten'. This right is not limited even when the processing is carried out for research or archiving purposes (unless it 'seriously impairs' these purposes); however, perhaps contrary to the common belief, the conditions for exercising this right are in fact very strict, and in practice seem to require some sort of prior violation of the GDPR on behalf of the controller. Perhaps most importantly, the data subject may request erasure if the data minimisation principle has been violated, i.e. the data are no longer necessary to achieve the purpose of the processing. This ground can be successfully used while requesting erasure e.g. from search engines, and possibly also from large, publicly accessible corpora. It is worth noting that if the request for erasure is well-founded, the controller shall also make reasonable steps (including technical measures) to inform other controllers who process the same information, so that they can also proceed with the erasure.

Criminal Law

Last but not least, criminal law, and more specifically the rules regarding defamation and defamation-related offences (slander, libel, insult…), may also require deletion of some parts of large corpora. This is especially relevant for newspapers or other press materials.

Since national rules may vary significantly (there is no harmonised law of defamation at the EU level), we will use § 186 of the German Criminal Code for illustration purposes. The text provides that whoever disseminates a fact related to another person which may defame him or negatively affect public opinion about him, is punished with a fine or imprisonment for up to one year (when the offence is committed by dissemination of written materials, the penalty increases to two years). Apart from that, the claimant may also obtain an injunction (i.e. a court order for the defendant to stop disseminating the material), even preliminary (i.e. applicable even before the final decision on the merits of the case is made by the court). Needless to say, a defamation claim, even ill-founded may lead to (at least temporary) deletion of parts of long-term archived corpora.

In the case of DeReKo, injunctions are by far the most common reason for the removal of individual texts, with about two incidents per week. The obligation to remove the texts is also stipulated in the license agreements with the right holders. There is a consensus within the German linguistic community that the removal of individual texts is unavoidable and typically irrelevant with respect to linguistic findings and should not pose an insoluble problem in terms of reproducibility of research results, which is reflected in the guidelines on legal aspects of handling corpora of the German Research Foundation DFG (Deutsche Forschungsgemeinschaft, 2015, p. 19), quoted here from its English translation:

> Regarding the still existing problem of persistence of research data, there is a certain pragmatic consensus within the scientific community: text deletions because of personality rights should be considered acceptable also epistemologically, since the replicability of important and methodically valid research results does not depend on individual texts. What is probably more important is de facto the organizational effort that can be caused by individual deletions. It is recommended to factor this into project costs in advance, if possible. (Wildgans et al., 2017, p. 20)

The three frameworks presented above may to a limited extent be derogated from by laws on public archives, such as the Bundesarchivgesetz or Landesarchivgesetze in Germany (so-called *Löschungssurrogat*), or Code du patrimoine in France.

[5] German Federal Data Protection Act (Bundesdatenschutzgesetz vom 30. Juni 2017 (BGBl. I S. 2097))

However, apart from them being heavily country-specific, their application is usually limited to 'official' registries or other documents of key importance for public administration, or to archiving by specifically designated institutions. Therefore, it is our opinion that the relevance of such laws on public archives for long-term archiving of language corpora is very limited and so they fall outside of the scope of this paper.

1.3. Versioning of COs

There are two cases in which COs change, only one of which constitutes a veritable challenge. The other one can be seen as unproblematic.

The unproblematic case arises when a resource is published in a new version, e.g. containing more annotations, but also correcting mistakes that will stay accessible in the previous version. The long-term archive will in this case simply issue a new version of the CO and the old version stays intact and accessible. There is a certain conflict of interests here: On the one hand, it may be interesting to users to see that there is a new version of an CO; on the other hand, integrating this information into the archive would most evidently be possible changing the metadata, a measure which evidently goes against the general guarantees of long-term archival. We suggest that in this case the latter point far outweighs the former: It is not necessary to point to the new version in the long-term archive and make changes to metadata. However, it is by perfectly admissible from an LTA perspective to point to old versions from the new ones – as long as the latter are archived after the former have been, as is normally the case. To improve usability, the presentation layer of the archive can invert these links without integrating them into the metadata proper.

The interesting case is when parts of a corpus must be altered; this generally occurs due to legal reasons: An alternative version of the respective CO and its LOs must be created, or the CO must even be deleted completely. In any case, the old version will no longer be accessible.

Caron et al.'s focus are single packages of digitised documents (images, text files) rather than growing, hierarchical corpora and partial modifications. Based on the OAIS terminology explained in our first section, they determine whether there is a new version or a new edition as follows (see their figure 3): Disregarding ingest failures, a modification or deletion of data (in their case, e.g., improved imaging) leads to a new version, while additional content or modification of metadata leads to a new edition.

1.4. Pointing to the Converted

Finally, it may happen that a certain file format falls out of use. For instance, in the area of video formats, Apple has retired its QuickTime format in macOS 10.15 (Catalina). In the area of text annotation, SGML (ISO8879:1986, 1986) has given way to XML (Bray et al., 1997). To anecdotically trace one of our migration paths: DeReKo used SGML/CES between 1999 and 2005, and was consequently converted to an XML-based format (for the history and the decisions involved, see Lüngen and Sperberg-McQueen, 2012), first based on the TEI's P3 recommendations (Sperberg-McQueen and Burnard, 1999), later converted to TEI P5 (Burnard and Bauman, 2020), the customization of the annotation is called I5. Similarly, in the area of spoken language annotation, the IDS is in the process of switching to the new ISO standard ISO-24624:2016 (ISO, 2016) for new annotations, s, and has also converted old annotations from a variety of formats including SGML, HTML and plain text. Also, the tools developed at the IDS, especially the EXMARaLDA family (Schmidt and Wörner, 2014)) are being adapted to work with this format.

One can argue that the older SGML and XML formats are still usable, but the modern formats provide much better interoperability at the current time. So while formats used in long-term archival are generally selected to minimize the chance of complete obsolescence or loss of readability, it may still be preferable to provide additional formats that are more readily supported by contemporary software. In this case, the original LO is not completely replaced, but it may be preferable to deliver an object in another format if one queries for the CO. Again it is inconvenient and misleading to modify the metadata of parent resources, as conceptually, not the CO but only the files realizing it have changed.

An important question is reversibility of transformations. For formats like video, where we store lossily compressed data, a transformation would probably not be reversible (CCSDS, 2012, p. 5-6f), as transcoding would introduce new compression COs. In case of the conversion of DeReKo from SGML to XML (cf. Lüngen and Sperberg-McQueen, 2012), however, the migration was reversible in the sense that all old data could be losslessly translated back. I5 has evolved further (as of this writing, the latest version is from 2020-03-05)[6] and accommodates features for new kinds of text (such as computer-mediated communication), new data may not be retranslatable to the old format. This effect of the migration path illustrates a further complication regarding growing corpora.

With respect to the OAIS model, we can model this as three cases: (1) First, the conversion is merely a change of the Dissemination Information Package (DIP), or (2) alternatively, it may constitute a migration, namely a a transformation (CCSDS, 2012, §5.1.3.4). By keeping the original data, we partly transcend the OAIS model.

2. Signposts: Dealing with Modified or Deleted Data in a Transparent Way

2.1. Example: Removing / Modifying Data

In DeReKo, the structure of the corpus has three levels: *corpus* (i.e., subcorpus), *document* and ultimately *text*. What the corpus and document levels correspond to, depends on the text type. For newspapers, for example, a year volume corresponds to a 'corpus' and a month to a 'document'. It is in newspaper and magazine documents that a removal may occur due to injunctions for privacy reasons. We have not yet had a case where a whole volume had to be removed.

For reasons of work effort, we have to retract the whole corpus release archive in case an injunction occurs. The next release archive of DeReKo, however, will contain a modified version of the document: We generally remove the body of a *text*, marking it as a gap in the XML annotation. (For technical reasons, it is necessary to have a `<div>` element after the `<gap>`.) Often, it is possible to keep the title of an article if it does not give away personal information.

[6] Information on the current state of the format can be found at https://www.ids-mannheim.de/kl/projekte/korpora/textmodell.html

```
<text>
    <body>
        <gap reason="injunction"/>
        <div type=""/>
    </body>
</text>
```

Figure 1 – An XML tombstone in IDS's I5 format, which is a selection of the TEI P5 recommendations. An injunction leads to a <gap> in the document.

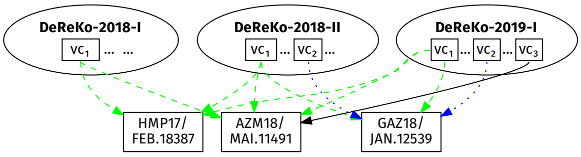

Figure 2 – Visualisation of the relationship between DeReKo releases, virtual sub-corpora and texts. DeReKo contains texts that are not only part of one corpus, but many (virtual) corpora. Considering the versioning, we find that text can be part of many different versions of many corpora.

2.2. Example: Versioning

Figure 2 shows the relationships between the DeReKo corpus releases DeReKo-2018-I – DeReKo-2019-I, three persistent virtual corpora $vc_{1,...,3}$, respectively initially defined on one of the releases, and three texts.[7] Based on DeReKo-2018-I, vc_1 was intensionally defined, already containing the texts HMP17/FEB.18387 and AZM18/MAI.11491. With DeReKo-2018-II, GAZ18/JAN.12539 was added to vc_1 because the text matches the intensional definition of vc_1. In addition, based on DeReKo-2018-II, vc_2 was defined, containing the text GAZ18/JAN.12539. Based on DeReKo-2019-I, then vc_3 was added, containing AZM18/MAI.1149. You can see here that texts in DeReKo can belong to many different corpora so that the removal of texts can have complex consequences.

2.3. General Discussion

Growing corpora are generally structured hierarchically, consisting of several subcorpora. The general approach is to model this as a containment relation, where the record of a parent resource refers directly to its constituent objects and also indicates specific information on the data, such as the file type. As Broeder et al. (2012) point out, the metadata should specify the MIME type, file size and potentially checksums, etc. In case of data removal, it is then possible to either modify the parent resource or to replace the object with a 'tombstone' which indicates removal of the original data (see fig. 3). This is what is suggested by Caron et al. (2017), but also implemented in systems like DSpace and Fedora with their tombstone features.[8]

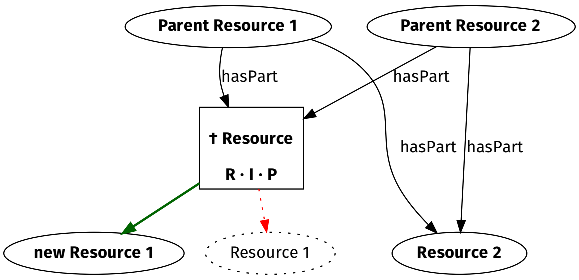

Figure 3 – A tombstone replaces a resource that has departed.

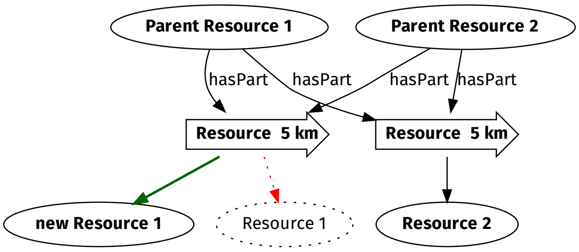

Figure 4 – A signpost just points to any resource – dead or alive!

However, this approach has several drawbacks: If it was possible before to address the removed object, e.g. by using a persistent identifier (PID), then only modifying the parent resource is insufficient, as it leaves the respective PID dangling. Therefore we now consider how to successfully implement the latter approach, namely the tombstones: We take it for granted that a tombstone object should be machine-readable and discernible as a tombstone rather than mere different data. In the general case, a replaced object may have represented an arbitrary kind of data, such as audio or audiovisual recordings or textual data. It is then evident that the tombstone object will not be of the same type as the replaced object. Hence, removal of objects has the effect that all metadata referring to the objects have to be modified as well by updating all information relating to the LO. Especially in the domain of growing corpora, this often constitutes a non-trivial change, and it again violates the precept of long-term archival that data will not be modified.

We suggest here that by introducing one layer of indirection, we can minimize modifications by introducing an intermediate object. It is customary to take the LOs which realize an CO to be the constituents of collections, and also to be referenced in metadata. We suggest that instead of the digital object, the CO be considered the referent of persistent identifiers, and its representation functions as a proxy or signpost. We use *CO* as an ontological category, and speak of *signpost* when we refer to a digital representation in a repository system. The general idea (see fig. 4) is to defer the specification of information on LOs representing a constituent CO to the signpost of the CO rather than to the record of the parent resource. The parent resource only contains general information on the CO and refers to the signpost. While the signpost can turn into a dead-end – analogously to a tombstone –, all data referring to it remain unchanged, which is advantageous in all cases mentioned above.

[7] Note that virtual corpora are a key concept of DeReKo's *primordial sample* design (Kupietz et al., 2010). They can be defined extensionally by a list of corpus, document or text IDs or intensionally, for example as *All available texts originating from "Der Spiegel" since the year 2000.*

[8] https://wiki.lyrasis.org/display/FEDORA4x/RESTful+HTTP+API; https://wiki.lyrasis.org/display/DSDOC6x/Functional+Overview

3. Signposts: Outsourcing the Object Metadata into a Separate Entity

If all metadata of parent resources only refers to the signpost, the metadata need not be modified if a resource becomes unavailable, or new formats are added.

As indicated above, there are consequences for metadata for parent resources, however, compared to the traditional approach: These metadata may only contain very general information on an CO then, e.g. whether it contains audio, audiovisual or textual data, but not whether it is a WAV file, an MPEG4 file with H.265 with AAC audio etc. If the metadata were this specific, a modification of the signpost would not be possible, and it would ultimately not offer added value.

The following features should be encoded in a signpost; we will make this more concrete in section 3.2.

the overall state: whether or not the CO is available.

the files: the files realizing the same CO; the description should include a MIME type, check sums and all other data that identifies the files.

a change log: the changes should be logged. To facilitate automatic processing of these files, the log entries must minimally specify a timestamp for the change, whether the change was a removal or an addition of files (we will address additions presently under the heading of conversions).

the next best version: In case of removal (or modification) of an CO, another version may be referenced, which comes as close to the modified CO as possible, in the optimal case only including legally required modifications, but if it is too costly to produce such a new adaptation of old COs, potentially also new additional data or corrections. For instance, in DeReKo, it may be that a certain year of a journal was first only included partially, but later was both cleared of illegal data and completed.

3.1. Delivering Signposts

We assume that delivery and processing of signposts are handled as outlined below. We would welcome a discussion of this approach.

In case an **unavailable CO** is retrieved, an HTTP 404 error is signalled, and signpost data is used to generate an error message that explains the situation and points the user to the next best version, if possible. As there are different use cases for the signpost, we suggest that besides delivering the signpost data or an error page as just described, the following convenience functions may be useful.

First, signposts may be used by human readers to access the CO. In this case a readable and human-friendly version of the information in the signpost may be presented. From there, access to the LO(s) will be possible. These will, however, under no circumstances be available via a persistent identifier, and there will be no guarantee that linking to them directly will have reliable result.

Secondly, a signpost may be used to automatically retrieve data. The focus of resource delivery should then be on machine-readability, namely as follows. If no additional requirements are signalled, the signpost file will be delivered, and an HTTP status of 300 *Multiple Choices* will be indicated, and a preferred choice will be signalled in the header. There is no long-term guarantee as to which format this will have, only that it is the format that, according to the expectations of the archive, suits the interest of a general user best. In case the user has an interest in retrieving specific file types, this may be signalled by asking for specific MIME types. If possible, the CO will be delivered in an object corresponding to the MIME type. If not, an HTTP 404 error will be signalled, and the signpost information will be used to generate a useful directions to get other realizations of the CO.

3.2. An Abstract Data Format for Signposts

In this section we present an abstract format for signposts. We also give a minimalist implementation of this format. We do not give an implementation in some existing format like CMDI, because this is a conceptual paper and we assume that the structure will be adapted after discussion in the community.

Remember that signposts are to be processed automatically as much as possible.

For a given CO, e.g. an audio recording or a transcription, the following is require (a simple XML grammar is provided at the end of the paper).

PID: the persistent identifier pointing to the CO, i.e. normally the URL of the signpost.

Pointers to LOs, e.g. to the audio file (or files of different formats) or to the transcription file (or the files of different formats), each pointer consisting of the following information:

- **State:** Every LO is either `"active"` or `"retired"`.
- **Creation and**, if applicable, **Retirement Dates:** It is thus reproducible what files were available at the time.
- **LO URL:** for retrieval
- **Format or MIME type:** to assure adequate processing
- **Information on the LOs:** like **size** or **check sums**

Log of Events in which the conceptual object was created and altered. This allows for reconstruction of availability and contributes to checking reproducibility.

- **Date** of the event.
- **Type of Change:** For conceptual objects, it can be seen when they were removed from the archive, and types of reasons are given using a closed vocabulary; we currently assume that creation, ingest, injunction and migration are sufficient.
- **Comment:** It is also possible to give more information in human-readable form.

Pointer to the Next Best Version For conceptual objects that are no longer available, a `<surrogate>` is presented which is only pointed at with a PID. This allows, theoretically, to chain signposts. While excessive use of this feature is not desirable, it is still a useful property in case injunctions are filed at greater temporal distance.

The first example points to a conceptual object like in the versioning example above. We assume we are in the year 2138. Let us assume that there are several metadata records pointing to our conceptual object, e.g. those of vc_4 and vc_5, as it may be part of different greater units. More importantly, the object was transcoded from the original MP4 Audio format to MP7 in 2028 and again, a hundred years later, to MP27. At the latter migration, the MP7 file was retired, as it is not the original and MP27 captures all significant properties of MP7 files. (The original MP4 file was kept, following the preservation policy of the IDS.) Programs that no longer can process the outdated MP4 audio can see that the object is available as MP27 as well and retrieve it. (As discussed above, implicit smartness can also be implemented.)

```
<?xml version="1.0" encoding="utf-8"?>
<signpost>
  <identity pid="http://PID-1"/>
  <logical-objects>
    <logical-object state="active"
      url="https://REPO/PATH/RECORDING-1"
      mime-type="application/mp4"
      creation-date="2021-07-07T02:00:00+02:00"
      byte-size="123456">
      <!-- check sum as element to allow different,
           non-hardcoded types -->
      <check-sum type="SHA-512" value="402550..."/>
    </logical-object>
    <logical-object
      url="https://REPO/PATH/RECORDING-1?format=mp7"
      mime-type="application/mp7"
      creation-date="2028-05-15T02:00:00+02:00"
      state="retired" retirement-date="2128-05-15
           T02:00:00+02:00"
      byte-size="23456">
      <check-sum type="SHA-512" value="31324..."/>
    </logical-object>
    <logical-object state="active"
      url="https://REPO/PATH/RECORDING-1?format=mp27"
      mime-type="application/mp27"
      creation-date="2128-05-15T02:00:00+02:00"
      byte-size="6789">
      <check-sum type="SHA-512" value="7a8b5a..."/>
    </logical-object>
  </logical-objects>
  <change-log>
    <entry date="2021-05-15T02:00:00+02:00"
      type="creation">File created</entry>
    <entry date="2021-07-07T02:00:00+02:00"
      type="ingest">File ingested into IDS LTA</entry>
    <entry date="2028-05-15T02:00:00+02:00"
      type="migration">converted to MP7</entry>
    <entry date="2128-05-15T02:00:00+02:00"
      type="migration">converted to MP27</entry>
  </change-log>
</signpost>
```

The second example may represent DeReKo data. Assume this is the 2020 volume of the *Postkutschenbote*. It was not yet completely digitized when it was ingested. Then an injunction was filed, making it necessary to remove the CO. For reasons of work economy, the reader is referred to a new edition (in the OAIS sense) of the work, which may already contain the full 2020 volume. As the CO as a whole is retired, all LOs have been, as well.

```
<?xml version="1.0" encoding="utf-8"?>
<signpost>
  <identity pid="http://PID-DEREKO-EXAMPLE-1-e1"/>
  <logical-objects>
    <logical-object url="https://REPO/PATH/NEWS-1"
      mime-type="application/tei+xml"
      creation-date="2021-07-07T13:37:23+02:00"
      state="retired"
      retirement-date="2021-08-08T13:37:23+12:05"
      byte-size="123456">
      <check-sum type="SHA-512" value="31324..."/>
    </logical-object>
  </logical-objects>
  <surrogate pid="http://PID-DEREKO-EXAMPLE-1-e2"
    type="edition">This version contains all original
         data except for the ones removed due to an
         injunction, and potentially more data.
  </surrogate>
  <change-log>
    <entry date="2021-01-01T13:37:23+02:00"
      type="creation">File created</entry>
    <entry date="2021-07-07T13:37:23+02:00"
      type="ingest">File ingested into IDS LTA</entry>
    <entry date="2021-08-08T13:37:23+12:05"
      type="injunction">File removed due to an
           injunction</entry>
  </change-log>
</signpost>
```

The third example concerns DeReKo data. Assume this is the 2019 volume of the *Mannheimer Spezielle Zeitung*. It was ingested, but an injunction was filed. As this hypothetical newspaper is one of the most-read in Germany and is particularly loved and used by corpus linguists for word usage statistics, a new version of this object was prepared, which is as close to the original data as possible. This should help maintain reproducibility as much as legally possible. On a terminological note, the surrogate does not constitute an OAIS version in this case, as the process is not the result of a migration. We still think that intuitive meaning of the term *version* comes closest to what we need here.

```
<?xml version="1.0" encoding="utf-8"?>
<signpost>
  <identity pid="http://PID-DEREKO-EXAMPLE-2-e1-v1"/>
  <logical-objects>
    <logical-object url="https://REPO/PATH/NEWS-2"
      mime-type="application/tei+xml"
      creation-date="2020-07-07T13:38:24+02:00"
      state="retired"
      retirement-date="2020-08-08T13:38:24+02:00"
      byte-size="123456">
      <check-sum type="SHA-512" value="31324..."/>
    </logical-object>
  </logical-objects>
  <surrogate pid="http://PID-DEREKO-EXAMPLE-2-e1-v2"
    type="version">This version contains all original
         data except for the ones removed due to an
         injunction.
  </surrogate>
  <change-log>
    <entry date="2020-01-01T13:38:24+01:00"
      type="creation">File created</entry>
    <entry date="2020-07-07T13:38:24+02:00"
      type="ingest">File ingested into IDS LTA</entry>
    <entry date="2020-08-08T13:38:24+02:00"
      type="injunction">File adapted due to an
           injunction</entry>
  </change-log>
</signpost>
```

4. Conclusion and Outlook

We assume that the concept of signpost is useful to address the problems of unavoidable data change in LTA, versioning of growing corpora and data migration as sketched in this paper. We discussed theoretical points and illustrated the use of signposts with concrete, if partly fictional examples.

Some details of the proposal should be discussed further, for instance:

Does the signpost need something like a title or a short human-readable summary of the conceptual object's place in the corpus? We decided against this in the examples we presented, as it cannot trivially be generated automatically.

Is it necessary to keep the information on logical objects that have been removed, especially in the case the conceptual object is no longer available? An argument in favour is that this may help to ensure reproducibility; this would only be useful, though, if there were standardized procedures for citing logical objects that include, e.g., the file checksums used in the signpost.

How adequate is the assumption that a presentation layer complements the metadata? We suggested above that pointers to later versions of an object can be implemented in the presentation layer to avoid adjustment of metadata; however, this conflates data modelling and presentation and hence introduces new challenges to data repositories.

Moreover, it may be useful to implement the signpost format in a way more compatible with established metadata standards, for instance CMDI (Broeder et al., 2012), or to define the vocabulary in a formal way such as using Semantic Web technologies like RDF(S) (see, e.g., McBride, 2003).

5. Acknowledgements

We thank Thorsten Trippel (SFS, University of Tübingen) for insightful discussions on metadata.

A Simple SignpostML grammar

A simple RelaxNG (Clark and Murata, 2001) grammar (using compact syntax) is provided below. The documents above are valid against this grammar.

```
start = element signpost {
  element identity { attribute pid { xsd:anyURI } },
  (AliveObject | DeadObject),
  element change-log {
    element entry {
      attribute date { xsd:dateTime },
      attribute type {
        "creation" | "ingest" |
        "injunction" | "migration" },
      text
    }+
  }
}

AliveObject = element logical-objects {
  AliveLO, (DeadLO*, AliveLO*)*
}

DeadObject =
element logical-objects { DeadLO+ },
element surrogate {
  attribute pid { xsd:anyURI },
  attribute type { "edition" | "version" },
  text
} ?

LOParts = attribute url { xsd:anyURI },
  attribute creation-date { xsd:dateTime },
  attribute mime-type { text },
  attribute byte-size { xsd:integer },
  element check-sum {
    attribute type { "SHA-512" },
    attribute value { text } }+

DeadLOAttributes = attribute state { "retired" },
  attribute retirement-date { xsd:dateTime }?

AliveLOAttributes = attribute state { "active" }

DeadLO = element logical-object {
  DeadLOAttributes, LOParts }

AliveLO = element logical-object {
  AliveLOAttributes, LOParts }
```

6. Bibliographical References

Bański, P., Bingel, J., Diewald, N., Frick, E., Hanl, M., Kupietz, M., Pezik, P., Schnober, C., and Witt, A. (2013). KorAP: the new corpus analysis platform at IDS Mannheim. In Vetulani, Z. and Uszkoreit, H., editors, *Human Language Technologies as a Challenge for Computer Science and Linguistics. Proceedings of the 6th Language and Technology Conference*, Poznań. Fundacja Uniwersytetu im. A. Mickiewicza.

Bodmer, F. (2005). COSMAS II. Recherchieren in den Korpora des IDS. *Sprachreport*, 3/2005:2–5.

Bray, T., Paoli, J., and Sperberg-McQueen, C. M. (1997). Extensible markup language XML. W3C Recommendation TR-XML, The World Wide Web Consortium.

Broeder, D., Windhouwer, M., Van Uytvanck, D., Goosen, T., and Trippel, T. (2012). CMDI: a component metadata infrastructure. In *Describing LRs with metadata: towards flexibility and interoperability in the documentation of LR workshop programme*, volume 1.

Burnard, L. and Bauman, S., editors (2020). *Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Initiative, Chicago, New York. version 1.0.0 2007; latest release 4.0.0 on 2020-02-13.

Caron, B., De La Houssaye, J., Ledoux, T., and Reecht, S. (2017). Life and death of an information package: Implementing the lifecycle in a multi-purpose preservation system. In *iPRES 2017 14th International Conference on Digital Preservation*, Kyōto, Japan.

CCSDS (2012). *Reference model for an open archival information system (OAIS)*. CCSDS, Washington, 2 edition.

Clark, J. and Murata, M., editors (2001). *RELAX NG Specification*. Organization for the Advancement of Structured Information Standards.

Conway, E., Giaretta, D., Lambert, S., and Matthews, B. (2011). Curating scientific research data for the long term: a preservation analysis method in context. *The International Journal of Digital Curation*, 6(2).

Deutsche Forschungsgemeinschaft (2015). Handreichung: Informationen zu rechtlichen Aspekten bei der Handhabung von Sprachkorpora. DFG-Leitlinien zum Umgang mit Forschungsdaten. http://www.dfg.de/foerderung/

antragstellung_begutachtung_entscheidung/antragstellende/antragstellung/nachnutzung_forschungsdaten/.

Digital Preservation Coalition (2015). *Digital Preservation Handbook*. Digital Preservation Coalition, 2 edition.

Fecher, B. and Friesike, S. (2014). Open science: one term, five schools of thought. In *Opening science*, pages 17–47. Springer.

ISO (2016). ISO 24624:2016 Language resource management – Transcription of spoken language. Technical report, ISO, Genève.

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

Kupietz, M., Belica, C., Keibel, H., and Witt, A. (2010). The German Reference Corpus DeReKo: A Primordial Sample for Linguistic Research. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 1848–1854, Valletta/Paris. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2010/pdf/414_Paper.pdf.

Kupietz, M., Lüngen, H., Bański, P., and Belica, C. (2014). Maximizing the potential of very large corpora: 50 years of big language data at IDS Mannheim. In Kupietz, M., Biber, H., Lüngen, H., Bański, P., Breiteneder, E., Mörth, K., Witt, A., and Takhsha, J., editors, *Proceedings of the LREC 2014 Workshop "Challenges in the Management of Large Corpora" (CMLC-2)*, pages 1–6, Reykjavik/Paris. European Language Resources Association (ELRA). http://nbn-resolving.de/urn:nbn:de:bsz:mh39-31634.

Kupietz, M., Lüngen, H., Kamocki, P., and Witt, A. (2018). The German Reference Corpus DeReKo: New Developments – New Opportunities. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC'18)*, pages 4353–4360, Miyazaki/Paris. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2018/pdf/737.pdf.

Lehr, U. and Thomae, H., editors (1987). *Formen seelischen Alterns*. Enke, Stuttgart.

Lüngen, H. and Sperberg-McQueen, C. M. (2012). A TEI P5 Document Grammar for the IDS Text Model. *Journal of the Text Encoding Initiative*, 3:1–18. http://jtei.revues.org/508.

McBride, B. (2003). The resource description framework (RDF) and its vocabulary description language RDFS. In Stab, S. and Studer, R., editors, *Handbook of Ontologies*, chapter 3, pages 29–50. Springer, Berlin, Heidelberg, New York.

Neuroth, H., Oßwald, A., Scheffel, R., Strathmann, S., and Jehn, M., editors (2009). *nestor Handbuch : eine kleine Enzyklopädie der digitalen Langzeitarchivierung*. nestor, version 2.0 [3/2010] edition.

Schmidt, T. (2016). Construction and dissemination of a corpus of spoken interaction – tools and workflows in the FOLK project. *Journal for Language Technology and Computational Linguistics (JLCL)*, 31(1):127–154.

Schmidt, T. (2017). DGD – die Datenbank für Gesprochenes Deutsch. Mündliche Korpora am Institut für Deutsche Sprache (IDS) in Mannheim. *Zeitschrift für germanistische Linguistik*, 45(3):451–463.

Schmidt, T. and Wörner, K. (2014). Exmaralda. In Durand, J., Gut, U., and Kristoffersen, G., editors, *The Oxford handbook of corpus phonology*. Oxford University Press, Oxford.

Sperberg-McQueen, C. M. and Burnard, L., editors (1999). *Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Initiative, Chicago, New York. initial release 1994-05-16; last version dated May 1999.

Teubert, W. and Belica, C. (2014). Von der Linguistischen Datenverarbeitung am IDS zur Mannheimer Schule der Korpuslinguistik. In Steinle, M. and Berens, F. J., editors, *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache*, pages 320–328. Institut für Deutsche Sprache, Mannheim.

Wildgans, J., Weitzmann, J., and Ketzan, E. (2017). Guidelines for building language corpora under German Law – by the DFG Review Board on Linguistics. https://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/guidelines_review_board_linguistics_corpora.pdf.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific data*, 3.

7. Language Resource References

Leibniz-Institut für Deutsche Sprache (2020). German Reference Corpus DeReKo. Deutsches Referenzkorpus, DeReKo-2020-I. PID: http://hdl.handle.net/10932/00-04B6-B898-AD1A-8101-4.

Schmidt, T. and Gasch, J. (2019). Datenbank für Gesprochenes Deutsch (DGD). Leibniz-Institut für Deutsche Sprache, 2.13. https://dgd.ids-mannheim.de.