Originally published in: Navarretta, Costanza/Eskevich, Maria (Eds.): Selected Papers from the CLARIN Annual Conference 2020. Virtual Event, 2020, 5-7 October. – Linköping: Linköping University Electronic Press, 2021. Pp. 93-103. (Linköping Electronic Conference Proceedings 180)
DOI: *https://doi.org/10.3384/ecp18011*

Towards Comprehensive Definitions of Data Quality for Audiovisual Annotated Language Resources

Hanna Hedeland
Leibniz-Institut für Deutsche Sprache
Mannheim, Germany
hedeland@ids-mannheim.de

Abstract

Though digital infrastructures such as CLARIN have been successfully established and now provide large collections of digital resources, the lack of widely accepted standards for data quality and documentation still makes re-use of research data a difficult endeavour, especially for more complex resource types. The article gives a detailed overview over relevant characteristics of audiovisual annotated language resources and reviews possible approaches to data quality in terms of their suitability for the current context. Conclusively, various strategies are suggested in order to arrive at comprehensive and adequate definitions of data quality for this specific resource type and possibly for digital language resources in general.

1 Introduction

The successful development of international digital research infrastructures such as CLARIN has enabled the sharing and re-use of language resources across geographic and, partly, disciplinary boundaries. This has led to a shift in focus from the technical means of data sharing towards the data itself and in particular its quality and fitness for re-use. However, while e.g in Germany, the German Council for Scientific Information Infrastructures (RfII) states in the latest of its published recommendations that ”securing and improving data quality is a fundamental value of good scientific practice” (RfII, 2020), widely acknowledged and adequate definitions of data quality for the various types of language resources provided through digital infrastructures are yet to be defined. Generic approaches such as the FAIR Principles (Wilkinson and others, 2016) or even the FAIR Metrics (Wilkinson et al., 2018) and similar approaches based on a comprehensive assessment and evaluation of the FAIR Principles do not provide detailed guidance for research data management for specific resource types or research methods related to specific disciplines. For example, regarding Reusability, the FAIR Metrics, as the FAIR Principles, only refer to resources that ”meet community standards”, the FAIRsFAIR Data Object Assessment Metrics (Devaraju et al., 2020) to standards and formats ”recommended by the target research community” and the Data Maturity Indicators of the FAIR Data Maturity Model WG (RDA FAIR Data Maturity Model Working Group, 2020) to data that ”complies with a (machine-understandable) community standard”. It is not possible to formulate more specific criteria for the data within these generic metrics, this task is delegated to the communities and those providing research data management services for them. For archives or research data centres aiming to make the deposited and hosted data FAIR, the existing frameworks could be used to design corresponding metrics or indicators for specific communities or resource types. Focusing on the resource type rather than the communities might be advisable, since as Bahim et al. (2021, p. 5) report after surveying ”communities” on topics related to FAIR data assessment, ”[d]espite being widely mentioned in the RDA context, the term ”community” remains unclear. The respondents are still questioning who these communities are and who are the stakeholders constituting them.”

Research data quality calls for adequate and comprehensive definitions, but this raises several – often overlooked – fundamental questions. Suitable quality criteria need to be transparent and operationalized, but also reflect the complexity of the subject matter, in our case: audiovisual annotated language data. A

Hanna Hedeland 2021. Towards Comprehensive Definitions of Data Quality for Audiovisual Annotated Language Resources. *Selected papers from the CLARIN Annual Conference 2020.* Linköping Electronic Conference Proceedings 180: 180 93–103.

first step is therefore a review of this resource type, before various approaches to defining data quality criteria can be evaluated in terms of their applicability.

2 Taking Stock of Audiovisual Annotated Language Data Resources

The various resource types subsumed under ”audiovisual annotated language resources” are highly heterogeneous but have in common that they comprise several data types and display a complex structure of abstract entities and data objects with different types of relations. A comprehensive description of these resources and the variation within the group is therefore an important first step. In figure 1, relevant data types and processing stages are pictured in detail in an approach similar to that proposed by Himmelmann (2012). For a specific data type, each stage or higher level can be based on all available data on the level(s) below, the most relevant input is however placed directly beneath it. Data types in filled dashed boxes are usually superseded by their counterparts on the next level and only retained for archival or reproducibility purposes. Data types in transparent boxes represent further processing and analysis stages based (primarily) on non-audiovisual sources, i.e. images or textual data. These data types are usually not created for audiovisual annotated language corpora, which rather only include annotations of audiovisual sources, i.e. recordings. In other types of research data within the humanities and social sciences, the data model expects coding to be performed on several sources of various modalities. While data models for these different types of annotations all include references to the prepared sources, the previous processing steps are only rarely recorded as (provenance) metadata. On each level, the notion of data quality has different implications, which takes the relevant research activities into account.

A better understanding and description of this resource type is one goal of the QUEST[1] project, which was based on the the existing cooperation of the CLARIN Knowledge Centre for Linguistic Diversity and Language Documentation[2] – including DCH/IfL (Cologne), ELAR/SWLI (London), HZSK/INEL (Hamburg) and ZAS (Berlin) – and extended by the German Sign Language Corpus project (Hamburg) and the Archive for Spoken German at IDS (Mannheim), adding their complementary expertise on sign language data and German data, respectively. ELAR and the Cologne Language Archive (LAC) allow self-deposit of resources with basic requirements on file formats and metadata, whereas the AGD and (previously) the HZSK curate resources to comply with data models and data consistency requirements. The data deposited with the AGD and the HZSK is mainly from projects working with qualitative methods only, for which the requirements regarding data modelling and consistency play a subordinate role. This is also reflected in the data to a varying extent. The resources in all four centres differ along several dimensions, which can be described as structural, methodological and content-based heterogeneity.

2.1 Structural Heterogeneity

Abstract data models for language resources such as EXMARaLDA (Schmidt and Wörner, 2014) or the DGD data model (Schmidt et al., 2013a) provide explicit requirements not only on the overall resource structure of abstract entities and data objects, but also regarding the resource content, i.e. consistency in tier structure and the conventions and schemes used for transcription and annotation and the identity of speakers. Even without an explicit data model, the resource structure is also defined by contextual data, including structurally relevant entities such as recording sessions, and by metadata on included files and their relations. However, not all resources found in archives today exploit such elaborate models or metadata formats. Some are limited to a set of audio and video recordings and individual transcripts, in some cases with no explicit information on the internal structure and only minimal metadata.

When it comes to the explicit or implicit data models describing collections of audiovisual linguistic research data, there are several differences as outlined in Figure 2, which shows various data models and metadata standards. In the figure, filled boxes represent the modelling of real world entities, e.g. participants, as entities, i.e. an independent object being referred to via a unique ID. Transparent boxes on the other hand represent complex or simple data types, which also have IDs, but these are not used for reference to avoid redundant listing of e.g. a participant with all its information for each session. The

[1] https://www.slm.uni-hamburg.de/en/ifuu/forschung/forschungsprojekte/quest.html

[2] http://ckld.uni-koeln.de/

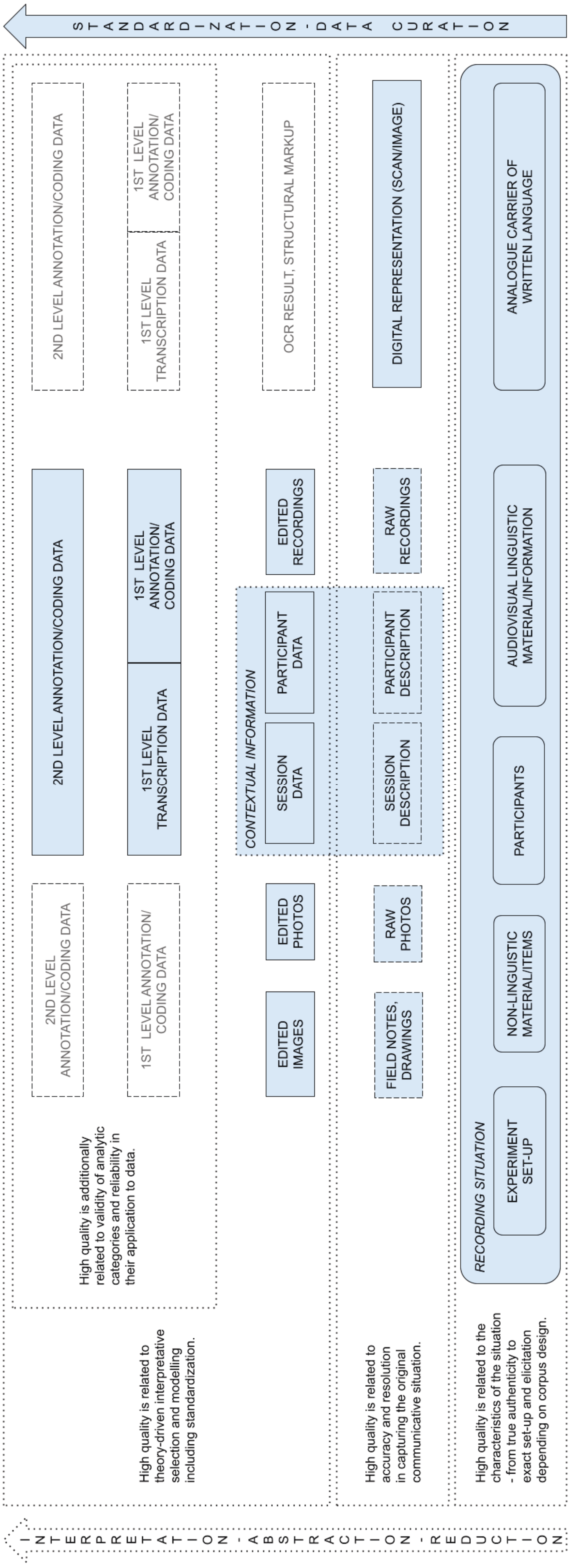

STANDARDIZATION - DATA CURATION
INTERPRETATION - ABSTRACTION - REDUCTION
High quality is related to theory-driven interpretative selection and modelling including standardization.
High quality is additionally related to validity of analytic categories and reliability in their application to data.
2ND LEVEL ANNOTATION/CODING DATA
1ST LEVEL ANNOTATION/CODING DATA
2ND LEVEL ANNOTATION/CODING DATA
1ST LEVEL TRANSCRIPTION DATA
1ST LEVEL ANNOTATION/ CODING DATA
2ND LEVEL ANNOTATION/CODING DATA
1ST LEVEL TRANSCRIPTION DATA
1ST LEVEL ANNOTATION/ CODING DATA
EDITED IMAGES
EDITED PHOTOS
CONTEXTUAL INFORMATION
SESSION DATA
PARTICIPANT DATA
EDITED RECORDINGS
OCR RESULT, STRUCTURAL MARKUP
High quality is related to accuracy and resolution in capturing the original communicative situation.
FIELD NOTES, DRAWINGS
RAW PHOTOS
SESSION DESCRIPTION
PARTICIPANT DESCRIPTION
RAW RECORDINGS
DIGITAL REPRESENTATION (SCAN/IMAGE)
High quality is related to the characteristics of the situation - from true authenticity to exact set-up and elicitation depending on corpus design.
RECORDING SITUATION
EXPERIMENT SET-UP
NON-LINGUISTIC MATERIAL/ITEMS
PARTICIPANTS
AUDIOVISUAL LINGUISTIC MATERIAL/INFORMATION
ANALOGUE CARRIER OF WRITTEN LANGUAGE

Figure 1: Data of various types are derived from existing data of a research project through various research activities during data creation and evaluation.

way participants are modelled is an example of which kind of information can be specified and how consistency can be controlled. The EXMARaLDA Corpus Manager models communications and Speakers as independent entities, and can therefore only model globally valid participant information. The IMDI and TEI formats, on the other hand, only model session specific participant information as participants do not exist independently of sessions. This allows for inclusion of relevant session specific information but makes data consistency more challenging. While participants can be listed in the `<teiHeader>` of the `<teiCorpus>` for TEI, there is no designated mechanism of referring to these instead of declaring them on the document level. The data model of the DGD also views Speakers as entities and can relate Speakers to Speech events, but the model is also capable of accommodating session specific information (such as the participant's role in the interaction or age at the date of recording) through the ”Speaker in Speech Event” relation. Another problematic aspect for interoperability across resources is the level(s) below the top resource level. If we disregard corpus design specific sublevels, e.g. ”L2 Speakers” vs. ”L1 Speakers” we still find differences and in many cases ambiguities regarding the meaning of structural units such as sessions. Basically, there are three relevant dimensions in this case: The time-based *recording session*, specific linguistically determined *communicative events*, possibly contained in a common recording session, and the *partition based on storage media*. While structuring a collection according to storage media (e.g. unit ”Tape 005, side A”) will most likely never be intended as a design choice, it is often encountered in legacy data, where analogue carriers have been digitized but not yet further processed and described. Due to the existing heterogeneity, it is not possible to decide on a certain recommended alternative, but the information should be included in the metadata. This also applies for cases where for technical reasons (e.g. handling of large files) communicative events have been arbitrarily split into separate units.

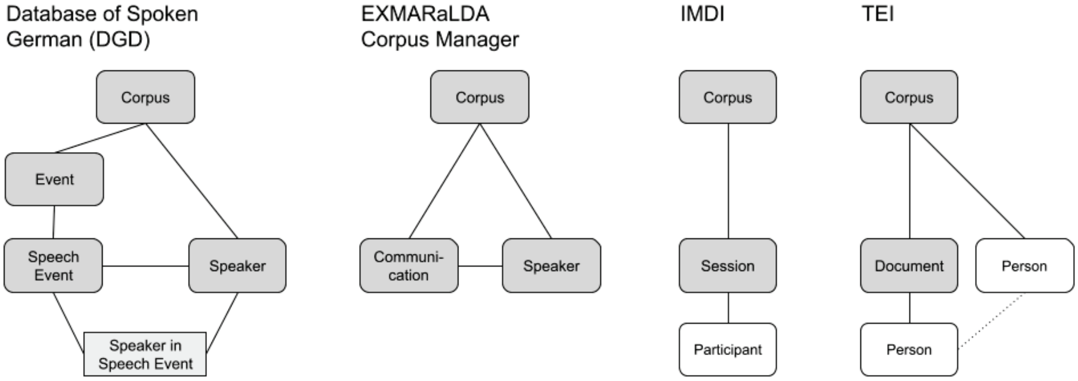


Figure 2: Various data models and metadata formats make explicit and implicit statements regarding resource structure.

Below the session level the heterogeneity is even greater, as the file formats used to encode transcription and annotation data vary quite a lot. In particular, they are either unstructured, i.e. provided in some (plain) text format, or (semi-)structured, as most XML transcription/annotation tool formats. Such structured formats show differences both in the macro-structure of tiers and speaker contributions and annotations and the micro structure related to annotation schemes and transcriptions conventions as explained in detail in Schmidt (2011). A comprehensive discussion is beyond the scope of this paper.

2.2 Methodological Heterogeneity

Differences on the transcript micro-level (Schmidt, 2011) depend on the research methods employed, especially in terms of qualitative or quantitative approaches. Annotations thus range from comprehensively applied systematic tags to selectively applied free comments. Since transcription conventions select and foreground certain aspects of language (Ochs, 1979), they also differ regarding units such as utterances or intonation phrases and the amount of linguistic information integrated into the basic transcription. Furthermore, not all transcription and annotation schemes lend themselves to automatic checking.

2.3 Content-related Heterogeneity

The content-related resource design plays a major role when it comes to visible differences due to choices regarding geographical and temporal coverage, and the selection of participants, topics, (multi)linguality types etc. for the data collection. Furthermore the amount and categories of contextual data describing recording sessions and participants also differ accordingly. The importance of complementary data types beyond recordings, annotations and contextual data, such as written or image material present in the recording situation also depend on the research question and resource design, i.e. the content.

3 Approaches to Data Quality and Possible Applicability for Language Resources

Since audiovisual annotated language resources are research data, which is a specific type of data in general, more generic approaches to data quality can provide valuable insights. These are therefore reviewed while evaluating the need to complement them with further more specific criteria.

3.1 Generic Approaches to Data Quality

Generic approaches do not restrict the types of data they are applicable to and thus recommendations remain general and abstract. (Wang and Strong, 1996) distinguish fundamental dimensions: intrinsic, contextual, representational and accessibility data quality, pertaining to the data itself, a particular usage context, and the systems providing data, respectively. This distinction between inherent and system-dependent data quality is also reflected in ISO/IEC 25012 - The Data Quality Model[3]. The W3C provide relevant input in their Best Practices for Data on the Web[4], both regarding the recommendations and the system used to disseminate them. However, these generic approaches do not provide directly applicable resource specific recommendations.

3.2 Approaches to Research Data Quality

Today, the main requirement for research data is to be FAIR. In terms of the generic data quality dimensions, not all of the FAIR principles and corresponding metrics or criteria are directly related to intrinsic characteristics of the data, but also to the infrastructure required to make data findable and accessible, and thus not under the control of the data creators. As outlined above, approaches aiming to operationalize the well-known principles also only refer to community-specific standards. The FAIRification process (Jacobsen et al., 2020) also still needs resource type specific requirements and workflows, but is a starting point to redefine data curation processes in line with FAIR concepts.

The concept of Data Maturity on the other hand, seems to be a suitable way of avoiding the word ”quality” altogether for dimensions related to data strucuturedness and machine-understandability, which might not be considered to be closely related to data quality at all by some humanities scholars. An important aspect beyond the scope of this paper, but which must be considered at all times, is the quality of research data as an artefact of research. The artefact can only be as good as the research (and vice versa). While this quality aspect can only be assessed with thorough documentation of the data and its provenance, relevant theoretic frameworks, research methods and analytic categories used, for metadata and data to be machine-readable or even machine-understandable is not a relevant requirement.

3.3 Resource Type Specific Approaches to Data Quality

Within CLARIN, there is work in progress to collect recommendations from all CLARIN B centres on standards and formats accepted for deposit[5]. Apart from the participants of the QUEST project, some centres providing detailed recommendations for audiovisual data are e.g. The Language Archive at the MPI in Nijmegen[6] and the Bavarian Archive for Speech Signals[7]. Furthermore, the German funder DFG

[3] https://www.iso.org/standard/35736.html

[4] https://www.w3.org/TR/dwbp/

[5] https://www.clarin.eu/content/standards-and-formats

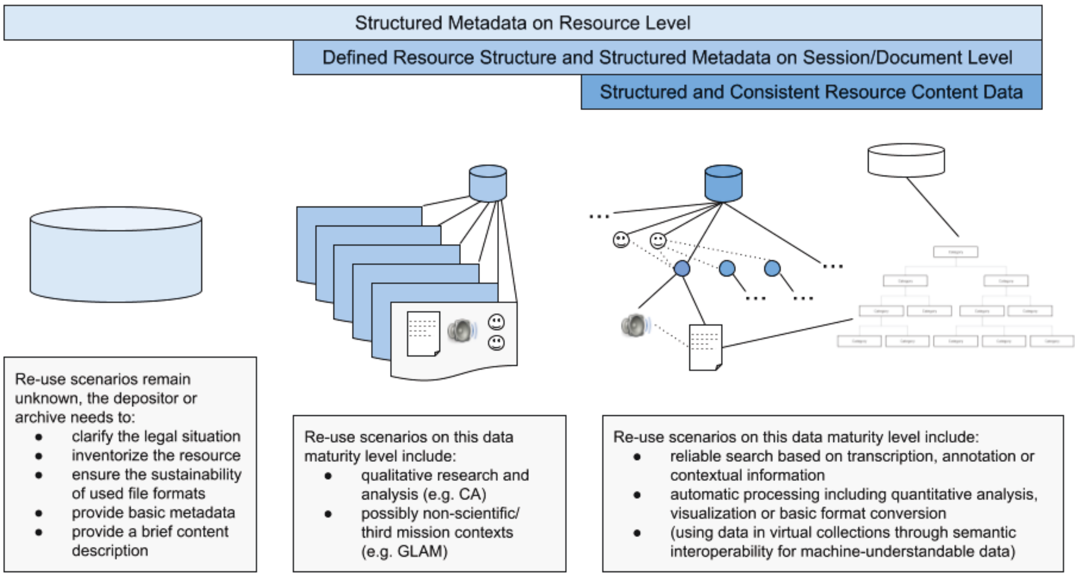
[6] https://archive.mpi.nl/tla/accepted-file-formats

[7] https://www.phonetik.uni-muenchen.de/Bas/BasInfoStandardsTemplateseng.html

has published recommendations for technical standards[8] collected through discussions within the relevant research communities. And still highly relevant after almost twenty years, (Bird and Simons, 2003) have described several aspects relevant for the long-time preservation and re-use of language documentation data. These recommendations are valuable resources for comprehensive definitions of resource type specific data quality.

4 Step One: Defining Data Maturity Levels for Audiovisual (Annotated) Language Resources

Considering the heterogeneity, it would be inappropriate to measure quality without regarding the conscious choices and trade-offs made by researchers that affect the machine-readability and consistency of the data, i.e. aspects of data maturity that might not be of direct value considering the chosen research method. The AGD and the HZSK have defined guidelines for deciding whether to perform data curation (Schmidt et al., 2013b). While curation of deposited data increases the re-use potential, the increasing deposit numbers without corresponding growth of capacities must necessarily run into a dead end. By allowing controlled variation, re-users know what to expect from language resources provided by others and more adequate goals for evaluation and curation can be defined. While the focus here is on audiovisual data, the following categories could also be applied to written language resources. The exact set of criteria for each level is still work in progress, as is the common system of quality and curation criteria for the QUEST project, but for now, three examples of criteria that have to be met are included for each level below. Some of these can be tested automatically, while others require manual (human) evaluation. The category names refer to prototypical resource subtypes and are not an integral part of the framework.


Structured Metadata on Resource Level
Defined Resource Structure and Structured Metadata on Session/Document Level
Structured and Consistent Resource Content Data
...
...
...
...

Re-use scenarios remain unknown, the depositor or archive needs to:

- clarify the legal situation
- inventorize the resource
- ensure the sustainability of used file formats
- provide basic metadata
- provide a brief content description

Re-use scenarios on this data maturity level include:

- qualitative research and analysis (e.g. CA)
- possibly non-scientific/ third mission contexts (e.g. GLAM)

Re-use scenarios on this data maturity level include:

- reliable search based on transcription, annotation or contextual information
- automatic processing including quantitative analysis, visualization or basic format conversion
- (using data in virtual collections through semantic interoperability for machine-understandable data)

Figure 3: The three levels of data maturity developed for audiovisual annotated language resources allow for adequate assessment of data quality and realistic data curation plans.

4.1 Deposits

Since research data centres will always be confronted with orphaned legacy data, there have to be minimal requirements for data which is by no means FAIR, but still, especially in the case of endangered languages or oral history data, is undoubtedly worth being archived. A deposit is thus a data set with minimal metadata clarifying the legal situation and providing basic information on the content.

[8] `https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_sprachkorpora.pdf`

- The licensing information provided is sufficient to decide which re-use is possible.
- There is a complete inventory describing all files included in the resource.
- All files come in well-documented formats that can be read by non-proprietary software.

4.2 Collections

On the next level, Collections comply with additional requirements on the resource structure level, including the completeness and consistency of metadata and relations between all resource parts. The requirement on completeness of metadata only pertain to standardized cataloguing metadata describing the linguistic resource and its provenance to make the data discoverable and comprehensible. Contextual data, e.g. information on participants, on the other hand can not be standardized, since this would interfere with research design. The textual language data of Collections is usually provided in various unstructured text formats suitable for human manual analysis, but there are no requirements on the completeness or even existence of transcripts.

- There is basic metadata for individual recording sessions including participants.
- The deposited materials share thematic characteristics.
- All files come in well-documented formats that can be read by non-proprietary software.

4.3 Corpora

Corpora fulfill all requirements of Collections and are additionally structured and consistent on the resource content level, i.e. in the use of tier structure, annotation schemes and transcription conventions, but also regarding contextual data such as participant identities across the resource. While the Corpus data is machine-readable and suitable for reliable automatic analysis, definitions of for instance tier content or annotation schemes are often not machine-understandable and interoperability is thus generally limited to syntactic aspects.

- Tier (annotation level) definitions are consistent across the resource.
- (Relevant) participants can be identified across the resource.
- There is a clear design principle for the resource and its parts.

5 Step Two: Data Curation as FAIRfication

Since important aspects of research data quality are reflected by the FAIR principles and various metrics based on them, data curation can also be considered FAIRification, the process resulting in FAIR data. In this process, beyond syntactical correctness, the semantic information needs to be made explicit; we need to ”define the semantic model”. The differences in the level of data maturity described above are relevant for this process, since for Deposits and Collections, which do not have structured transcription/annotation data, machine-readable definitions enabling the data to become machine-understandable and further semantic enrichment and linked open data features are only possible for metadata. For Corpora on the other hand, the structured data allows for the semantic model to be defined more fine-grained, by linking tiers and annotations to controlled vocabularies and ontologies, but this option has rarely been used, e.g. the option to reference ISO Data Categories available in ELAN (Sloetjes, 2014) seems to play no role in the ELAN annotation data (EAF) currently found in archives (von Prince and Nordhoff, 2020).

As described in Schmidt (2011), the ISO standard for Transcription of Spoken Language[9] provides more semantic information on units and information types as part of the underlying data model than most widely used formats for transcription/annotation data, which do not define the notation of e.g. participants' contributions, noise or pauses. As the standard was developed with this idea in mind, conversion into this format (Schmidt et al., 2017) would be one step towards semantic interoperability, while still not enforcing any semantics for the theory-dependent micro-structure. This is an important aspect,

[9] https://www.iso.org/standard/37338.html

since standardization efforts based on defining the micro-structure, such as the CHAT format, are bound to be restricted to the specific theoretical frameworks and usage scenarios they reflect.

Although the benefits of this approach and the required next step of providing reliable parsers for various micro-structures were outlined already in Schmidt (2011), almost ten years later, despite undeniable progress in digital research data management for audiovisual language resources, some basic modules and functionality are still lacking beyond conversion of widely used tool formats into the ISO/TEI format for this type of data to become truly FAIR. For this, agreement is necessary on how to describe both the information types in various annotation levels and tiers, and the rules for parsing the micro-structure in a machine-understandable way. Though alternatives such as OLiA[10] exist, there is still no designated widely used method to include machine-readable references for tiers or individual annotations in TEI-based formats. Additional conventions would allow for a proper definition of the semantics of individual data sets and increase the options for re-use, especially for automatic processing and enrichment. Another aspect purposefully not treated in depth in the ISO/TEI approach is semantic interoperability on the level of the resource, or even means of encoding basic contextual data in a standardized manner. The generic TEI standard was not primarily developed for spoken corpora and the TEI Corpus is a set of documents. It would be possible to simply include relevant parts of existing formats[11], but these are also not interoperable, even though, as shown in 2, they share basic characteristics in the same way as transcription data does, simply because these are models of a common reality.

6 Step Three: Adding the ”Fit for Purpose” Dimension

While the aspects of FAIRification (from data structuredness to semantic enrichment and linking) are generic, data quality is to a great extent a question of the data being fit for particular purposes or usage scenarios – and not all usage scenarios improve directly by using more structured data, i.e. a higher level of data maturity. Since it is not feasible for research projects creating language resources to consider all possible re-use scenarios, explicit and formalized definitions of re-use scenarios would allow projects to comply with specific re-use scenarios. Re-users would also be able to quickly judge whether the data is suitable for their purpose, which is often difficult to tell today, especially in the case of interdisciplinary re-use, e.g. between linguistics and education sciences, partly also due to the use of different terminology.

The definition and implementation of criteria for such interdisciplinary re-use scenarios are further important goals of the QUEST project, complementing the technical and intrinsic aspects of data quality. Within the QUEST project, four main re-use scenarios are being investigated and systematically described on various levels ranging from the general legal situation to the interoperability with specific data formats and the use of certain annotation schemes or transcription conventions. For example, to enable re-use of research data from linguistic research projects within third mission contexts, e.g. as audiovisual augmentation in museums, the legal situation must allow (parts of) the data to be made available to the public, and specific linguistic information will have to be removed from transcripts to make them readable to laymen.

When considering the data maturity levels of audiovisual resources described above, it also becomes clear how they enable various forms of re-use. While audio files might be available in any case, reliable metadata on individual recording level, as required for Collections, is necessary to make a selection. With structural speaker assignment and alignment of the transcripts with the audio, more options to tailor the material become available and only structured data, as required for Corpora, can be automatically enriched in a reliable manner, converted, aggregated or visualized to suit the needs of the re-using institution.

Within the QUEST project, two re-use scenarios are used as pilot cases for semantic interoperability using the ISO/TEI standard. The first is the kind of qualitative data created within many areas of non-linguistic humanities research by so-called CAQDAS, computer assisted qualitative data analysis software. Though such resources bear resemblance to the resource pictured in Figure 1, and often include coding on images and textual content in addition to audiovisual sources, the transcript data of such

[10] `http://www.acoli.informatik.uni-frankfurt.de/resources/olia/`

[11] E.g. by using the xenoData element of the teiHeader.

resources is rarely structured. This also applies for the new open standard in development, REFI-QDA[12], which will serve as an exchange format between widely used proprietary tools such as ATLAS.ti[13] or NVivo[14]. Since this format is only targeted at plain text transcripts, a proof-of-concept conversion scenario for this type of data to the ISO/TEI standard was implemented based on a TEI format developed at the UK Data Service QualiBank[15]. TEI as a structured transcription format is allowed in their CAQDAS exchange format, QuDEx[16] developed several years ago.

The second pilot case is the morphosyntactic description of interlinear glossed text in language documentation data. While there are several conventions in use, many are very similar and only have slight variation on the level of symbols used to denote various concepts. Within the INEL project and the HZSK, an extension for this type of data was developed (Arkhangelskiy et al., 2019), however, the TEI class att.linguistic[17] (Bański et al., 2018) could possibly provide a simpler solution. This TEI class is used for token-based annotation in the DGD, the MTAS-based platform developed within the ZuMult project[18] and also for written language, e.g. in the DTA Basis format[19].

Both pilot cases share their relevance beyond communities merely interested in the linguistic features of the data, since they can often both be considered useful for oral history related research. The combination of TEI and RDF or linked data is on the one hand common but has on the other hand been solved in various ways, with a uniform solution based on RDFa currently being developed[20].

7 Discussion

Though seemingly trivial, fundamental questions regarding the structure and content of annotated audiovisual language resources created as research data within various disciplines have yet to be thoroughly discussed and answered. The characteristics of such resources need to be systematically described in order to define suitable criteria for data quality. When defining such criteria, we need to acknowledge that data maturity might seem irrelevant to many humanities researchers and in some cases really is[21]. The characteristics of the research data is the result of the research methods, hence a research data centre should be able to handle different data maturity levels while providing comprehensive examples of how increasing the level of data maturity has benefits for the individual researcher. As discussed, standardization can only apply to certain aspects of the data, the metadata and the documentation. Machine-understandable expression and documentation of the micro-structure would allow for conversion into the ISO/TEI format, which would increase re-use options and allow for the use of TEI compatible services and tools where this makes sense from the researcher's perspective. However, this it not a task for humanities researchers but requires the expertise of research data managers or data stewards in close cooperation with the research projects.

Providing generic quality criteria applicable to resources with various data maturity levels is one aim of the QUEST project, another is to provide additional criteria for formalized re-use scenarios. To allow data creators to comply with these criteria, the project will also provide software solutions to evaluate various types of resources according to such generic and specific criteria (Arkhangelskiy et al., 2020). This evaluation will ideally be performed continuously during data creation as a part of the project's data quality assurance methods. In combination, the definitions and evaluation mechanisms developed within

[12] https://www.qdasoftware.org/about/

[13] https://atlasti.com/

[14] https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home

[15] https://ukdataservice.ac.uk/get-data/explore-online/qualibank/qualibank.aspx

[16] https://www.data-archive.ac.uk/managing-data/standards-and-procedures/metadata-standards/qudex/

[17] https://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-att.linguistic.html

[18] http://zumult.ids-mannheim.de

[19] https://www.deutschestextarchiv.de/doku/basisformat/

[20] https://github.com/TEIC/TEI/issues/1860

[21] Cf. RDA FAIR Data Maturity Model Working Group (2020, p. 10)"[D]data coming from humanities fields, especially from outside of Digital Humanities, will often not be expressed in a machine understandable knowledge representation (RDF, SKOS or LOD) by nature but instead, it is often expressed in natural language, even if encoded using machine readable methods (e.g. TEI). Therefore, it becomes quite clear that the indicator treating machine-understandable knowledge representation will be less relevant according to the humanities."

the QUEST project will hopefully make data depositing and re-use more transparent and fruitful within and across disciplines.

References

Timofey Arkhangelskiy, Anne Ferger, and Hanna Hedeland. 2019. Uralic multimedia corpora: ISO/TEI corpus data in the project INEL. In *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages*, pages 115–124, Tartu, Estonia, January. Association for Computational Linguistics.

Timofey Arkhangelskiy, Hanna Hedeland, and Aleksandr Riaposov. 2020. Evaluating and assuring research data quality for audiovisual annotated language data. In *Proceedings of the CLARIN Annual Conference 2020*. CLARIN ERIC.

Christophe Bahim, Makx Dekkers, Edit Herczog, Keith Russell, and Shelley Stall. 2021. Survey on bridging the gap between funders and communities — perspectives on benefits and challenges of fair assessments. v1.0.

Piotr Bański, Susanne Haaf, and Martin Mueller. 2018. Lightweight grammatical annotation in the TEI: New perspectives. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018), 7-12 May 2018, Miyazaki, Japan*, pages 1795 – 1802, Paris, France. European language resources association (ELRA).

Steven Bird and Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language*, 79(3):557–582, September.

Anusuriya Devaraju, Robert Huber, Mustapha Mokrane, Patricia Herterich, Linas Cepinskas, Jerry de Vries, Herve L'Hours, Joy Davidson, and Angus White. 2020. FAIRsFAIR Data Object Assessment Metrics. October.

Nikolaus P. Himmelmann. 2012. Linguistic data types and the interface between language documentation and description. *Language Documentation and Conservation*, 6:187–207.

Annika Jacobsen, Rajaram Kaliyaperumal, Luiz Olavo Bonino da Silva Santos, Barend Mons, Erik Schultes, Marco Roos, and Mark Thompson. 2020. A generic workflow for the data FAIRification process. *Data Intelligence*, 2:56–65.

Elinor Ochs. 1979. Transcription as theory. In E. Ochs and B.B. Schieffelin, editors, *Developmental pragmatics*, pages 43–72. Academic Press, New York.

RDA FAIR Data Maturity Model Working Group. 2020. Fair data maturity model: specification and guidelines. June.

RfII. 2020. The Data Quality Challenge. Recommendations for Sustainable Research in the Digital Turn.

Thomas Schmidt and Kai Wörner. 2014. EXMARaLDA. In Ulrike Gut Jacques Durand and Gjert Kristoffersen, editors, *Handbook on Corpus Phonology*, pages 402–419. Oxford University Press.

Thomas Schmidt, Sylvia Dickgießer, and Joachim Gasch. 2013a. Die datenbank für gesprochenes deutsch - DGD2.

Thomas Schmidt, Kai Wörner, Hanna Hedeland, and Timm Lehmberg. 2013b. Leitfaden zur beurteilung von aufbereitungsaufwand und nachnutzbarkeit von korpora gesprochener sprache.

Thomas Schmidt, Hanna Hedeland, and Daniel Jettka. 2017. Conversion and annotation web services for spoken language data in clarin. In *Selected papers from the CLARIN Annual Conference*, pages 113–130, Aix-en-Provence, France. Linköping University Electronic Press, Linköpings Universitet.

Thomas Schmidt. 2011. A TEI-based Approach to Standardising Spoken Language Transcription. *Journal of the Text Encoding Initiative*, 1, 06.

Han Sloetjes. 2014. ELAN: Multimedia annotation application. In Jacques Durand, Ulrike Gut, and Gjert Kristoffersen, editors, *Handbook on Corpus Phonology*, pages 305–320. Oxford University Press.

Kilu von Prince and Sebastian Nordhoff. 2020. An empirical evaluation of annotation practices in corpora from language documentation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2778–2787, Marseille, France, May. European Language Resources Association.

Richard Y. Wang and Diane M. Strong. 1996. Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12(4):5–33.

Mark D. Wilkinson et al. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3:160018–, March.

Mark Wilkinson, Susanna-Assunta Sansone, Erik Schultes, Peter Doorn, Luiz Olavo Bonino da Silva Santos, and Michel Dumontier. 2018. A design framework and exemplar metrics for FAIRness. *Scientific Data*, 5:180118, 06.