Published in: Navarretta, Costanza/Eskevich, Maria (Eds.): Proceedings of CLARIN Annual Conference 2020.
05 – 07 October 2020, Online Edition. - Utrecht: CLARIN, 2020. Pp. 136-140.
https://office.clarin.eu/v/CE-2020-1738-CLARIN2020_ConferenceProceedings.pdf

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 particular resource type.

1 Introduction

The successful development of large 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 the German Council for Scientific Information Infrastructures (RfII) states in the latest of their 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 still lacking. Generic approaches such as the FAIR Principles (Wilkinson and others, 2016) or even the FAIR Metrics (Wilkinson et al., 2018) do not provide detailed guidance for research data management for specific resource types or research methods related to specific disciplines. The metrics only refer to data formats "recommended by the target research community" and since the metrics are not resource type or discipline specific, it is not possible to formulate more specific criteria for the data within these generic metrics.

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, audiovisual annotated language data. A first step is therefore a review of this resource type, before various approaches to defining data quality criteria can in turn 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. This is one goal of the German QUEST project, based on the the existing cooperation of the CLARIN Knowledge Centre for Linguistic Diversity and Language Documentation (CKLD)[1] – including DCH/IfL (Cologne),

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

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 German data and sign language data, respectively.

The participating data centres partly allow self-deposit of resources with basic requirements on file formats and metadata – ELAR and the Cologne Language Archive (LAC) –, and partly curate resources to comply with corpus data models and achieve data consistency – the AGD and the HZSK. Since the data deposited with the AGD and the HZSK is often from projects working with qualitative methods only, all requirements are not really relevant to the original research, which is obviously 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 structural requirements on not only the macro-level of abstract entities and data files, but also regarding the micro-level, with consistency in tier structure and content and regarding 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, e.g. IMDI[2] or CMDI[3] metadata. However, not all resources exploit such models or schemes, but simply amount to a set of audio and video recordings and individual transcripts, in some cases with no explicit information on the internal structure available and only minimal metadata. Though resources typically contain the same type of abstract and data objects, the corresponding file formats vary. In particular, they are either unstructured, e.g. as transcription data provided in PDF or plain text format, or structured, e.g. as XML transcription/annotation tool formats.

2.2 Methodological Heterogeneity

Differences on the micro-level are strongly dependent on the research methods employed, especially whether qualitative or quantitative/automatic analysis have been used. Annotation thus range from tags from a controlled scheme added in a systematic and comprehensive way to interpretative free text comments added to relevant parts only. Since transcription conventions capture and stress certain aspects of language, they also differ with respect to 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 in use lend themselves to automatic syntax 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 is research data, which is in turn data, more generic approaches to data quality can provide valuable insights and 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 too and thus recommendations remain general and abstract. (Wang and Strong, 1996) distinguish fundamental dimensions: intrinsic,

[2] https://archive.mpi.nl/forums/t/imdi-metadata-information/2639/2

[3] https://www.clarin.eu/cmdi

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[4]. The W3C provide relevant input in their Best Practices for Data on the Web[5], 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, for research data to be FAIR is the main requirement. Even though the FAIR metrics aim to operationalize the well-known principles, they also only refer to community-specific standards. The FAIRification process (Jacobsen et al., 2020) also requires resource type specific requirements and workflows, but is a starting point to redefine data curation processes in line with FAIR concepts.

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[6]. 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[7] and the Bavarian Archive for Speech Signals[8]. Furthermore, the German funder DFG has published recommendations for technical standards[9] 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 are valuable resources for definitions of data quality.

An important aspect which is beyond the scope of technical recommendations on standards and formats, but at the same time must be considered at all times, is the quality of research data as an artefact of research, which can only be as good as the research (and vice versa).

4 Step One: Defining Classes of Audiovisual Resources

Considering the heterogeneity, it would be inappropriate to measure quality without regarding the conscious choices and trade-offs made by researchers leading to the encountered differences. The AGD and the HZSK have defined guidelines for deciding whether to perform data curation (Schmidt et al., 2013b) and while curation of deposited data increases the re-use potential, with increasing deposit numbers, the task becomes impossible. By allowing controlled variation, re-users know what to expect from resources data 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.

4.1 Deposits

Since research data centres will always be confronted with orphaned legacy data, there needs 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, has to be archived. A deposit is thus a data set with minimal metadata clarifying the legal situation and providing basic information on the content.

4.2 Collections

On the next level, Collections comply with additional requirements on the macro-level, including the completeness and consistency of metadata and relations between all resource parts. The completeness of metadata only pertain to standardized cataloguing metadata describing the linguistic resource and its

[4] https://www.iso.org/standard/35736.html
[5] https://www.w3.org/TR/dwbp/
[6] cf. https://www.clarin.eu/content/standards-and-formats
[7] https://archive.mpi.nl/tla/accepted-file-formats
[8] https://www.phonetik.uni-muenchen.de/Bas/BasInfoStandardsTemplateseng.html
[9] https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_sprachkorpora.pdf

provenance to make the data comprehensible, since contextual data, e.g. information on participants, can not be standardized without interfering with research design. The language data of Collections is provided in various unstructured text formats suitable for human manual analysis.

4.3 Corpora

Corpora fulfill all requirements of Collections and are additionally structured and consistent on the micro-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 e.g. tier content or annotation schemes are often not machine-readable and interoperability is thus limited to syntactical interoperability.

5 Step Two: Data Curation as FAIRfication

Since important aspects of research data quality are reflected by the FAIR principles and metrics, data curation can also be considered FAIRification, a 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 structuredness described above are relevant for this process, since for Deposits and Collections, which do not have structured transcription/annotation data, machine-readable definitions including semantic enrichment and linked open data features are possible on the macro-level only. For Corpora on the other hand, the structured data on the micro-level allows for the semantic model to be defined more fine-grained, but this option has rarely been used, e.g. the option to reference ISO Data Categories available in ELAN (Sloetjes, 2014) seem to play no role in the ELAN annotation data (EAF) currently found in archives (von Prince and Nordhoff, 2020).

The ISO standard for Transcription of Spoken Language[10] 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 would be one step towards semantic interoperability, though still there is no designated method to include machine-readable references for tiers or individual annotations in this TEI-based format. Additional conventions would allow for a proper definition of the semantics of individual data sets and increase the options for re-use, especially within NLP contexts.

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

While the aspects of FAIRification (from 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 by using more structured data.

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 recognize 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 is another important goal 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.

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

When considering the classes 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 reliably automatically enriched, converted, aggregated or visualized to suit the needs of the re-using institution.

7 Outlook

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. Providing generic quality criteria applicable to resources with various levels of curation and structuredness 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 accordingly and preferably continuously during data creation. In combination, the definitions and evaluation mechanisms developed within the QUEST project will hopefully make data depositing and re-use more transparent and fruitful within and across disciplines.

References

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

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.

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.

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.