QUEST: Guidelines and Specifications for the Assessment of Audiovisual, Annotated Language Data

Anna Wamprechtshammer, Jocelyn Aznar, Elena Arestau, Hanna Hedeland, Amy Isard, Ilya Khait, Herbert Lange, Nicole Majka, Felix Rau

Anna Wamprechtshammer – Elena Arestau – Jocelyn Aznar – Hanna Hedeland – Amy Isard – Ilya Khait– Herbert Lange – Nicole Majka – Felix Rau

QUEST:

Guidelines and Specifications for the Assessment of Audiovisual, Annotated Language Data

Version 1.0: November 2022


Working Papers in Corpus Linguistics and Digital Technologies: Analyses and Methodology

Vol. 8

Anna Wamprechtshammer – Elena Arestau – Jocelyn Aznar – Hanna Hedeland – Amy Isard – Ilya Khait– Herbert Lange – Nicole Majka – Felix Rau

QUEST:

Guidelines and Specifications for the Assessment of Audiovisual, Annotated Language Data

Working Papers in Corpus Linguistics and Digital Technologies:

Analyses and Methodology

Vol. 8

Szeged – Hamburg

2022

Working Papers in Corpus Linguistics and Digital Technologies: Analyses and methodology

Vol. 8

WPCL issues do not appear according to strict schedule.

Vol. 8 (2022)

Published by
University of Szeged, Department of Finno-Ugric Studies
Egyetem utca 2. 6722 Szeged
Universität Hamburg, Zentrum für Sprachkorpora
Max-Brauer-Allee 60 22765 Hamburg

Published 2022

ISBN 978-963-306-910-3 (pdf)

Contents

Contents v

Diagrams and tables vii

1. Introduction 3

1.1 Quality Assurance for Audiovisual, Annotated Language Data 3

1.2. About QUEST 4

1.3 Use of the Document 5

1.4 Objectives 5

1.4.1 Use Case I: Evaluation of Data Deposit Projects 6

1.4.2 Use Case II: Certification & Archiving 8

1.4.3 Use Case III: Certification of Archived Data 10

2. Framework 11

2.1 Data Maturity Levels 11

2.2 Quality Criteria 12

2.3 Evaluation Method: Quality Assurance Measures and Intended Outcome 13

3. Quality Criteria 14

3.1 Overview 14

3.2 Module A: Generic Recommendations for Audiovisual Data 17

3.2.1 A0: Considerations on Legal Aspects 17

3.2.2 A1: Data Recommendations 18

3.2.3 A2: Metadata Recommendations 27

3.3 Module B: Discipline-specific Recommendations for Audiovisual Data 44

3.3.1 Module B1: Language Documentation/RefCo 45

3.3.3 Module B3: Interpreted Communication 56

3.3.4 Module B4: Anonymisation of Multimodal Corpora 62

3.3.5 Module B5: Sign Language Corpora 63

3.3.6 Module B6: Language Community 68

3.3.7 Module B7: Ethnography 71

3.3.8 Module B8: Dissemination of Oral History (meta)data via Europeana and libraries 73

4. Implementation 80

4.1 Workflows 80
4.2 Web application 80
4.2.1 Stand-alone application 80
4.3 Git workflow 81
4.4 Development, Deployment, and Maintenance 81
4.5 Automatic Checks 81
5 Related Documents 82
References 83

Diagrams and tables

Diagram 1 *Use Case I: Evaluation of Data Deposit Projects*

Diagram 2 *Use Case II: Certification & Archiving*

Diagram 3 *Use Case III: Certification of Archived Data*

Table 1 *Overview of Quality Criteria*

1. Introduction

1.1 Quality Assurance for Audiovisual, Annotated Language Data

As digitization continues to advance, it is easier than ever to collect, process and archive larger collections of digital language resources. At the same time, quality assurance measures are playing an ever more important role in the handling of larger amounts of research data and are increasingly seen as one of the major challenges for researchers and research institutions (cf. Förderung Kurationskriterien und Qualitätsstandards von Forschungsdaten, BMBF). Accordingly, the German Council for Scientific Information Infrastructures (RfII) states that „securing and improving data quality is a fundamental value of good scientific practice “(RfII, 2019).

However, even though large collections of digital resources are easily produced with the help of digital tools nowadays and increasingly made available by large infrastructures such as CLARIAH-DE[1], much of this data does not meet overall standards for data quality, which often makes their reuse a difficult endeavour.

The need for widely accepted and adequate definitions of data quality for different types of linguistic resources is especially important for data centres and archives. These institutions play a central role in the sustainable and permanent archiving and provision of spoken language corpora. The most important task of such centres is to take over corpora from completed projects and to prepare them in such a way that they can be made permanently available and archived. This task is made more difficult by the fact that in the field of audiovisual language data, it is often not possible to fall back on uniform standards. An issue, which is, among other things, due to the heterogeneity of the resource type. With respect to efforts of increasing data quality, the main requirement for research data is to be widely accessible. Because of these requirements, there is increasing reference to the FAIR[2] principles (according to the FAIR principles, data should be "Findable, Accessible, Interoperable and Re-usable") and a plea for greater consideration of these in the context of data collection and processing. However, since the FAIR principles only inadequately define how a state of 'FAIRness' can be achieved, the question of how the FAIR principles should be implemented leads to extremely different interpretations in different research disciplines.

Together with the increasing importance of the FAIR principles for the management of research data, several initiatives and projects have been developed tools for the manual or automatic assessment of data FAIRness[3]. Existing approaches based on the FAIR principles, which in addition to the development of metrics, are mostly generic and aim at the evaluation of research data in general. They do not provide detailed guidance on research data management for specific resource types regarding individual disciplines but simply reference the standards of a community, without specifying them further. For the FAIRification process[4], however, opertionalizable, resource specific requirements are also necessary.

---

[1] https://www.clariah.de/

[2] Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

[3] The FAIRassist.org website provides information on existing tools and an overview of existing FAIR assessment tools and approaches created by the RDA FAIR Data Maturity Working Group.

[4] https://www.go-fair.org/fair-principles/fairification-process/

The joint project QUEST[5] addresses this desideratum for audiovisual, annotated language data and develops both generic quality criteria and reuse-specific curation criteria as well as evaluation mechanisms tailored to the specified resource types. In addition to quality criteria, that are defined in accordance with the FAIR principles, a system of semi-automatic quality control is provided in the second step to ensure that the elaborated specifications are adhered to, and that the data remain permanently correct and consistent.

1.2. About QUEST

Funded by the German Federal Ministry of Research and Education (BMBF) as part of an interdisciplinary effort to sustainably improve data quality and general reuse potential of research data, the collaborative project QUEST[6] is developing approaches to measure and improve the quality of audiovisual, annotated language data in terms of its suitability in a specific research context.

Audiovisual, annotated language data are currently characterized by a great deal of heterogeneity regarding data constitution practices in the area of metadata, transcription and annotation. Defining uniform standards relevant for the communities already working on such language data is challenging. This is largely due to the highly discipline-specific characteristics of the resources or the discipline-specific research interests that guide the creation and indexing of resources. However, the aim of the QUEST project is not to standardise the creation of audiovisual language resources. We rather take stock of the existing heterogeneity and promote such standards and formats in use that lend themselves to preferably automatic quality control (cf. Hedeland, 2021). We pay special regard to research data generated within empirical research in the fields of language documentation, language typology, multilingualism research, oral history, and sign language. QUEST is therefore on the one hand elaborating basic generic standards in relation to data type-specific, technical standards for the various resource types and metadata relevant within the scope of the project. On the other hand, curation criteria, i.e., procedures for data enrichment and transformation, are being developed for various subject and usage specific scenarios. Developed generic quality standards and use-case-specific curation criteria are thus applied to assess the reusability of audiovisual, annotated language data.

In order to ensure continuous quality assurance of data and to support researchers in complying with quality standards and curation criteria, a review will be implemented as a data-related service. Therefore, QUEST combines different approaches to assessing data quality within a differentiated evaluation system. A data review process is provided to the different actors engaged in the quality process, which is composed of a staggered set of instruments consisting of guided online surveys, semi-automatic quality checks and discipline-specific reviewing. Internal quality assurance procedures are thus complemented by external subject-specific reviewing processes and procedures. After successfully passing through and completing the evaluation system, data producers or corpus creators are given the opportunity to have their data certified.[7]

For this purpose, the collaborative project uses expertise from an existing collaboration within the CLARIN Knowledge Centre for Linguistic Diversity and Language Documentation (CKLD)[8], involved

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

[6] Full project title: “QUEST: Quality – Established: Qualitätsstandards und Kurationskriterien für audiovisuelle, annotierte Sprachdaten”.

[7] QUEST only provides the theoretical framework for data certification practices.

[8] https://www.clarin.eu/content/ckld-clarin-knowledge-centre-linguistic-diversity-and-language-documentation

parties are the Data Centre for the Humanities (DCH)[9] and the Institute for Linguistics (IfL) in Cologne, the Endangered Language Archive (ELAR)[10] and the SOAS World Languages Institute (SWLI)[11] (London), the Hamburg Centre for Language Corpora (HZSK)[12] and the academy project INEL[13], both located in Hamburg. The network is complemented by the Hamburg project DGS-Korpus[14] (German Sign Language Corpus) and the Archive for Spoken German (AGD)[15], located at the Institute for German Language in Mannheim (IDS).

1.3 Use of the Document

This guide provides a detailed overview of the quality criteria elaborated in QUEST, which are intended to provide information on the reuse potential of audiovisual, annotated language data. It is primarily aimed at data centres and archives that wish to implement the evaluation process established within QUEST in order to make decisions about the reusability of research data within the framework of data depositing's for archiving as well as certification purposes.
In addition, based on the guide and the quality checks to be implemented, data centres and archives can already offer researchers the evaluation of their data as a service during the compilation process.

To this end, the aim of the guide is to define and record criteria for assessing the quality or reusability of audiovisual, annotated language data. The aspects of long-term accessibility as well as opening it up to broad scholarly and non-scholarly use are assessed to ensure baseline requirements for digitally driven research.

The document provides definitions and examples for each criterion and aims to give a clear overview of objects and workflows of the evaluation system, i.e., to link the quality standards and curation criteria to the data maturity levels and to make suggestions on how to evaluate each criterion.

1.4 Objectives

In this document you will find recommendations and good practices for producing audiovisual corpora of high quality, i.e., to ensure that they can be archived and reused properly. Along with an online questionnaire, automated quality checks, and a final discipline-specific review process, a complex coordinated evaluation framework is described. Using the evaluation processes and tools developed, the pursuit of three different objectives is possible: evaluating data deposit projects, archiving research data, and certifying the data for reuse. These purposes are to be presented based on three use cases: "Evaluation of Data Deposit Projects", "Archiving and Certification" and "Certification of Archived Data".

[9] https://dch.phil-fak.uni-koeln.de/

[10] https://www.elararchive.org/

[11] https://www.soas.ac.uk/courseunits/languages-world

[12] https://corpora.uni-hamburg.de/hzsk/

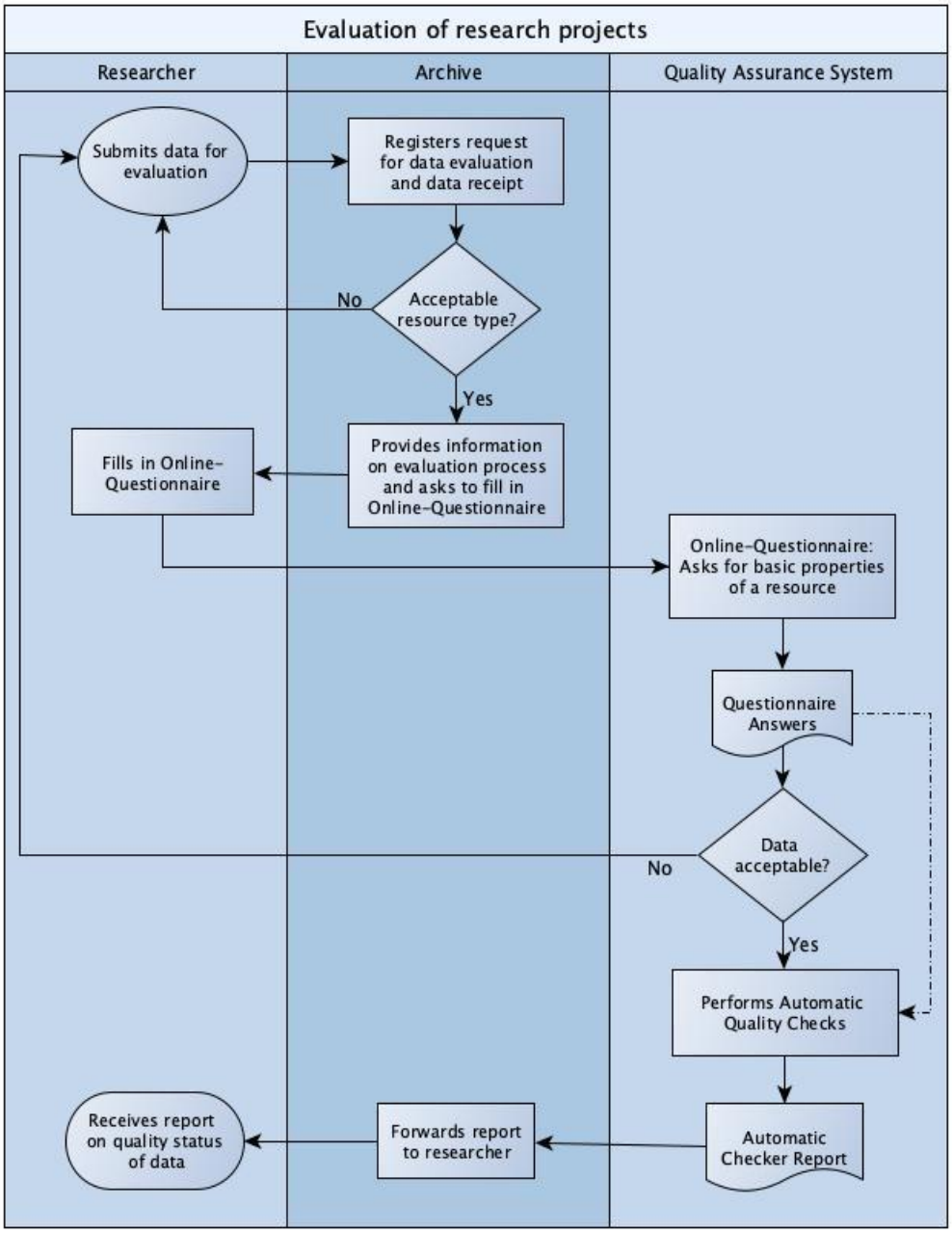
[13] https://www.slm.uni-hamburg.de/inel.html

[14] https://www.sign-lang.uni-hamburg.de/dgs-korpus/index.php/dgs-korpus.html

[15] https://agd.ids-mannheim.de/index.shtml

1.4.1 Use Case I: Evaluation of Data Deposit Projects

Diagram 1 *Use Case I: Evaluation of Data Deposit Projects*



The first of the three different use cases assumes that researchers, during or after the completion of a data deposit project, request a cooperation with a suitable data centre or an archive to have their data evaluated.

To begin, the researchers first approach an entity, which will then register the request for an evaluation. After a brief check to ensure the suitability of the resource type for the application of the quality assurance tools[16], the entity provided the information necessary to the applicant to go through the

[16]At this stage, apart from generic data and metadata standards, quality assurance in the field of audiovisual language data is only possible based on specific use cases.

evaluation process. As a next step, an online questionnaire has to be filled out. With the online questionnaire starts the configuration of the evaluation. This questionnaire contains questions that will allow the evaluation system to assess whether the submitted data are suitable or not to the entity hosting the data. By going through the questionnaire, information to configure the automatic checks are collected. The questions determine the basic characteristics of the applicant's dataset: its topic and title, its audience, the different data types (audio, video, annotations, texts) it contains, copyrights, etc. At the same time, the current data maturity level[17] of a resource is ascertained based on the answers. The data maturity level is then used as a reference for further evaluation of the resource within the quality checks. The evaluation of the data maturity level associated with a dataset quality is therefore carried out while considering the specificities of the dataset's domain (see section 2.1 below).

If the resource meets basic quality requirements, i.e., no issues are left to be fixed by the researcher, the data can be evaluated in the next step. The next step is based on a machine-readable settings file, which contains a summary of the answers from the questionnaire.

The settings file is then transferred to the automatic quality checks. Even if the setting file specifies which checks should be done, the evaluation entity still can have the final word and decide which checks will be performed. During the second step of the evaluation, data is checked by a web-service or a stand-alone application. A web page or a stand-alone application that can be downloaded and run locally, enables the entity to upload the respective dataset and, if applicable, the settings file.

The evaluation performed by an entity applying the processes devised by QUEST on the content and the metadata of a dataset ensures that it follows both generic quality standards and discipline-specific curation criteria according to the subject-specific assignment of a resource. After uploading the dataset and the settings file, the evaluation entity runs the checks, and a report is generated. This report is meant to be sent to the applicant or the entity hosting the datasets and contains the results of the quality assessment of the applicant's data. The automatic checks are based on the Corpus Services[18] framework initially developed by HZSK.

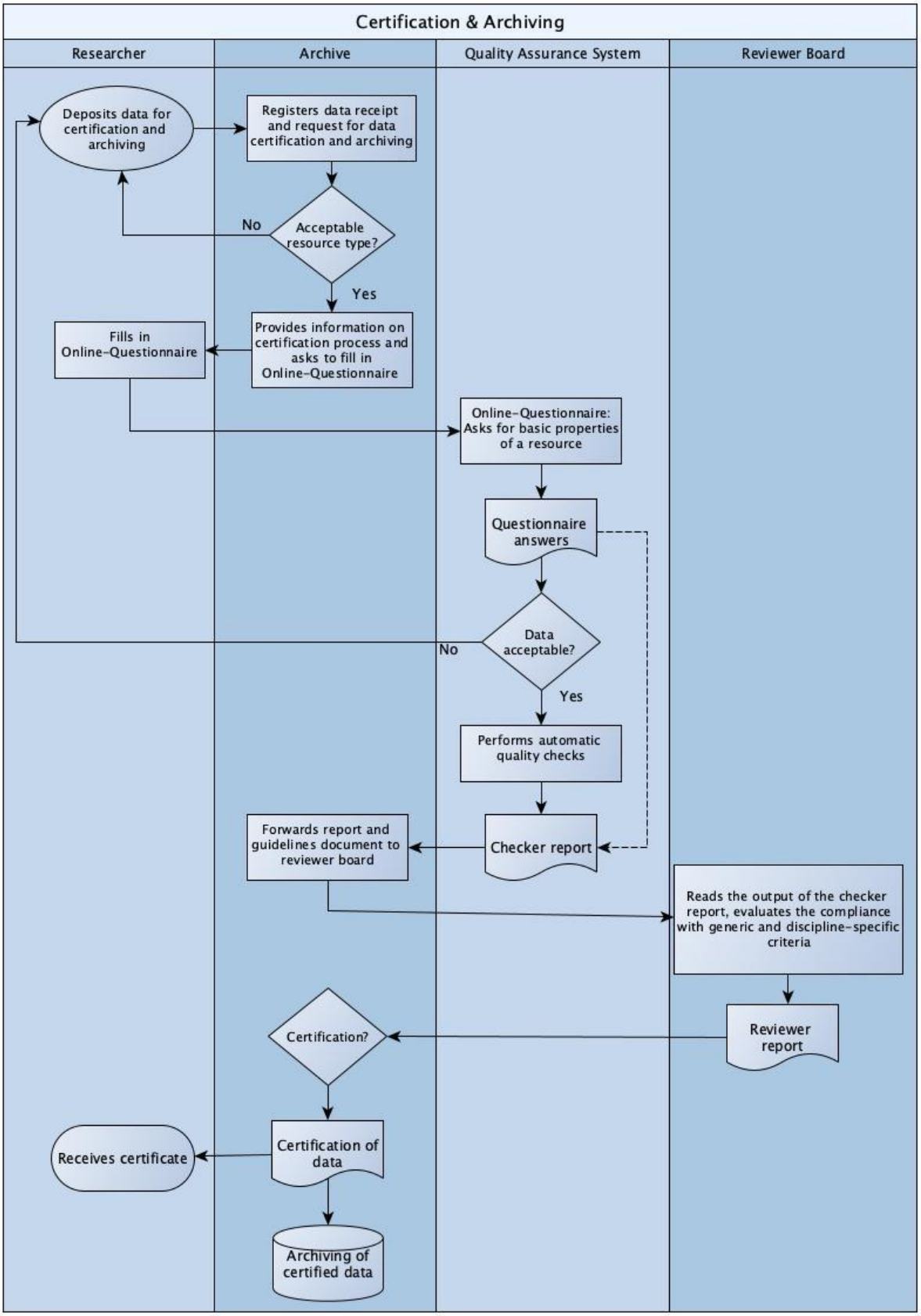
Within this use case, only automatically verifiable criteria are evaluated. Criteria that cannot be evaluated automatically are not carried out by the archive's staff.

[17]As QUEST provides services for ensuring and improving the quality of the data, the maturity level of a dataset/resource can improve by going through the evaluation process.

[18]https://gitlab.rrz.uni-hamburg.de/corpus-services/corpus-services

1.4.2 Use Case II: Certification & Archiving

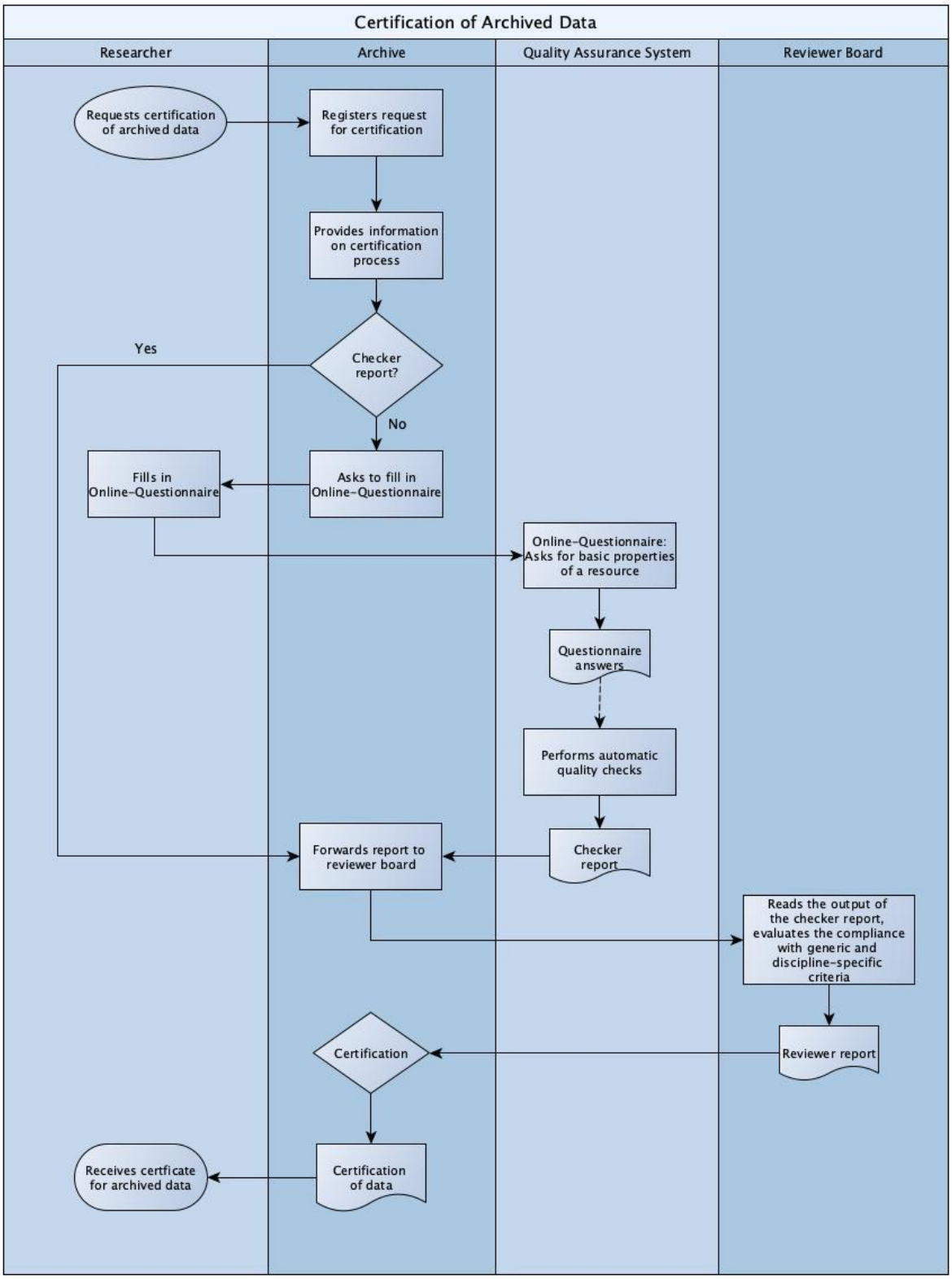
Diagram 2 *Use Case II: Certification & Archiving*



In the second use case, the evaluation system is used for data certification and archiving purposes. Consequently, a researcher approaches an archive with the intention of having some research data certified and archived. The archive again checks the general suitability of the resource, and this time initiates a certification process with filling out the online questionnaire in the first step. After checking the compliance with basic quality features, quality criteria are applied according to the discipline-specific classification of the resource with the help of the automatic checks. The report from the web-based quality checks is then forwarded to a discipline-specific reviewer board. The board checks the output and, if necessary, uses the guidelines to manually check the discipline-specific features of the resource and makes a recommendation regarding the certification of the data. Based on the reviewer report, the archive or the affiliated certification body makes a decision on certification. If, based on the output of the quality check and the manual review by the reviewers, the data do not meet the necessary requirements for certification, the researcher has the option of improving the data and having them re-evaluated. If the data is certified, it is to be archived at the same time. The researcher also receives a certificate.

1.4.3 Use Case III: Certification of Archived Data

Diagram 3 *Use Case III: Certification of Archived Data*



A third possible application of the evaluation system is its use for the purpose of certifying archived data. Provided that archiving has not already been preceded by a quality check with the tools described

and an up-to-date automatic checker report is not available, the data also pass through the online questionnaire and the automatic quality checks. Following the systematic evaluation process, a reviewer board examines the discipline-specific characteristics of the resource based on the automatic checker report and the guidelines. Finally, the archive decides on the certification of the resource.

2. Framework

The evaluation framework developed by QUEST consists of three elements:

1. Data Maturity Levels
2. Quality Criteria
3. Evaluation Methods

2.1 Data Maturity Levels

Reusability of research data is generally considered in relation to data quality, i.e., the properties of a resource to be reused. The relevance of this has recently come increasingly into focus while dealing with research data. In addition to the Council for Information Structures (RfII), which in its publication "The Data Quality Challenge. Recommendations for Sustainable Research in the Digital Turn" identifies the safeguarding and enhancement of data quality as a fundamental value of good scientific practice (RfII, 2020: 3), the DFG[19] also describes in its Impulse Paper on the Digital Transformation in the Sciences and Humanities the discussion of quality criteria and the establishment of metadata standards as an elementary component of research practice (DFG, 2020:9).

In the context of data quality, research data is nowadays often required to be FAIR. However, including the data quality dimensions, not all FAIR principles and the corresponding metrics or criteria are primarily related to the intrinsic properties of data, but also, for example, to the infrastructure required to make data discoverable and accessible, i.e., aspects that are beyond the control of the data producer (cf. Hedeland, 2022). In general, what is understood by data quality can be seen as highly contentious. With the concept of data maturity levels, introduced by Hedeland (2021), an approach is being pursued in the QUEST project that makes it possible to operationalize reusability. The concept considers the heterogeneity of the resource type dealt with and makes it possible to compare the reusability of language data according to their certain level of structuredness.

For this reason, the starting point for the assessment of data quality in the context of the QUEST project's quality initiative is the assignment of audiovisual, annotated language data to three different data maturity levels ranging from 0 to II. With reference to structuredness of data, this division has also been introduced to avoid a too strong focus on certain quality criteria preventing researchers from depositing or even assessing their data.

For more detailed information on this topic see Hedeland, 2021.

[19] Deutsche Forschungsgemeinschaft

The following tentative definitions of the three data maturity levels refer to generic quality criteria. For discipline-specific curation criteria, separate requirements for the different levels of data structuredness are defined and asserted.

Data Maturity Level 0 – Unstructured Data Set

- Basic structured metadata on the resource level, in particular legal aspects and provenance information on source and resource levels
- Data in readable, sustainable formats

Data Maturity Level I – Structured Data Set

In addition to level 0:

- Information on sampling and (in)completeness of annotations, transcriptions, etc.
- Basic structured metadata on resource part level including information on the recording situation and participants, structural metadata explaining the relations between parts
- Transcription/Annotation data is human-readable

Data Maturity Level II – Structured Data Set + Structured Data

In addition to level I:

- Machine-readable structural metadata with resolving paths to all resource parts
- Machine-readable consistent contextual information (on the recording situation and participants) allowing for the creation of subsets, including identification of design relevant elements for stratification
- Participants and their contributions can be identified across the resource
- Data and metadata are in formats recommended for the respective functional domains
- Transcription/Annotation data is machine-readable, syntactically correct and consistent
- Schema-based annotation levels are syntactically correct
- Basic transcription level is distinguishable if existing
- Transcription makes different information types explicit through machine-readable conventions or annotation structure
- Tokenization is possible through machine-readable conventions or annotation structure

2.2 Quality Criteria

Quality criteria established within QUEST are on the one hand based on generic quality standards for the different relevant resource types within the scope of audiovisual, annotated language data and their metadata regardless of an intended usage scenario. On the other hand, specific curation criteria are established and tailored to reuse scenarios related to individual disciplines and/or research methods.

Quality standards and curation criteria which are defined in this way can be used as minimum standards for the discipline specific reusability in the design and appraisal of future projects and thus provide information on the reuse potential of resources.

Compliance of a certain dataset with the quality criteria is to be evaluated on a modular basis. Therefore, QUEST presents its quality standards and curation criteria in generic and discipline-specific modules.

2.3 Evaluation Method: Quality Assurance Measures and Intended Outcome

Quality Criteria are intended to be used as a grounding set for the evaluation methodologies.
Each criterion has to contain the description both of an evaluation method and which results of this method are considered acceptable. The aim is to specify automatic evaluation methods where possible and resort to human evaluation otherwise. Automatic check procedures can provide results in three different categories: critical, warning or correct.

- Critical problems result in a failure to meet the criterion and the problems have to be resolved by the submitter.
- Warnings result in a temporary failure to meet the criterion and require additional human intervention. They can be overruled by human judgement.
- Correct results mean that the criterion is met by the submitted data, at least in respect to the automatic evaluation.

The results of the automatic checker together with a potential human review decide on the acceptance of the submitted data. In case of issues, a reviewer can review the reasons, propose fixes or suggest the classification as a different maturity level for which the criteria are met.

3. Quality Criteria

3.1 Overview

Table 1 *Overview of Quality Criteria*

Number	**Criterion**	**Data Maturity Level** (0 – II)	**Evaluation Method** (Automatic/Manual)
Generic Data and Metadata Recommendations for Audiovisual, Annotated Language Data			
A1.1	Data is provided in formats suitable for long term preservation	0	Automatic
A1.2	Transcription and/or annotation is provided in machine-readable or human-readable formats	I	Automatic & Manual
A1.3	Participants and their contributions can be identified across the resource	II	Automatic
A1.4	Data and metadata are in formats recommended for the respective functional domains	II	Automatic
A1.5	Transcription/Annotation data is machine-readable, syntactically correct and consistent	II	Automatic
A1.6	Schema-based annotation levels are syntactically correct	II	Automatic
A1.7	Basic transcription level is identifiable if existing	II	Automatic
A1.8	Transcription makes different information types explicit through machine-readable conventions or annotation structure	II	Automatic
A1.9	Tokenization is possible through machine-readable conventions or annotation structure	II	Automatic
A2.1	Basis metadata recommendations for audiovisual, language data	0	Automatic
A2.2	Information on sampling, (in)completeness and other relevant design aspects	I	Manual

A2.3	Basic structured metadata on dataset part level including information on the recording situation and participants	I	Automatic/Manual
A2.4	Structural metadata explaining the relations between parts	I	Automatic/Manual
A2.5	Machine-readable structural metadata with resolving links to all files	II	Automatic
A2.6	Machine-readable consistent contextual information	II	Semi-automatic
Discipline-Specific Recommendations: Language Documentation/RefCO			
B1.1	File format in the dataset should be portable	I	Automatic
B1.2	Naming files using explicit semantic conventions	II	Manual
B1.3	Providing an overview over the corpus	II	Automatic
B1.4	Describing the corpus compositum	II	Manual
B1.5	Describing the conventions associated with the tiers in a dataset	II	Manual
B1.6	Describing the glosses, abbreviations and symbols used in the dataset	II	Automatic/Manual
Discipline-Specific Recommendations: Learner Corpora			
B2.1	Design considerations for building learner corpora are met	/	Automatic
B2.2	Data is transcribed using standardised conventions and the transcription is documented	/	Automatic
B2.3	Annotation tageset is documented	/	Automatic/Manual
B2.4	Metadata recommendations for spoken learner corpora	/	Manual
Discipline-Specific Recommendations: Interpreted Communication			
B3.1	Transcription conventions are documented	/	Automatic

B3.2	The tier structure of the annotation file is described including annotation of multilingual phenomena	/	Automatic
B3.3	Corpus data are consistently translated	/	Automatic
B3.4	The languages of the corpus are consistently annotated	/	Automatic
B3.5	Metadata recommendations for interpreted corpora	/	Manual
Discipline-specific Recommendations: Sign Language Corpora			
B5.1	Annotation recommendations for sign language corpora are met	/	Automatic/Manual
B5.2	Metadata recommendations for sign language corpora are met	/	Manual
Discipline-specific Recommendations: Language Community			
B6.1	Data-recommendations for reuse by language communities	II	Automatic/Manual
B6.2	Dataset metadata recommendations for reuse by language communities	I	Automatic/Manual
B6.3	Individual recording metadata recommendations for reuse by language communities	II	Automatic/Manual
Discipline-specific Recommendations: Ethnography			
B7.1	Individual recording metadata recommendations for reuse in Ethnography	II	Automatic/Manual
Discipline-specific Recommendations: Dissemination of Oral History (meta)data via Europeana and libraries			
B8.1	Relevant (meta)data can be made publicly available	I	Automatic/Manual
B8.2	Data covers the desired content and its authenticity and integrity is guaranteed	I	Manual
B8.3	DDB/Europeana metadata requirements are met	I	Manual

B8.4	Library catalogue metadata requirements are met	I	Manual
B8.5	Transcripts include NE/EL annotation	II	Manual

3.2 Module A: Generic Recommendations for Audiovisual Data

Module **A** provides generic recommendations on data standards, such as technical standards (e.g., file formats), the structural and formal correctness of the data as well as content-related and methodological aspects of their presentation, for the various mentioned resource types and their associated metadata handled by QUEST. They should be considered when creating and curating audiovisual language data in order to ensure optimal and long-term reusability of the data.

After a brief note on the consideration of legal issues in relation to the reuse of language corpora, a comprehensive presentation of quality criteria at the data level follows, taking into account the different data maturity levels. Under A.2, criteria at metadata level are presented.

3.2.1 A0: Considerations on Legal Aspects

Exploiting the possibilities of reuse of audiovisual, annotated language data and, more generally, issues related to increasing data quality are linked to legal frameworks.

Due to their heterogeneity and sometimes very discipline-specific characteristics, audiovisual language data place high demands on the definition of standards and evaluation criteria regarding legal issues. In principle, however, legal and ethical aspects of the reuse and transfer of linguistic resources must definitely be taken into account and, if possible, clarified depending on the individual case. With a few exceptions, it is necessary in every case to ensure in advance that informed consent is obtained from all relevant stakeholders. Because of this and other aspects to be considered as relevant, we strongly recommend contacting a repository or a research data centre in advance.

For an basic overview of legal and ethical questions (see CARE-Principles[20]) that might arise in connection with the creation, usage and archiving of audiovisual, annotated corpora, we also refer to helpful resources such as the DFG handout "Informationen zu rechtlichen Aspekten bei der Handhabung von Sprachkorpora "[21] (Information about legal aspects for handling language corpora).The handout provides information on legal aspects that should be taken into account to ensure good data reusability. Further recommendations for the evaluation of legal aspects in the use and dissemination of spoken language corpora based on best practices can also be found, for example, in the shared Guidelines document from the AGD/HZSK[22].

Infrastructures such as CLARIAH-DE also address legal issues for research data in the humanities in the areas of personal rights, data protection and copyright law, among others, as well as ethical principles. Although CLARIAH-DE[23] itself does not offer binding legal advice, it lists helpful websites and documents from other partners on legally compliant handling of data protection, anonymization of data, rights of use etc.

Helpful user support in legal matters is also provided by a helpdesk[24] offered by CLARIAH-De.

---

[20] https://www.gida-global.org/care

[21] https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_recht.pdf

[22] https://corpora.uni hamburg.de/pdf/Leitfaden_Aufbereitungsaufwand_und_Nachnutzbarkeit_von_Korpora.pdf

[23] https://www.clariah.de/beratung-und-schulung/rechtliches-ethisches-lizenzen

[24] https://www.clarah.de/support

3.2.2 A1: Data Recommendations

(A1.1) Data is provided in formats suitable for long term preservation (Level 0)

Description

Data needs to be in formats suitable for long term preservation, i.e., they need to be readable, open and, if structured, well-documented. Since this type of resource will usually not be integrated into the local infrastructure and services, format recommendations from generic sources on digital preservation, such as the Library of Congress, are also valid.

Possible Approach to Implementation

For textual content, PDF/A or formats based on text (TXT, CSV, TSV) and XML are suitable, others as agreed. Audio data should be in e.g., LPCM-WAV, video data in e.g., MP4 (MPEG-4, Advanced Video Coding (Part 10) (H.264)). The exact parameters of audiovisual data should be in agreement with the requirements of the archive or research data centre when possible.

Example

Legacy scans of handwritten transcripts were converted from PDF to PDF/A.

How to measure it?

Automatic format validation to point out possibly non-acceptable formats.

What is a valid result?

All files must be accepted in the validation process.

Benefits

This data is not FAIR, but it can be provided in compliance with domain-relevant community standards (R1.3).

Related Resources

Format recommendations provided by CLARIN centres: https://clarin.ids-mannheim.de/standards/views/recommended-formats-with-search.xq

Format recommendations (in German) by the DFG: https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_sprachkorpora.pdf

Format descriptions by the Library of Congress: https://www.loc.gov/preservation/digital/formats/fdd/

(A1.2) Transcription and/or annotation data is provided in machine-readable or human-readable formats (Level I)

Description

Transcription and/or annotation data needs to be in formats that can be read and understood by humans. Designated annotation formats in TXT or XML are additionally machine-readable, but for this level various document formats such as MS Word (.docx) or PDF, but also plain text (TXT) formats only readable by interpreting humans are accepted. Transcription conventions and annotation schemas must be standardized in external documentation or documented for humans to be able to understand and reuse the data. The participants and the djaraziation must be inferable for the human reader. Machine-readability or reliable searching/querying is not required.

Possible Approach to Implementation

Apart from designated annotation formats, all standard document formats (.docx, odt, .pdf and txt formats) are acceptable. Images of textual (transcription/annotation) content in standard/open formats (e.g., GIF, JPEG, PNG, TIFF) are also acceptable if no text-based format is available and/or the content is handwritten.

Example

Transcripts were created with MS Word and the project-specific transcription conventions were thoroughly documented.

How to measure it?

Automatic format validation can point out non-acceptable file formats (cf. Level 0), but a manual assessment of the comprehensibility of transcription and annotation data is necessary for this criterion.

What is a valid result?

All files must be accepted in the validation process.

Benefits

This data is not FAIR, but it can be provided in compliance with domain-relevant community standards (R1.3).

Related Resources

Format recommendations provided by CLARIN centres: https://clarin.ids-mannheim.de/standards/views/recommended-formats-with-search.xq

Format recommendations (in German) by the DFG: https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_sprachkorpora.pdf

Format descriptions by the Library of Congress: https://www.loc.gov/preservation/digital/formats/fdd/

(A1.3) Participants and their contributions can be identified across the resource (Level II)

Description

All participants need to have unique names/codes that allow for identification of individual participants in transcription/annotation data and also, through an adequate data format, for automatic retrieval of their respective contributions.

Possible Approach to Implementation

Recommended formats for transcription/annotation data created by common tools, e.g., ELAN (EAF), EXMARaLDA (EXB), FOLKER etc. can fulfil this requirement, additionally a list or registry of the participants is necessary for comparison.

Example

Transcripts were created with EXMARaLDA Transcription and Annotation Tool while using the EXMARaLDA Corpus Manager to reliably manage participants across participants. The EXMARaLDA data model includes information on participants for all contributions (where applicable).

How to measure it?

EXMARaLDA has an option for automatic validation that assesses whether participant codes are used in transcripts that are not in the Corpus information. For other widely used formats, this needs to be checked by the Corpus Services. It is only possible to assess participant listing and consistency, not whether the correct codes have been used or participants have been mixed up.

What is a valid result?

All participants occurring in the transcription/annotation data are defined and described. All contributions include information on the participant (where applicable).

Benefits

This data is not FAIR, but provided in compliance with domain-relevant community standards (R1.3).

Related Resources

Format recommendations provided by CLARIN centres: https://clarin.ids-mannheim.de/standards/views/recommended-formats-with-search.xq

Format recommendations (in German) by the DFG: https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_sprachkorpora.pdf

(A1.4) Data and metadata are in formats recommended for the respective functional domains (Level II)

Description

All provided files need to be in the formats recommended for their respective functional domains, i.e., transcription/annotation data should be in preferably standardized formats tailored for this purpose and not in general purpose document formats.

Possible Approach to Implementation

Recommended formats for transcription/annotation data are those created by common tools, for the QUEST context mainly ELAN (EAF), EXMARaLDA (EXB) and FOLKER, and the ISO/TEI standard for transcription of spoken language. Structured metadata needs to be provided in specific formats according to the QUEST guidelines.

Example

Transcripts with annotations were created with ELAN and metadata is provided in the IMDI format.

How to measure it?

Format validation depending on the functional domain of the individual files can be done automatically if information on the purpose of individual files is provided.

What is a valid result?

All files are provided in formats recommended for their functional domain.

Benefits

This data is not FAIR but provided in compliance with domain-relevant community standards (R1.3).

Related Resources

Format recommendations provided by CLARIN centres: https://clarin.ids-mannheim.de/standards/views/recommended-formats-with-search.xq

Format recommendations (in German) by the DFG: https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_sprachkorpora.pdf

(A1.5) Transcription/Annotation data is machine-readable, syntactically correct and consistent (Level II)

Description

All files containing the transcription/annotation data need to be machine-readable, syntactically correct, and consistent to allow for reliable (automatic) querying and analysis. This applies both to the structural information or file format/data model and the tier structure.

Possible Approach to Implementation

Recommended formats for transcription/annotation data are structured and machine-readable. The transcription and annotation tool of the EXMARaLDA system provides validation of the generic tier structure, i.e., the data model. EXMARaLDA transcription files that pass structural assessment fulfil this requirement. FOLKER files that pass syntax checks also fulfil this requirement. For EAF data, tier structure needs to be validated. ISO/TEI data complying with the schema contains the relevant information as part of the structural information and thus also fulfils this requirement.

Example

The EXMARaLDA transcription and annotation tool was used to create and check transcription files, which all passed the tests.

How to measure it?

For EXMARaLDA, checks for structure, alignment, and segmentation (syntax) errors are part of the software. FOLKER also provides checks for syntax and alignment errors and does not allow for any further annotation tiers, thus will not contain structure errors. In ELAN, EAF files can be validated and a template can be used to dry-run the adaptation of data to its structure, i.e. to check tier structure. In all cases, tests for consistency of tier structure across the dataset are required.

What is a valid result?

All files pass all relevant tests.

Benefits

This data is not FAIR, but provided in compliance with domain-relevant community standards (R1.3).

Related Resources

Format recommendations provided by CLARIN centres: https://clarin.ids-mannheim.de/standards/views/recommended-formats-with-search.xq

Format recommendations (in German) by the DFG: https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_sprachkorpora.pdf

(A1.6) Schema-based annotation levels are syntactically correct (Level II)

Description

All annotation levels based on schemas/vocabularies only use the defined categories. If complex annotations such as glosses have been used, these should be parsable according to existing documented conventions.

Possible Approach to Implementation

For EXMARaLDA files with schema-based annotations, the use of an annotation specification file allows for consistent annotation. FOLKER has no annotation tiers. For EAF data, controlled vocabularies assist in creating consistent annotation.

Example

Annotations were created using controlled vocabularies in ELAN.

How to measure it?

For EXMARaLDA, Corpus Services provide a check against an annotation specification file or the generation of such a file from the transcription data to evaluate existing annotation. FOLKER does not allow for any further annotation tiers. In ELAN, controlled vocabularies can be evaluated. Complex annotations such as glosses require specific syntax checks.

What is a valid result?

All files pass all relevant tests.

Benefits

This data is not FAIR, but provided in compliance with domain-relevant community standards (R1.3).

Related Resources

Format recommendations provided by CLARIN centres: https://clarin.ids-mannheim.de/standards/views/recommended-formats-with-search.xq

Format recommendations (in German) by the DFG: https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_sprachkorpora.pdf

(A1.7) Basic transcription level is identifiable if existing (Level II)

Description

For all files containing transcription and annotation data, the basic orthographic transcription needs to be recognizable as such and separable from other annotation levels. This is a machine-readability requirement, i.e., the basic transcription level needs to be automatically recognizable.

Possible Approach to Implementation

In all recommended formats for transcription/annotation data, this can be achieved. In EXMARaLDA and FOLKER, the information is part of the data model, in EAF the information can be added to the tier definition. The structure of ISO/TEI is based on the transcription level.

Example

The EXMARaLDA transcription and annotation tool was used to create and evaluate transcription files, which all passed the structure tests.

How to measure it?

For EXMARaLDA and FOLKER, there are checks in place, in ELAN, the information can be added to the tier, but this is rarely done. ISO/TEI is built around the transcription level and can be validated as XML.

What is a valid result?

All files contain the relevant information and pass all relevant tests.

Benefits

This data is not FAIR, but provided in compliance with domain-relevant community standards (R1.3).

Related Resources

Format recommendations provided by CLARIN centres: https://clarin.ids-mannheim.de/standards/views/recommended-formats-with-search.xq

Format recommendations (in German) by the DFG: https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_sprachkorpora.pdf

(A1.8) Transcription makes different information types explicit through machine-readable conventions or annotation structure (Level II)

Description

The transcription usually contains more than words, e.g., pauses, non-verbal behaviour and sometimes comments on prosody or pronunciation. The conventions must ensure that these non-tokens are recognizable as such.

Possible Approach to Implementation

EXMARaLDA supports several transcription systems, FOLKER is based on the GAT system, and both tools produce data that fulfils this requirement if it passes the transcription convention/syntax checks. For EAF data, transcription conventions need to be evaluated or the information has to be provided in different tiers. Tokenized ISO/TEI data complying with the schema differentiates between these types of information as part of the structural information.

Example

The EXMARaLDA transcription and annotation tool was used to create and evaluate transcription files using the cGAT segmentation algorithm, and all files passed the tests/segmentation.

How to measure it?

For EXMARaLDA and FOLKER, checks for the supported transcription systems are part of the software. In ELAN, the different types of information might be in different tiers, otherwise the transcription conventions need to make the different types of information clear and there needs to be an evaluation of the syntax of the transcription convention.

What is a valid result?

All files pass all relevant tests.

Benefits

This data is not FAIR, but provided in compliance with domain-relevant community standards (R1.3).

Related Resources

Format recommendations provided by CLARIN centres: https://clarin.ids-mannheim.de/standards/views/recommended-formats-with-search.xq

Format recommendations (in German) by the DFG: https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_sprachkorpora.pdf

(A1.9) Tokenization is possible through machine-readable conventions or annotation structure (Level II)

Description

All transcription data should be tokenizable in the sense that words and other elements and larger segments such as intonation phrases or utterances should either be recognizable from machine-readable conventions, or by structuring the transcription tier using events/slots, or a tokenization should be provided in a separate (dependent) tier. The tokenization should include the units defined by the transcription conventions, usually words/tokens, non-token material, and some larger unit such as utterances or intonation phrases. If a tokenization is not provided, the transcription conventions need to be documented to allow for automatic tokenization of the transcribed text. Only machine-readable conventions can be automatically tokenized.

Possible Approach to Implementation

EXMARaLDA supports several transcription systems, FOLKER is based on the GAT system, and both tools produce data that fulfils this requirement if it passes the transcription convention/syntax checks. For EAF data, transcription conventions need to be evaluated or the information has provided in different tiers. Tokenized ISO/TEI data complying with the schema includes this information as part of the structural information.

Example

The EXMARaLDA transcription and annotation tool was used to create and evaluate transcription files using the cGAT segmentation algorithm, and all files passed the tests/segmentation.

How to measure it?

For EXMARaLDA and FOLKER, checks for the supported transcription systems are part of the software. ELAN data is more often tokenized by tier structure than by transcription conventions, but the transcription conventions should still be used consistently. If they provide tokenization information, automatic checks and processing is required. Automatic checks of transcription systems in turn require machine-readable conventions, i.e. not conventions that rely on human interpretation to decide which of several meanings a symbol has at a certain position.

What is a valid result?

All files pass all relevant tests.

Benefits

This data is not FAIR, but provided in compliance with domain-relevant community standards (R1.3).

Related Resources

Format recommendations provided by CLARIN centres: https://clarin.ids-mannheim.de/standards/views/recommended-formats-with-search.xq

Format recommendations (in German) by the DFG: https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_sprachkorpora.pdf

3.2.3 A2: Metadata Recommendations

(A2.1) Basic metadata recommendations for audiovisual, language data (Level 0)

Description

This set of metadata recommendations for assessing the quality and interoperability of metadata for audiovisual language data resources does not specify a particular metadata format nor do they require any particular serialisation, as various research communities use metadata formats and serialisation techniques that are established in their specific field, which can differ from established formats in other fields.

The current document *Metadata Recommendations for Audio-Visual Language Data* (DOI: 10.5281/zenodo.7346840) contains the modules Generic Research Data and Language Data that specify recommendations for dataset level metadata for these respective aspects of research data.

The recommendations consist of three modules: Generic (basic), Generic (extended) and Language Data. The module Generic (basic) groups the use cases DOI/DataCite, GDS/Schema.org, Citation and Legal together. We consider the Use Cases DOI/DataCite, Citation and Legal essential for any dataset regardless of its data type and scientific background. The requirements of the GDS/Schema.org scenario sit within the scope of this module and are integrated in it. The module Generic (extended) supplements the first module by optional properties from these use cases. The module Language Data covers the requirements from the use cases concerning the visibility in the language data-specific search portals OLAC and VLO. The following set of basic metadata would make an ideal metadata sheet. Mandatory elements will have a (M) behind them:

- Generic (basic): Identifier (M), Title (M), Description (M), Version, Keywords, Licence (M), Rightsholder, Access Rights, Publication Year (M), Publisher (M), Creator (M) and Contributor.
- Generic (extended): sameAs, isPartOf, hasPart, isBasedOn and Project.
- Language Data: Object Language (M), Linguistic Data Type and Modality.

Additionally, human readable information about provenance and structure of the data should be included in some way. This can be done either in the metadata, a PDF or text document or written on a sheet of paper, that will be given to the archive with the data. Information about provenance should report on when, where and by whom the data was recorded. When describing the structure of the data, it is important to include information on how individual recordings and/or documents are related to each other. Do individual recordings make up one bigger recording session, but were separated because of various elicitation tasks? Are there any documents like annotations or transcripts that need to be seen in relation to certain recordings?

Possible Approach to Implementation

Every archive or depository has their own internal processes to assess whether a data set is suitable to be archived or deposited, usually by validating serialised XML-files. Additionally, some archives and depositors offer tools that help to create uniform and suitable metadata according to their own standards. DataCite offers Fabrica specifically for members, but there are also free tools like EXMARaLDA.

Example

This example represents a metadata file of an imaginary dataset that would resemble an ideal file (both mandatory and optional fields filled out). Here we use OLAC which is widely used in Language

Documentation. Other metadata schemas are also viable. How certain fields translate across the schemas is described in our recommendation document (link). Some schemas do not have all elements mentioned above. OLAC for example has no special tag to denote the modality of a resource or to contact details of a person or organisation.

<olac:olacxsi:schemaLocation="http://www.language-archives.org/OLAC/1.1/ http://www.language-archives.org/OLAC/1.1/olac.xsd">
<dc:identifier xsi:type="dcterms:URI"> http://dx.doi.org/immaginaryDOI1
</dc:identifier>
<title>Recordings of Swabian spoken in Stuttgart</title>
<creator>Imanginatrix, Ingrid</creator>
<subject>Dialect description</subject>
<subject xsi:type="olac:language" olac:code="swg">Swabian</subject>
<language xsi:type="olac:language" olac:code="de">German</language>
<language xsi:type="olac:language" olac:code="en">English</language>
<description>Adult speakers producing speech in Swabian</description>
<publisher>IMG Publishing</publisher>
<contributor xsi:type="olac:role" olac:code="author">Imanginatrix, Ingrid</contributor>
<contributor xsi:type="olac:role" olac:code="speaker">Mustermann, Max</contributor>
<contributor xsi:type="olac:role" olac:code="annotator">Notata, Nina</contributor>
<dcterms:isPartOf>http://dx.doi.org/immaginaryDOI2</dcterms:isPartOf>
<dc:date xsi:type="dcterms:W3CDTF">2010</dc:date>
<dcterms:license>http://creativecommons.org/licenses/by-sa/3.0/
</dcterms:license>
<dc:type xsi:type="dcterms:DCMIType">Collection</dc:type>
<type xsi:type="olac:linguistic-type" olac:code="language_description"/>
</olac:olac>

How to measure it?

To check whether these recommendations are met by your metadata, our automated checker can be used to scan whether the necessary tags are used and whether they are filled with valid values. This automated checker provides a list of various issues if any were found. There are critical and non-critical issues that the data provider can look over, but also a peer review will be set in place to ensure the metadata fulfils community standards, if necessary.

What is a valid result?

Any dataset that provides the metadata recommended in this document in a formally identifiable and machine-readable way will be considered compliant with these recommendations. Metadata that does not follow the recommendations, does not automatically fail the certification process. As there are corpora and collections that do not follow these standards, which does not mean that they are unscientific. Some metadata schemata or previous best practices do not make it possible to follow

the recommendations exactly as we stated them. In such cases corpora and collections have to be reviewed manually and guidelines for this specific case will be provided to the data provider, so they can still have their corpus or collection certified.

Benefits

Findability is addressed by considering the requirements of search portals and data portals such as the CLARIN Virtual Language Observatory, Open Language Archives Catalogue and Search, as well as Google Dataset Search. The recommendations concerning licensing, access rights as well as referencing with persistent identifiers address **accessibility** and **reusability**. To ensure the greatest possible **interoperability**, the recommendations rely as much as possible on pre-existing well-defined metadata properties from widely used vocabularies. The recommendations utilise properties defined by Dublin Core Terms and Schema.org, FOAF, and SKOS.

Related Resources

https://dublincore.org/specifications/dublin-core/dcmi-terms/
http://purl.org/dc/terms
https://schema.org/
https://schema.datacite.org/
https://www.w3.org/TR/vocab-dcat-2/
http://www.language-archives.org/OLAC/metadata.html
https://web.archive.org/web/20090411054359/
https://www.mpi.nl/IMDI/documents/Pro-posals/IMDI_MetaData_3.0.4.pdf
http://www.meta-share.org/assets/pdf/META-SHARE_Documentation_v3.1.pdf
https://vlo.clarin.eu/search?1
https://catalog.clarin.eu/ds/ComponentRegis-try/#/?itemId=clarin.eu:cr1:p_1427452477080&registrySpace=published
http://agd.ids-mannheim.de/download/metadaten_schemata_DGD_2.0_2009-09-01.pdf
https://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-teiHeader.html
https://talkbank.org/manuals/CHAT.html#_Toc40195866
https://www.wikidata.org/

(A2.2) Information on sampling, (in)completeness and other relevant design aspects (Level I)

Description

Since it is expected that this resource was compiled for a purpose or has some kind of thematic distinction, the sampling decisions, known aspects of (in)completeness and other relevant design aspects should be described.

Possible Approach to Implementation

The description element of the metadata contains the relevant information.

Example

<dc:description>[...] Spontaneous and elicited speech data was collected from three geographical areas (North, Centre and South) from two age groups (10-30 and 31-80). Not all older participants could perform the task as noted in [...]</dc:description>

How to measure it?

This needs to be done manually.

What is a valid result?

All information is available and understandable.

Benefits

This data is not FAIR, but serves F2, R1 and R1.3.

Related Resources

See Level 0.

(A2.3) Basic structured metadata on dataset part level including information on the recording situation and participants (Level I)

Description

On this level, each session or recording with or without a transcript requires its own individual description with information on the recording situation and the participants. Apart from the basic metadata required on resource level for data maturity level 0, this should also cover the content of the dataset part itself. There is no requirement on consistent description of participants or the use of reliable codes or identifiers that identify the participants across the dataset.

Possible Approach to Implementation

Each transcript comes with a table header including values for “setting”, “participants” and “language use” to describe the content, and “recording information” to give information on how the recording was done and which parts are represented in which transcripts.
Alternatively, if the dataset is structured into dataset parts (bundles, sessions) that are individually described with their own metadata file additional information in the transcripts is not necessary.

Example

Some metadata formats allow to include that information in dedicated tags, in other instances project metadata conventions were extended to include the relevant information, which was managed by using the description tag (OLAC) or the transcript documents (COMA).
The following example describes an imaginary resource using the IMDI-Schema:

```
<?xml version="1.0" encoding="UTF-8"?>
<METATRANSCRIPT xmlns="http://www.mpi.nl/IMDI/Schema/IMDI"
                xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance”
                ArchiveHandle="hdl:ImaginaryHandle"
                Date="2022-01-10"
                FormatId="IMDI 3.03"
                Originator="Hand"
                Type="SESSION"
                Version="1"
        xsi:schemaLocation="http://www.mpi.nl/IMDI/Schema/IMDI ./IMDI_3.0.xsd"
>
    <Session>
    <!--... all the relevant metadata from the previous criteria...-->
        <Content>
            <Genre>Discourse</Genre>
            <SubGenre>Conversation</SubGenre>
            <CommunicationContext>
                <Interactivity>Interactive</Interactivity>
                <PlanningType>Semi-spontaneous</PlanningType>
                <Involvement>Non-elicited</Involvement>
                <SocialContext>Controlled Environment</SocialContext>
```

```
                    <EventStructure>Dialogue</EventStructure>
                    <Channel>Face to Face</Channel>
                </CommunicationContext>
                <Task>Travel planning</Task>
                <Modalities>Speech</Modalities>
                <Subject>Planning a trip to Madrid</Subject>
                <Languages>
                    <Language>
                        <Name>Swabian</Name>
                        <Id>swg</Id>
                        <Dominant>True</Dominant>
                    </Language>
                    <Language>
                        <Name>German</Name>
                        <Id>deu</Id>
                        <Dominant>False</Dominant>
                    </Language>
                </Languages>
                <Description></Description>
            </Content>
            <Actors>
                <Actor>
                    <Role>Author</Role>
                    <Name>Ingrid Immaginatrix</Name>
                    <Contact>ingrid@immaginatrix.de</Contact>
                </Actor>
                <Actor>
                    <Role>Speaker</Role>
                    <Name>Adult01</Name>
                    <Code>Sib1</Code>
                    <FamilySocialRole>Sibling</FamilySocialRole>
                    <Languages>
                        <Language>
                            <Id>swg</Id>
                            <Name>Swabian</Name>
                            <MotherTongue>True</MotherTongue>
                            <PrimaryLanguage>True</PrimaryLanguage>
                        </Language>
                        <Language>
                            <Id>deu</Id>
                            <Name>German</Name>
                            <MotherTongue>True</MotherTongue>
```

```
                                        <PrimaryLanguage>False</PrimaryLanguage>
                                    </Language>
                                </Languages>
                                <EthnicGroup>White</EthnicGroup>
                                <Age>30</Age>
                                <Sex>Male</Sex>
                                <Education>Abitur, equivalent to A-levels</Education>
                                <Anonymized>True</Anonymized>
                            </Actor>
                            <Actor>
                                <Role>Speaker</Role>
                                <Name>Adult02</Name>
                                <Code>Sib2</Code>
                                <FamilySocialRole>Sibling</FamilySocialRole>
                                <Languages>
                                    <Language>
                                        <Id>swg</Id>
                                        <Name>Swabian</Name>
                                        <MotherTongue>True</MotherTongue>
                                        <PrimaryLanguage>True</PrimaryLanguage>
                                    </Language>
                                    <Language>
                                        <Id>deu</Id>
                                        <Name>German</Name>
                                        <MotherTongue>True</MotherTongue>
                                        <PrimaryLanguage>False</PrimaryLanguage>
                                    </Language>
                                </Languages>
                                <EthnicGroup>White</EthnicGroup>
                                <Age>28</Age>
                                <Sex>Male</Sex>
                                <Education>Abitur, equivalent to A-levels</Education>
                                <Anonymized>True</Anonymized>
                            </Actor>
                        </Actors>
                        <!-- any other relevant metadata -->
                    </Session>
</METATRANSCRIPT>
```

How to measure it?

Automatic checks will not be enough since there is no standardisation of how to document the relevant information across various communities. Peer reviews will be necessary to ensure all relevant information is included in the metadata.

What is a valid result?

All information is available and understandable.

Benefits

This data is not FAIR, but serves F2, R1 and R1.3.

(A2.4) Structural metadata explaining the relations between parts (Level I)

Description

On this level, each session or recording with or without a transcript requires its own individual description. Apart from the basic metadata required on dataset level for data maturity level 0, this should also describe the structure of the dataset part (e.g., a Session/Bundle/Communication) itself, i.e. what files it comprises and how these are related to another, and its relation to other dataset parts.

Possible Approach to Implementation

Choosing a metadata format that has dedicated tags for the relevant information or a document like a Word file or an Excel sheet containing an inventory with file locations and paths and describing the files' naming scheme. Any other relevant information, like split recordings, for example motivated by a change of elicitation task or conversation topic, should also be explained in these documents.

Additionally, it is necessary to be consequent with data pairs/bundles/sessions which should contain only one annotation per recording.

Example

Some metadata formats allow to include that information in dedicated tags, in other instances project metadata conventions were extended to include the relevant information, which was managed by using the transcript documents.

Here, an imaginary DataCite example will show how to include information on the relation between several parts of a resource:

```
<resource xsi:schemaLocation="http://datacite.org/schema/kernel-4
https://schema.datacite.org/meta/kernel-4.4/metadata.xsd">
    <identifier identifierType="DOI">10.856132/imaginary-example</identifier>
    <!-- relevant metadata mentioned in the other criteria-->
    <resourceType resourceTypeGeneral="Text">Imaginary Metadata<resourceType>
    <relatedItems>
        <relatedItem relationType="IsPartOf" relatedItemType="Collection">
            <relatedItemIdentifier relatedItemIdentifierType="DOI">10.856132/im-
            aginary-collection<relatedItemIdentifier>
            <titles>
                <title>The Swabian Corpus</title>
            </titles>
            <publicationYear>2022</publicationYear>
        </relatedItem>
        <relatedItem relationType="IsMetadataFor" relatedItemType="Audiovisual">
            <relatedItemIdentifier relatedItemIdentifierType="DOI">10.856132/im-
            aginary-recording>
            <titles>
                <title>Two adults speaking Swabian</title>
```

```
                    </titles>
                    <publicationYear>2022</publicationYear>
                </relatedItem>
            </relatedItems>
            <!--other relevant metadata-->
</resource>
```

How to measure it?

Automatic checks will not be enough since there is no standardisation of the relevant information. Some only use certain metadata schemata that might not contain dedicated tags, but use Word files or Excel sheets instead. Also these documents are used differently across various communities, which is why a peer review will be necessary to ensure the metadata follows community standards.

What is a valid result?

All information is available and understandable.

Benefits

This data is not FAIR, but serves F2, R1 and R1.3.

(A.2.5) Machine-readable structural metadata with resolving links to all files (Level II)

Description

The structure of the resource should be described by machine-readable metadata including resolvable links to all files. DOI, for example, are types of URIs that identify objects like academic articles or datasets and point towards its location, thus making it findable and retrievable.

Possible Approach to Implementation

The dataset metadata file includes information on the types and characteristics of the resource parts and holds a resolvable link to each file.

Example

The EXMARaLDA Corpus Manager includes the Element <NSLink> for all files of abstract Transcription elements. IMDI has a section in their metadata schema that is dedicated to listing all relevant files of a recording session and making it possible to insert a direct link to the file.

```
<?xml version="1.0" encoding="UTF-8"?>
<METATRANSCRIPT xmlns="http://www.mpi.nl/IMDI/Schema/IMDI"
                xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
                ArchiveHandle="hdl:ImaginaryHandle"
                Date="2022-01-10"
                FormatId="IMDI 3.03"
                Originator="Hand"
                Type="SESSION"
                Version="1"
            xsi:schemaLocation="http://www.mpi.nl/IMDI/Schema/IMDI ./IMDI_3.0.xsd"
>
    <Session>
        <!-- relevant metadata mentioned in the other criteria-->
        <Resources>
            <MediaFile>
                <Type>Video</Type>
                <Format>Video/MPEG4</Format>
                <ResourceLink>www.immaginaryswabiancorpus.de/recording123/video.mpeg4</ResourceLink>
                <Quality>4</Quality>
                <RecordingConditions>To record the session a Panasonic HC-X1E has been used in a room with foam material on the walls to ensure good audio quality</RecordingConditions>
                <Access>
                    <Availability>CC BY</Availability>
                    <Date>2022-01-10</Date>
                    <Owner>Ingrid Immaginatrix</Owner>
                    <Publisher>Universität Tübingen</Publisher>
```

```
                                    <Contact>
                                        <Name>Ingrid Immaginatrix</Name>
                                        <Address> </Address>
                                        <Email>ingrid@immaginatrix.de</Email>
                                        <Organisation>Universität Tübingen</Organi-
                                        sation>
                                    </Contact>
                                </Access>
                            </MediaFile>
                        </Resources>
                        <!--other relevant metadata-->
                    </Session>
</METATRANSCRIPT>
```

How to measure it?

Several formats can be automatically validated by testing all links. There should be no files without links and descriptions in the deposit and no non-resolving resource links in the metadata. The formats also need to include information on the type of object the files represent, i.e. if it is the recording, a transcript or some photography documenting the recording session etc.

What is a valid result?

The format includes the relevant information, and all files are available.

Benefits

This data is not FAIR but provided in compliance with domain-relevant community standards (R1.3).

(A2.6) Machine-readable consistent contextual information (Level II)

Description

For this level, the contextual information on the recording situation and the participant needs to be machine-readable and consistent. This information describes any relevant elements (about the participants for example) that help to derive subsets which can be used for further research (e.g., investigating only recordings of a certain age group).

Possible Approach to Implementation

The contextual information is encoded into the metadata format using controlled vocabularies where possible and data types for numbers and dates to ensure comparability. The design-relevant elements are included using e.g., the age groups or geographical distinctions defined for the creation of the resource.

Example

The EXMARaLDA Corpus Manager can be used to encode key-value pairs using the project's vocabularies. But also other metadata schemas like IMDI or CMDI can be used.

```
<?xml version="1.0" encoding="UTF-8"?>
<METATRANSCRIPT xmlns="http://www.mpi.nl/IMDI/Schema/IMDI"
                xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
                ArchiveHandle="hdl:ImmaginaryHandle"
                Date="2022-01-10"
                FormatId="IMDI 3.03"
                Originator="Hand"
                Type="SESSION"
                Version="1"
                xsi:schemaLocation="http://www.mpi.nl/IMDI/Schema/IMDI ./IMDI_3.0.xsd"
>
    <Session>
        <Name>Swab_Adult01</Name>
        <Title>Two adults conversing in Swabian</Title>
        <Date>2022-01-10</Date>
        <Description>This is an imaginary recording of two adults who have grown up in
        Stuttgart and have been given an elicitation task, where they plan an imaginary trip
        to Madrid.</Description>
        <Project>
            <Name>SwabianCorpusStuttgart</Name>
            <Title>The Swabian of Stuttgart</Title>
            <Id>Swab01</Id>
            <Contact>Deutsches Seminar
            Wilhelmstraße 50
            72074 Tübingen</Contact>
            <Organisation>Universität Tübingen</Organisation>
```

```
        <Description>A corpus of Swabian spoken in Stuttgart. Fifty people aged
        25-55 were given various tasks (planning tasks, frog story, etc.) to elicit
        natural speech. All participants grew up in Stuttgart.</Description>
</Project>
<Content>
        <Genre>Discourse</Genre>
        <SubGenre>Conversation</SubGenre>
        <CommunicationContext>
                <Interactivity>Interactive</Interactivity>
                <PlanningType>Semi-spontaneous</PlanningType>
                <Involvement>Non-elicited</Involvement>
                <SocialContext>Controlled Environment</SocialContext>
                <EventStructure>Dialogue</EventStructure>
                <Channel>Face to Face</Channel>
        </CommunicationContext>
        <Task>Travel planning</Task>
        <Modalities>Speech</Modalities>
        <Subject>Planning a trip to Madrid</Subject>
        <Languages>
                <Language>
                        <Name>Swabian</Name>
                        <Id>swg</Id>
                        <Dominant>True</Dominant>
                </Language>
                <Language>
                        <Name>German</Name>
                        <Id>deu</Id>
                        <Dominant>False</Dominant>
                </Language>
        </Languages>
        <Description></Description>
</Content>
<Actors>
        <Actor>
                <Role>Author</Role>
                <Name>Ingrid Immaginatrix</Name>
                <Contact>ingrid@immaginatrix.de</Contact>
        </Actor>
        <Actor>
                <Role>Speaker</Role>
                <Name>Adult01</Name>
                <Code>Sib1</Code>
```

```
            <FamilySocialRole>Sibling</FamilySocialRole>
            <Languages>
                <Language>
                    <Id>swg</Id>
                    <Name>Swabian</Name>
                    <MotherTongue>True</MotherTongue>
                    <PrimaryLanguage>True</PrimaryLanguage>
                </Language>
                <Language>
                    <Id>deu</Id>
                    <Name>German</Name>
                    <MotherTongue>True</MotherTongue>
                    <PrimaryLanguage>False</PrimaryLanguage>
                </Language>
            </Languages>
            <EthnicGroup>White</EthnicGroup>
            <Age>30</Age>
            <Sex>Male</Sex>
            <Education>Abitur, equivalent to A-levels</Education>
            <Anonymized>True</Anonymized>
        </Actor>
        <Actor>
            <Role>Speaker</Role>
            <Name>Adult02</Name>
            <Code>Sib2</Code>
            <FamilySocialRole>Sibling</FamilySocialRole>
            <Languages>
                <Language>
                    <Id>swg</Id>
                    <Name>Swabian</Name>
                    <MotherTongue>True</MotherTongue>
                    <PrimaryLanguage>True</PrimaryLanguage>
                </Language>
                <Language>
                    <Id>deu</Id>
                    <Name>German</Name>
                    <MotherTongue>True</MotherTongue>
                    <PrimaryLanguage>False</PrimaryLanguage>
                </Language>
            </Languages>
            <EthnicGroup>White</EthnicGroup>
            <Age>28</Age>
```

```
                    <Sex>Male</Sex>
                    <Education>Abitur, equivalent to A-levels</Education>
                    <Anonymized>True</Anonymized>
                </Actor>
            </Actors>
            <Resources>
                <MediaFile>
                    <Type>Video</Type>
                    <Format>Video/MPEG4</Format>
                    <ResourceLink>www.immaginaryswabiancorpus.de/record-
                    ing123/video.mpeg4</ResourceLink>
                    <Quality>4</Quality>
                    <RecordingConditions>To record the session a Panasonic HC-X1E
                    has been used in a room with foam material on the walls to ensure
                    good audio quality</RecordingConditions>
                    <Access>
                        <Availability>CC BY</Availability>
                        <Date>2022-01-10</Date>
                        <Owner>Ingrid Immaginatrix</Owner>
                        <Publisher>Universität Tübingen</Publisher>
                        <Contact>
                            <Name>Ingrid Immaginatrix</Name>
                            <Address></Address>
                            <Email>ingrid@immaginatrix.de</Email>
                            <Organisation>Universität Tübingen</Organi-
                            sation>
                        </Contact>
                    </Access>
                </MediaFile>
                <WrittenResource>
                    <Type>Annotation</Type>
                    <SubType>Phonology</SubType>
                    <ResourceLink>www.immaginaryswabiancorpus.de/record-
                    ing123/annotation.txt</ResourceLink>
                    <Validation>
                        <Methodology>Semi-Automatic</Methodology>
                    </Validation>
                    <Derivation>Annotation</Derivation>
                    <Access>
                        <Availability>CC BY</Availability>
                        <Date>2022-01-10</Date>
                        <Owner>Ingrid Immaginatrix</Owner>
```

```
                                    <Publisher>Universität Tübingen</Publisher>
                                    <Contact>
                                        <Name>Ingrid Immaginatrix</Name>
                                        <Address></Address>
                                        <Email>ingrid@immaginatrix.de</Email>
                                        <Organisation>Universität Tübingen</Organi-
                                        sation>
                                    </Contact>
                                </Access>
                            </WrittenResource>
                        </Resources>
                    </Session>
</METATRANSCRIPT>
```

How to measure it?

Consistency can be measured semi-automatically, automatic validation is possible when vocabularies are also available and machine-readable.

What is a valid result?

Information is available and consistent.

Benefits

This data is not FAIR, but serves F2, R1 and R1.3.

3.3 Module B: Discipline-specific Recommendations for Audiovisual Data

The common core of quality standards that should be met by all types of resources in the field of audiovisual language data (Module A) is extended in the modules related to B by discipline-specific criteria based on specific use cases in the broad disciplines of language documentation, language typology, sign language and multilingualism research.

The modules contain specific curation criteria[25] that should be applied depending on the respective use case and whose definition serves to exploit the reuse potential of research data across disciplinary boundaries. In addition to the discipline-specific criteria for the reuse scenarios 'Language Documentation', 'Learner Corpora', 'Interpreted Communication' and 'Sign Language', curation criteria are defined for the non-scientific (Third-Mission) reuse ('Oral History', 'Anthropology', 'Ethnography') for cultural and other social purposes.

[25] Curation criteria are operationalizable criteria that allow statements about the reuse potential of resources in specific scenarios.

3.3.1 Module B1: Language Documentation/RefCo

Language documentation datasets are sets of audio and video recordings and their associated annotation, whose purpose is to provide a representation of a language. Their subject are communities for which few records, if any, are available. One of the core goals of any language documentation dataset is to produce recordings and annotations that will be reusable, either by other researchers or the community that was documented. While language documentation projects often focus on a specific speech community, it is not a requirement, and some projects can for instance focus on a particular topic, such as speech genre or theme. There is no restrictions regarding the topics of a language documentation dataset, as long as the documentation provides recordings and linguistic annotations that are following the current practice in the field.

Data Recommendations

Data in the context of a language documentation dataset can be divided into two types: recordings, whether they are written, audio or audiovisual, and annotation data, such as transcription and glossing (morphological, syntactical or another linguistic information). The recommendations we are providing here deal with annotation data and the conventions used to produce them. Regarding the production of recordings, as QUEST and RefCo are not evaluating the quality of video or audio recordings, one should consult dedicated resources on how to select the proper recording equipment and file formats currently used in the field, such as:

- Seyfeddinipur, M., & Rau, F. (2020). Keeping it real: Video data in language documentation and language archiving. *Language Documentation & Conservation*, *14*, 503-519.

Data Maturity Level 0

At this level, any set of recordings, with or without annotations, regardless of their quality, as long as it is representing an under-documented language, is acceptable. As any data regarding under-documented language is valuable, it is important to handle datasets which have not been documented.

Data Maturity Level I

For this level, the annotations within a dataset should be recording using machine-readable file formats. The recordings should be transcribed, glossed and translated. Audio and audiovisual recordings should at least have been transcribed and translated.

Data Maturity Level II

In order to reach the level II, a dataset should have been produced following to a conventionalized structure and have a unified machine-readable documentation. This documentation should include metadata regarding the represented languages, the corpus design, and statistics describing the size of the corpus, transcription and annotation guidelines and the various annotation schemes used. Regarding the annotations, in addition to a transcription and translation, the transcription should be glossed morphologically.

(B1.1) File format in the dataset should be portable (Level I)

Description

The portability of the file formats used in a dataset consists first, in ensuring that files are open-access and documented, so that any entity can implement the readability of the format on the long term. Then, it also requires a file to be valid in regard of their own file format standard. Finally, one should pay attention to whether the file formats used in a dataset are maintained, that is if there exists any entity ensuring that these formats can be updated to a version currently recognized by the application reading it; or ensuring that there exists an application that can run on current operating systems and read the files in the dataset.

Possible Approach to Implementation

One should check that each file format in the dataset is a portable file, that is, at least that the file formats are open-access and documented. If possible, one should also verify that their maintenance is ensured on the long term.

How to measure it?

Each file in the dataset should be checked in order to know whether their format is:
Open-access,
documented,
maintained.
The Facile platform provides an interface for validating that files are well formatted. They also ensure the portability of these formats.

What is a valid result?

'Failure state': Not all the files in the dataset are using an open-access format.

'Acceptable state': All file formats used in the dataset are open-access, but their portability is not assured by an organism.

'Desirable state': All the file formats used in the dataset are open-access and portable.

Benefits

Ensuring the portability of a dataset ensures that, potential reusers will be able to find a suitable software to access and read the content of a dataset in the future.

Related Resources

S Bird, G Simons – Language, 2003, Seven dimensions of portability for language documentation and description, https://arxiv.org/pdf/cs/0204020

The Facile platform developed by the CINES provides a service for ensuring that files are valid according to the conventions of their format and that their portability is ensured: https://facile.cines.fr/

(B1.2) Naming files using explicit semantic conventions (Level II)

Description

When reusing a dataset, one hast to start selecting the files within it that will be relevant to their study. In order to facilitate the identification of those files, an issue really important when one is reusing various files from multiple datasets, proper file naming practices should be followed.

Possible Approach to Implementation

A possible approach for dataset submitter to implement a file naming policy would be to use systematically a certain amount of semantic tags for their files names. These tags could be meta information such as, the date (YYYYMMDD), the speech genre, the place of recording, and finally a title.

How to measure it?

A manual reviewer should read the file names in the submitter dataset to ensure that they follow some systematic conventions and that they would enable a potential reuser to identify them uniquely within the dataset.

What is a valid result?

'Failure state': File names without any systematic conventions and without any obvious semantic meaning, such as "a.eaf", or "2.eaf".

'Acceptable state': A systematic file naming convention across the dataset.

'Desirable state': A systematic file naming convention across the dataset using semantic tags.

Benefits

Potential reusers will be able to quickly identify the files in a dataset, have an idea of their content, and be able to differentiate them from other files, whether they come from the same dataset or another.

Metadata Recommendations

(B1.3) Providing an overview over the corpus (Level II)

Description

The overview metadata should provide enough information about the dataset so that potential re-users can decide whether the dataset is relevant and reusable for their research project, without having to actually read the content of the set.

Possible Approach to Implementation

One should provide the following information regarding their dataset:

Dataset creator name, Dataset creator contact, Dataset creator institution, stable URL towards the dataset, license for the data, license for the annotation.

How to measure it?

RefCo provided instructions for evaluation the metadata given by a dataset creator. An automatic check is provided by QUEST to ensure the relevance of the metadata provided.

What is a valid result?

'Failure state': No metadata for an overview of the dataset content.

'Acceptable state': A description of the dataset containing metadata described in the Possible Approach to Implementation section.

'Desirable state': A machine-readable description of the metadata.

Benefits

Potential reusers will be able to quickly evaluate, through the size of the dataset, its content, and its access rights, whether a given dataset is relevant or not for their projects.

Related Resources

For RefCo, see https://zenodo.org/record/7380448, and in particular the Corpus Documentation template, and its Overview tab.

(B1.4) Describing the corpus composition (Level II)

Description

A dataset should be provided with a set of metadata that describe how the files are related to each other. Often, files within a language documentation dataset are organised into recording sessions, but it is not always the case. Still, regardless of the strategy adopted by the corpus creator, each recording session should specify where and when it was done, with who, and the list of files associated with the session.

Possible Approach to Implementation

A corpus creator should provide the following information regarding the recordings in their dataset: Sessions, a listing of the files (recordings, annotations, notes, etc.) associated with each sessions, place of recording, speech genre, speakers, their age at the time of the recording, their gender.

The CorpusComposition tab of the RefCo corpus documentation or Lameta[26] or SayMore provide an interface for describing this set of metadata.

How to measure it?

If the description has been designed by the corpus submitter, check whether or not the information provided allows to fill all of the required set of metadata. If the description has been made using Lameta, SayMore or RefCo, ensure that all the metadata required by these software have been actually provided.

What is a valid result?

'Failure state': No information given regarding the corpus composition, or incomplete information.

'Acceptable state': A description designed by the corpus creator that provides the set of metadata required by the RefCo corpus Documentation.

'Desirable state': A machine-readable description of the corpus composition, such as the one produced with Lameta, SayMore or the RefCo Corpus Documentation.

Benefits

Dataset reusers will be able to identify, for each recording session, which are the file pertaining to it, as well as have access to a minimal set of metadata describing each recording situation and the persons who were recorded.

Related Resources

Lameta is software which helps the fieldworker to organize their collections and annotation files. https://github.com/laMETA

SayMore was developed in order to help fieldworkers organizing their collections of recordings and annotation files, as well as producing BOLD annotation on them: https://software.sil.org/saymore

RefCo Corpus Documentation also provides a tab for describing these information, see https://zenodo.org/record/7380448#.Y4tsPH2ZNPY

(B1.5) Describing the conventions associated with the tiers used in a dataset (Level II)

Description

To be understandable by someone else than the corpus creator, the conventions used for naming and organizing the annotation tiers of a dataset has to be described. This is particularly important when a corpus creator defines their own original tiers. Furthermore, a corpus should follow only one annotation structure, that is, a unique description should be sufficient for the whole corpus.

Possible Approach to Implementation

The AnnotationTiers tab of the RefCo Corpus Documentation's spreadsheet provides a set of metadata for describing a dataset's tier structure.

How to measure it?

A reviewer should check whether the tier names are meaningful, and that the corpus submitter provided a description of the function associated with these tiers.

What is a valid result?

'Failure state': No description has been provided by the corpus submitter.

'Acceptable state': A corpus provided with an ad-hoc file describing the functions associated to each tier of a corpus.

'Desirable state': A machine-readable description of the tiers names and their functions has been provided with the dataset by the corpus submitter.

Benefits

A reuser will be able to clearly understand the functions associated with the annotation tiers in a given corpus.

Related Resources

For RefCo Documentation, see https://zenodo.org/record/7380448#.Y4tsPH2ZNPY

[26] https://sites.google.com/site/metadatatooldiscussion/home

(B1.6) Describing the glosses, abbreviations and symbols used in the dataset (Level II)

Description

There exists various conventions regarding the glosses, abbreviations and symbols used for annotating language corpora, such as Leipzig Glossing Rules, Jefferson's convention, Universal Dependencies, etc. But whatever are the conventions a corpus creator uses, they can often have to come up with new conventions, or even their own particular scheme, suitable for their project. In order to ensure the reusability of their annotations, they should provide a description of the glosses, abbreviations and symbols used in their dataset.

Possible Approach to Implementation

The RefCo's Corpus Documentation provides an implementation or reference. The corpus submitter should fill the Glosses, Punctuation and Transcription tabs of the RefCo Corpus Documentation's spreadsheet in order to provide the required information.

How to measure it?

A reviewer should ensure that the list of all glosses, abbreviations and symbols and, if relevant[27], their meaning variations depending are describing the linguistic units has been provided. They should check first that all the glosses used in the corpus are described, and then that only glosses found in the corpus are described. If a corpus submitted provided a RefCo corpus documentation, the RefCo checker will be able to check the adequacy of the list and report errors whenever it is not the case.

What is a valid result?

'Failure state': No description given or not associated with the dataset content.

'Acceptable state': A description of the glossing conventions using well-defined criteria or well-known standards (like LGR).

'Desirable state': A machine-readable description of all the symbols, glosses and abbreviations used.

Benefits

Potential reusers will be able to unambiguously interpret the glosses, abbreviations and symbols used in the dataset. If a machine-readable description is provided, then it will be possible to perform automatic checks on the data to ensure the coherency between the description and the data.

Related Resources

See for instance the Leipzig Glossing Rules for recommendations that should be applied regarding the glossing of annotation tiers: https://www.eva.mpg.de/lingua/resources/glossing-rules.php

Jefferson transcription conventions: https://www.universitytranscriptions.co.uk/jefferson-transcription-system-a-guide-to-the-symbols/

IPA: https://www.internationalphoneticassociation.org/content/full-ipa-chart

X-SAMPA: https://www.phon.ucl.ac.uk/home/sampa/x-sampa.htm

3.3.2 Module B2: Learner Corpora

Learner corpora consist of electronic collections of natural or near natural written or/and oral production by L2 or L3/Lx learners or users which are built and published according to certain design criteria. Such samples can be used to answer a variety of research questions not only in second language research and pedagogy but also in other disciplines such sociology, psychology and even neurology.

This module contains recommendations for learner corpora datasets including metadata, annotation and documentation that are crucial for making learner corpora findable, accessible, interoperable and reusable (cf. FAIR-principles).

Data Recommendations

Data Maturity Level 0

Data on this level is not relevant for this reuse scenario.

Data Maturity Level I

At this level a dataset should be documented in machine-readable form according to the conventions used within the field of learner language research. The recordings should be orthographically transcribed and anonymized. At least transcripts of the audio files should be available.

Data Maturity Level II

At this level a dataset should have a unified documentation including represented languages, metadata, corpus design, statistics of the corpus, transcription and annotation guidelines, annotation schemes used and a unified search interface to query the corpus (König et al. 2021).

[27]It is often the case that a symbol, such as a dash "-", changes of meaning depending on the annotation tier it is found. Every different usage has to be properly defined.

(B2.1) Design considerations for building learner corpora are met

Description

Learner corpora need to be specified according to strict, transparent and systematic design criteria. All decisions and specifications of the design criteria should be well documented and made available to the corpus users to increase the reusability of a corpus. If data is gathered without documentation of learner-, language- and task variables, the resulting corpus will be not of much use.

Possible Approach to Implementation

For this to be implemented we recommend to take into account the following decisions[28]:

- The decision about the target population: e.g. learning environment, age, nationality, mother tongue background of the learner group
- The decision about how the data is collected: naturalistic production (e.g. recorded natural speech), elicited production (e.g. role play) or experimental production
- The decision about how to sample the particular learner group: how much from one particular learner group and how often and for how long to be sampled
- The decisions about possible variations in learner corpus design: language-related criteria (e.g. mode, genre, style, topic), task-related criteria (data collection, data elicitation, time limitations, use of references), learner-related criteria (e.g. age, motivation and attitude, learning context, L1 background, L2 proficiency)

Example

Language-related feature:
Mode (written, spoken, multimodal)
Genre (letter, diary, fiction, essay)
Style (narration, argumentation)
Topic (general, leisure..)

Task-related feature:
Data collection (cross-sectional, longitudinal)
Elicitation (spontaneous, prepared)
Use of references (dictionary, source text)
Time limitation (fixed, free, homework)

Learner-related features:
Internal-cognitive (age, cognitive style)
Internal-affective (motivation, attitude)
L1 background
L2 environment (ESL/EFL, level of school)
L2 proficiency (standard test score)

How to measure it?

[28] Considered also as the best recommendations for the major categories proposed by Tono, 2003.

The presence of documentation can be automatically checked. Documentation about corpus design (corpus design metadata), about the decisions taken during the design phase: specification of the design criteria.

What is a valid result?

'Failure state': Documentation is missing or not complete, metadata about the corpus design is missing or incomplete.

'Desirable state': Documentation both about corpus design and about corpus design metadata.

(B2.2) Data is transcribed using standardised conventions and (accordingly documented) the transcription is documented

Description

For the analysis to be possible, spoken learner corpora should be transcribed and the presence of transcription should be documented. Standardized conventions should be used to ensure consistency in the transcription of data and conversion into other formats.

For instance:
CHAT transcription format is frequently used in the community and recommended in a number of articles (MacWhinney 2017a). The CHILDES-handbook offers important information about this transcription format. Transcripts in CHAT format can be automatically converted into the formats required for Praat (praat.org), Phon (phonbank.talkbank.org), ELAN (tla.mpi.nl/tools/elan), CoNLL, ANVIL (anvil-software.org), EXMARaLDA (exmaralda.org), LIPP (ihsys.com), SALT (saltsoftware.com), LENA (lenafoundation.org), Transcriber (trans.sourceforge.net), and ANNIS (corpustools.org/ANNIS) (cf. https://talkbank.org/manuals/CHAT.pdf)

Possible Approach to Implementation

This criterion can be automatically checked, if the transcription file is documented and has a function "transcription". Then it can be checked, whether the format used is suitable for archiving.

What is a valid result?

'Desirable state': Transcription file available and accordingly characterised.

Benefits

Increase collaboration and corpora sharing.

Related Resources

https://talkbank.org/manuals/CHAT.pdf

(B2.3) Annotation tagset is documented

Description

Linguistic annotation makes it possible to sort and compare learner corpora. For learner corpora the annotation needs to be documented in the file list with function "annotation". It is possible to retrieve linguistic patterns such as errors or grammatical categories, and it can support the identification of learner language use. It can be done manually or automatically. One type of manual annotation is error annotation. For the error tagging/annotation a multi-layer corpus standoff architecture is very useful (cf. Lüdeling et al. 2005). The most frequent type of automatic annotation is POS-tagging and Parsing. The annotation should be consistent and accurate and the raw text should always remain recoverable.

An important step is the evaluation of annotation: one has a gold standard and evaluates it against this corpus or one uses several annotators to annotate the same sub corpus using the same tagset and guidelines and evaluates how often and where they agree (called inter-annotator agreement). This step assures the consistency of annotation and is very important.

How to measure it?

For this to be measured we recommend the documentation of the annotation conventions (tagset). If the annotation file is documented in the data list "function", it will allow the automatic check of the consistency of the annotation and then a respective checker can be used.
For the validation of the annotations, manual checking is needed.

What is a valid result?

'Desirable state': Annotation tagset is documented in the data list and is in appropriate format (tagset is correct).

Benefits

Reusability and comparability

Metadata Recommendations

(B2.4) Metadata set recommendations for spoken learner corpora

Description

There is no standardised set for learner corpus metadata. There are some efforts: Granger and Paquot (2017) proposed a core metadata set for learner corpora (L2). They design a flexible system that allows, depending on the focus of research, description of the main variables that have an influence on the learner use of L2 with the possibility to add or to expand the elements in the metadata set.

This set consists of five main components: administrative metadata, corpus design metadata, corpus annotation metadata, task and learner metadata. Some categories are obligatory and essential for the learner language research such as target language (languages) of the corpus, L1(s) and other L2 language(s) in the learner's environment, proficiency level, nationality of the participants, place and institution (cf. Stemle et al. 2019).

The administrative metadata comprises general information on the data (corpus title) and its provenance including information about the data collectors and all persons involved in the production and processing of the corpus, also the information about the availability, information about a licence or other legal agreements, character and markup information.

The corpus design metadata describe the research design of the corpus.

Corpus annotation metadata give information about the conducted annotation and processing activities (e.g. error annotation, linguistic annotation).

Learner metadata describe the learner language(s), language background, learner language proficiency and their characteristics.

Possible Approach to Implementation

Learner corpora metadata are present in a different format and must be manually checked. The automatic evaluation is not possible.

3.3.3 Module B3: Interpreted Communication

Community Interpreting Corpora contain audio and/or video recordings of various types of community interpreted discourse (e.g. medical interpreting (doctor-patient communication, simulated doctor-patient communication) and legal interpreting (e. g. courtroom communication)) conducted with the help of the consecutive or simultaneous interpreting or both.

Data Recommendations

(B3.1) The transcription conventions are documented

Description

This description should provide the transcription convention used and documentation of the transcription conventions and the tools for the transcription (like ELAN or EXMARaLDA). This criterion is useful for further analysis of the corpus. There are several transcription systems available to the researchers with important theoretical variations. (cf. CHAT, HIAT, LIDES) The overall recommendation is to try to transcribe in the simplest way possible and be non-theory dependent.

Possible Approach to Implementation

A list to describe the conventions.

Example

CHAT, HIAT as well-known standards and LIDES coding manual (LIPPS Group 2000), Speech recognition software - depending on how to segment the turns of speakers and to use existing documented conventions.

How to measure it?

The transcription tier contains an orthographic transcription of the recording (orthographic rules depend on language-specific conventions). Non-verbal elements - project-specific conventions. The transcription in EXMARaLDA-Format can be checked, whether there is a particular transcription convention. Otherwise, this will be indicated as an error. The list of the extended conventions: CHAT, HIAT, LIDES, GAT.

What is a valid result?

'Failure State': No transcription convention provided.

'Desirable state': Documentation of the transcription conventions or using well-known transcription conventions (like CHAT, HIAT, LIDES, GAT, cGAT… coding manual)

Related Resources

Gardner-Chloros P., Moyer M., Sebba M. (2007) Coding and Analysing Multilingual Data: The LIDES Project. In: Beal J.C., Corrigan K.P., Moisl H.L. (eds) Creating and Digitizing Language Corpora. Palgrave Macmillan, London. https://doi.org/10.1057/9780230223936_5

Rehbein, J.; Schmidt, T.; Meyer, B.; Watzke, F. & Herkenrath, A. (2004). Handbuch für das computergestützte Transkribieren nach HIAT. In: Arbeiten zur Mehrsprachigkeit, Folge B 56.

CHAT: https://talkbank.org/manuals/CHAT.pdf

(B3.2) The tier structure of the annotation file is described including annotation of multilingual phenomena

Description

This criterion describes a kind of annotation applied to the corpus and the transparent documentation of the annotation set and additionally the annotation of the multilingual phenomena e.g. code-switching, code-mixing, interjections.

How to measure it?

The consistency between the corpus annotation and the according corpus documentation can be checked.

What is a valid result?

'Failure state': No description of the annotation set used.

'Desirable state': Annotation is consistent according to the documented annotation set.

(B3.3) Corpus data are consistently translated

Description

The corpus should be at least translated in more widely accessible languages. The annotations in such a case should match source and target pairs.

Possible Approach to Implementation

The translation tiers are named.

Example

A corpus tier structure should contain a tier for translation.

How to measure it?

If the translation tier is named, it can be checked, whether this tier is consistent all over the corpus.

(B3.4) The languages of the corpus are consistently annotated

Description

This criterion should describe the languages involved in the interaction according to common conventions in the research area.

Possible Approach to Implementation

All the languages in the corpus should be annotated.

Example

Lang tier in the ComInDat Corpus.

How to measure it?

If the languages are annotated, it is possible to check the consistency of the annotation.

Metadata Recommendations

(B3.5) Metadata recommendations for interpreting corpora

Metadata of the community interpreted corpora depends on the corpus type, purpose and the public status of the given corpus.

Description

This set of metadata describes the event and discourse context / types, qualifications of interpreters and preparation of interpreters, spontaneity index.

Possible Approach to Implementation

Example

Background information
Communication type
Discourse type
Project name
Communication location: type, duration, time
Language(s): languageCode, Language status (institution language, interpreting language)

Metadata describing speaker(s) and interpreters

This set of metadata gives information about speakers, the role in the communication, the status of the language(s), regional variety of the language
Interpreters: gender, level of expertise, native language, language combination

Metadata describing speech event

Description

This set of metadata describes the communication event, its location, languages, mode, topic, the quality of sound and image, setting, speed of delivery

Possible Approach to Implementation

Example

Background information
Communication type
Discourse type
Project name
Communication location: type, duration, time
Language(s): language code, Language status (institution language, interpreting language)

Metadata describing translation status

The translation languages (original language and translation language, interpreted language) should be documented and the translation conventions status, translation modality and translation mode. When possible dialectal variation within one language should also be described

Metadata describing language of an utterance

The status of the languages should be indicated: source and target

Information of the language affiliation of each utterance

Categorisation of phenomena of bilingual speech (definitions)

3.3.4 Module B4: Anonymisation of Multimodal Corpora

The reason that corpora are anonymised is to make sure that no personal information is shared without the informed consent of the person concerned. Informed consent itself is not a simple concept, and the issues around it vary depending for instance on the size of the community in which the corpus is collected, the nature of the corpus content and the technological background of the subjects. (Crasborn, 2010; Rock, 2001; McEnery and Hardie, 2011; Singleton et al., 2014; Schembri et al., 2013).

In addition to personal information pertaining to the participants in a corpus, there are often mentions of third parties, who have not been asked for or given any kind of consent for information about them to be shared publicly. This is particularly problematic with small communities, where it is often easy to identify a person from minimal amounts of information.

The process of anonymisation is expensive and time-consuming. Many corpus projects take the decision to only release all or parts of the data to researchers who have signed a confidentiality agreement which also specifies how the data may be used.

If anonymisation of a multimodal corpus is going to be carried out, it can be applied to all the modalities which make up the data: video, audio and annotations.[29]

Video

What to anonymise

There may be signs in the background which give away locations, or documents on the screen which contain personal information pertaining to participants. The entities which should be anonymised are personal names and place names which can be used to identify a person. In some cases, particularly where the recording is of a member of a small community, more data may need to be anonymised such as professions or unusual personal characteristics (e.g. pink hair, pronounced limp etc). Natural Language Processing (NLP) methods for Named Entity Recognition can be run over the transcription of a corpus to identify the names of people and places which may need to be anonymised. It is often considered not necessary to obscure the names of famous people and large places (e.g. countries, cities), so further pre-processing may be applied to exclude these (Bleicken et al, 2013).

Sign language corpora

For a sign language corpus, it is impossible to hide the faces of the participants, as the mouthing and other facial gestures play a significant communicative role. The face, hands and/or body can be temporarily obscured in order to hide parts of the conversation which concern personal information about the participants or third parties.

Spoken language corpora

Facial expressions also play a role in spoken communication, but depending on the type of corpus it might be possible to hide the participants' faces while still retaining some useful video information. If parts of the audio are being anonymized, it is possible to also obscure the speaker's mouth at this point to avoid anyone being able to lip-read sensitive information.

[29]These considerations for anonymisation were presented at LREC 2022.

How to anonymise

Video can be anonymized with blurring, pixelation, video filters or by adding black shapes over parts of the video.

Blurring, pixelation and filters can be added using a variety of video processing software.

The creators of the CASE corpus of Skype dialogues wanted to hide the identity of the participants. They experimented with video anonymisation using various pixel, art, and transformation filters in Adobe Premiere, and chose a contour filter. They carried out control tests and concluded that when this filter was used, subjects did not recognise themselves (Diemer et al., 2016).

Researchers are working with video processing techniques which can anonymise a participant in a corpus by changing their appearance. Xia et al (2022) recently reported on a model which can hide the identity of signers, but evaluation studies have not yet been carried out.

Audio

Audio can be anonymized by playing a sound over the words which need to be obscured. This can consist of a "beep" or of white or brown noise, which many people find less disturbing while listening to speech (Schiel and Kisler, 2019; Kisler et al., 2017).

Annotations

When names are anonymised in the annotations, they can either be replaced with alternative names (e.g. Mary changed to Sarah) or by types, which can optionally include a number so that it is clear when a single entity is referred to multiple times (e.g. Mary changed to NAME or NAME23). Where a profession is anonymised, these can also be replaced with names (e.g. butcher changed to baker or to PROFESSION1). Personal characteristics are more complicated and should be dealt with on a case by case basis, and whole phrases may need to be removed (e.g. the man with the limp changed to PERSONAL-DESCRIPTION).

3.3.5 Module B5: Sign Language Corpora

Sign language corpora consist of videos made of one or more participants by one camera or multiple cameras filming simultaneously from different angles. There are most often two participants taking part in a dialogue, sometimes with a moderator present, or a single person telling a story or recounting an experience.

Data Recommendations

(B5.1) Recommendations for the annotation of sign language corpora

Description

There is no standard for the annotation of sign language corpora, but ID-glosses have emerged as the most important base annotation layer (Johnston, 2010). Other annotations depend on the research aims of the group involved in the annotation effort. Sign language corpora are most often annotated using the ELAN (Wittenburg et al., 2006) or iLex (Hanke & Storz, 2008) software.
There is no transcription standard for sign languages; there are several notation systems but all are very time-consuming to apply and are not used as the base transcription for corpora. The most-used systems are Hamnosys (Hanke, 2004) and Signwriting (Sutton, 1995 and Sutton, 2009).
Sign language corpora are often translated into the regionally collocated spoken language and also English, if that is not the regional language. This can be free translation or word-by-word translation.

How to measure it?

Gloss checking has been implemented for some corpora. If glossing conventions have been made available and they have been integrated into the tool developed in A1.5, we can check if the glosses in a corpus follow the appropriate convention.

What is a valid result?

'Failure state': if there is a gloss checker for this corpus and the tests are failed.

'Desirable state': if there is a gloss checker for this corpus and the tests are passed.

Related Resources

Hanke, T. (2004). HamNoSys – Representing Sign Language Data in Language Resources and Language Processing Contexts. Proceedings of the LREC2004 Workshop on the Representation and Processing of Sign Languages: From SignWriting to Image Processing. Information Techniques and Their Implications for Teaching, Documentation and Communication, 1–6.

Hanke, T. & Storz, J. (2008). iLex - A Database Tool for Integrating Sign Language Corpus Linguistics and Sign Language Lexicography. In Proceedings of the 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora. (pp. 64-67). ELRA.

Prillwitz, S. et al., 1987. HamNoSys. Hamburg Notation System for Sign Languages. An introduction. Hamburg: Zentrum für Deutsche Gebärdensprache.

Johnston, T. (2010). From archive to corpus: Transcription and annotation in the creation of signed language corpora. International Journal of Corpus Linguistics, 15(1), 106–131.

Sutton, V. (1995). Lessons in sign writing. SignWriting.

Sutton, V. (2009). Signwriting: sign languages are written languages. Center for Sutton Movement Writing, CSMW, Tech. Rep.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: a professional framework for multimodality research. In Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation. (pp. 1556-1559). ELRA.

Metadata Recommendations

(B5.2) Metadata recommendations for sign language corpora

Description

There are extra requirements for sign language corpus metadata in addition to the requirements of language corpora in general. Metadata standards were generally designed with spoken language in mind.

Possible Approach to Implementation

As a result of the extra requirements, the European Cultural Heritage Online (ECHO) project held workshops from 2003-2007 with the goal of establishing a Sign Language Metadata standard, using the ISLE Metadata Initiative (IMDI) scheme. They established a set of properties in the "Content" and "Actors" sections of the IMDI scheme, listed below. More details of the properties and their allowed values can be found in Crasborn and Hanke, 2003b and Crasborn et al, 2007.

4. Content
   * Language Variety Elicitation Method
   * Interpreting Source
   * Interpreting Target
   * Interpreting Visibility
   * Interpreting Audience

5. Actors
   * Deafness: Status
   * Deafness: Aid Type
   * Sign Language Experience: Exposure Age
   * Sign Language Experience: Acquisition Location
   * Sign Language Experience: Sign Teaching
   * Family: [Mother/Father/Partner]: Deafness

The IMDI standard was lacking some features which are important for multimodal and particularly sign language data, so more recently a Component Metadata Initiative (CMDI) standard was proposed (Freigang et al 2014, ISO 24622-1:2015 (2015)) based on ISOcat data categories for describing signed language resources compiled and implemented by Crasborn and colleagues (Crasborn and Hanke, 2003a; Crasborn and Hanke, 2003b; Crasborn and Windhouwer, 2012). It includes more properties which refer to the multimodal aspects of sign language such as participant handedness, and detailed elements on the language background of participants. It also allows more elaborate descriptions of the technologies used in recordings, as for example HD videos. Information can be also found in the Clarin Concept Registry, where "Sign Language" can be selected as a Concept Scheme https://concepts.clarin.eu/ccr/browser/index.php?key=&termsOr=true&matchTermsExact=true&facet0=ALL&facet1_10=http%3A%2F%2Fhdl.handle.net%2F11459%2FCCR_P-SignLanguage_8f51ce1b-211d-9682-a8e7-aeb0f2a79e03&facet3=ALL&facet4=meertens

However, there are still some aspects missing from the existing metadata standards. For instance, if multiple cameras are used, which is standard for the recording of a sign language corpus, there is no standard for stating the angles at which the cameras are placed.

Many corpora, including sign language corpora, have parts, which are available to the public, or to researchers, and other parts which are private and used only by the research team which collected them. In these cases, there is usually personal metadata about the participants which is not released when parts of the corpus are made public. It would be beneficial to have a flag on certain metadata properties to signify that these should be omitted when the metadata is made public. For instance this could include exact ages, which would instead become age ranges.
Many sign language corpora do not have metadata in a standard format, or do not make the details of their metadata publicly available. The necessary complexity of the IMDI and CMDI formats means that the creation of the metadata for a corpus requires significant expertise and effort which is often not available to smaller projects.

A list of European sign language corpora, which includes information on their metadata if they have made it available, can be found in Kopf, Schulder and Hanke, 2021.

The IMDI standard is used among others by the Australian Sign Language (Auslan) corps (https://www.auslan.org.au/about/corpus), the Dogon Sign Language corpus (https://archive.mpi.nl/tla/islandora/object/tla%3A1839_00_0000_0000_0016_2E9E_4) the British Sign Language (BSL) corpus (https://bslcorpusproject.org), the Sign Language of the Netherlands (NGT) corpus (https://www.ru.nl/cls/our-research/researchgroups/sign-language-linguistics/completedprojects/completed-projects/corpus-ngt), the corpus of Finnish Sign Language (FinSL and FinSSL) (https://www.jyu.fi/hytk/fi/laitokset/kivi/opiskelu/tutkintoohjelmat-ja-oppiaineet/viittomakieli/tutkimus-2/suomenviittomakielten-korpusprojekti), the CREAGEST corpus of French Sign Language (LSF) (https://www.ortolang.fr/market/corpora/ortolang-000926), the Dicta-Sign corpus of BSL, German Sign Language (DGS), Greek Sign Language and LSF (https://www.sign-lang.uni-hamburg.de/dicta-sign/portal), the ECHO corpus of NGT, BSL and Swedish Sign Language (STS) (http://sign-lang.ruhosting.nl/echo) and the Vidi Sign Space Corpus of DGS and Turkish Sign Language (TİD) (https://www.nwo.nl/en/projects/276-70-009),

The CMDI standard is used among others by the German Sign Language DGS corpus (http://dgs-korpus.de), the Giving Cognition a Hand corpus of Turkish Sign Language (TİD) and Sign Language of the Netherlands (NGT), the IPROSLA corpus of NGT (https://www.ru.nl/cls/our-research/researchgroups/sign-language-linguistics/completedprojects/completed-projects/iprosla-gebaren-groei) and the Visibase corpus of NGT (https://hdl.handle.net/1839/00-0000-0000-0004-DF8F-4).

How to measure it?
If CMDI or IMDI metadata are available, these can be automatically checked with the same procedure as for the generic metadata. If metadata are present in a different format, they must be manually checked.

What is a valid result?
'Failure state': the metadata do not pass the checks

'Desirable state': the metadata pass the checks

Benefits

Adequate metadata improve the chances of corpus reuse.

Related Resources

Crasborn, O. (2010). The Sign Linguistics Corpora Network: Towards standards for signed language resources. Proceedings of the Seventh International Conference on Language Resources and Evaluation ({LREC}'10), 4. http://www.lrec-conf.org/proceedings/lrec2010/pdf/25_Paper.pdf

Crasborn, O. and Hanke, T. (2003a). Additions to the IMDI metadata set for sign language corpora. Agreements at an ECHO workshop, May 8–9, 2003, Radboud University, Nijmegen. http://sign-lang.ruhosting.nl/echo/docs/SignMetadata_Oct2003.pdf

Crasborn, O. & Hanke, T. (2003b). Metadata for sign language corpora. Background docu- ment for an ECHO workshop, 8–9 May 2003, Nijmegen University. Available at: http:// www.let.ru.nl/sign-lang/echo/docs/Metadata_SL.doc

Crasborn, O. and Menzo Windhouwer (2012). ISOcat data categories for signed language resources. In Efthimiou, Eleni and Kouroupetroglou, Georgios and Fotinea, Stavroula-Evita, editor, Gestures in embodied communication and human-computer interaction, pages 118–128. Springer.

Crasborn, O. A., Mesch, J., Waters, D., Nonhebel, A., Kooij, E. van der, Woll, B., & Bergman, B. (2007). Sharing sign language data online: Experiences from the ECHO project. International Journal of Corpus Linguistics, 12(4), 535–562. https://doi.org/10/gjn7v9

Freigang F, Priesters M, Nishio R, Bergmann K (2014). Sharing Multimodal Data: A Novel Metadata Session Profile for Multimodal Corpora. In Selected Papers from the CLARIN 2014 Conference. Odijk J (Ed); Linköping Electronic Conference Proceedings: 25-35. https://pub.uni-bielefeld.de/record/2771649

ISO 24622-1:2015. (2015). Language resource management — Component Metadata Infrastructure (CMDI) — Part 1: The Component Metadata Model. Standard, International Organization for Standardization, Geneva, Switzerland. https://www.iso.org/standard/37336.html

Kopf, Maria, Schulder, Marc and Hanke, Thomas (2021). Overview of Datasets for the Sign Languages of Europe. EASIER Project Deliverable D6.1. https://doi.org/10.25592/uhhfdm.9561

3.3.6 Module B6: Language Community

Digital language archives materials in endangered or marginalised languages that are of value to speakers or their descendants. Facilitating access and (re)usability to these materials for the original language community members is of great value for sustenance and/or revitalization of the language. Beyond that, recordings can have great cultural and personal importance to the group they originated in. Hence, it is important to make the use of linguistic archival materials by communities as available and simple as possible.

Data Recommendations

Data Maturity Level 0

Data on this level is not relevant for this re-use scenario.

Data Maturity Level I

Data on this level is not relevant for this re-use scenario.

Data Maturity Level II

(B6.1) Data recommendations for reuse by language communities (Level II)

Description

With regard to supporting use and revitalisation of marginalised languages, recordings should be made accessible and attractive to members of the speakers' community. In order to promote the literate form of the language, the data has to be supplemented with **transcription** (if applicable, in standard or common orthography). In addition, a **translation of the materials to a commonly spoken major regional language** should assist non-native-speakers in better comprehending the language.

Possible Approach to Implementation

For audiovisual materials, it is recommended to reserve distinguishable, properly named tiers for transcription and translation using common tools. Textual content would be supplemented by the corresponding data within the same document (for general information on formats and tools, see module A.1).

Example

An ELAN file supplementing a video recording in Mehri, an endangered language in Oman, should contain a tier with transliteration of Mehri in Arabic script and a translation tier in Standard or Omani Arabic.

How to measure it?

Availability check for transcription and translation tiers of audiovisual annotation data can be automated to some extent if they are properly named. Data quality in general requires manual evaluation.

What is a valid result?

For acceptable state, the materials contain a translation of the studied language into a major language spoken in the community and a transcription; audiovisual materials should use tiers.

Benefits

This is not about FAIRness, but serves R1.

Metadata Recommendations

Data Maturity Level 0

Data on this level is not relevant for this reuse scenario.

Data Maturity Level I

Data on this level is also dependent on general WP 2.4 criteria (see Module B.8: 'Dissemination of Oral History (meta)data via Europeana and libraries', level I).

(B6.2) Dataset metadata recommendations for by reuse by language communities (Level I)

Description

In order to make **dataset level metadata** understandable to the language community, it should be supplemented with translation to **the community's language** and/or a **major regional language** known to the community. In other regards, requirements to the translated metadata match (A2.1) *Basic metadata recommendations of audiovisual language data.*

Benefits

This is not about FAIRness, but serves R1.

(B6.3) Individual recording metadata recommendations for by reuse by language communities (Level II)

Description

In order to make **individual recordings metadata** understandable to the language community, it should be supplemented with translation to **the community's language** and/or a **major regional language** known to the community. In other regards, requirements to the translated metadata match (A2.3) Basic structured metadata on resource part level including information on the recoding situation and participants.

Benefits

This is not about FAIRness, but serves R1.

3.3.7 Module B7: Ethnography

Besides their value for Linguistics, recordings of endangered languages often contain materials that are of significance for Ethnography (as well as, to a certain extent, Anthropology, Folklore Studies and other related disciplines in the Humanities). In order to facilitate further use of this data in ethnographic and related research, recordings must be made available, discoverable and re-usable. This requires matching a few additional metadata criteria.

Metadata Recommendations

Data Maturity Level 0

Data on this level is not relevant for this re-use scenario.

Data Maturity Level I

Data on this level is dependent on general level I WP 2.4 criteria (see Module B8: 'Dissemination of Oral History (meta)data via Europeana and libraries').

(B7.1) Individual recording metadata recommendations for reuse in Ethnography (Level II)

Description

Reuse of linguistic materials by Ethnographers requires an appropriate set of metadata to enable the discoverability of content. This requires using a standard commonly used Linked Data metadata terminology to describe the content of individual recordings.

Possible Approach to Implementation

The description of content should be accomplished using metadata in an RDF-based format. It is recommended to use the standard terminology offered by the American Folklore Society Ethnographic Thesaurus (AFSET). **Relevant concepts and their subterms** include but not limited to:

- performance
- narratives
- verbal art and literature
- legend
- poetry
- ritual
- music
- belief
- food
- migration and settlement
- work
- occupations
- manufacturing processes
- objects
- social dynamics
- space and place
- time

It is recommended to list five terms or more per recording. It is preferable to use narrower terms if such are available and applicable.

Example

Recording containing narration of a folktale about a rabbit outwitting a fox would have the following metadata:

- info:lc/vocabulary/ethnographicTerms/afset007214 (folktale)
- info:lc/vocabulary/ethnographicTerms/afset000584 (animal tales)
- info:lc/vocabulary/ethnographicTerms/afset020704 (trickster tales)
- info:lc/vocabulary/ethnographicTerms/afset014823 (rabbits)
- info:lc/vocabulary/ethnographicTerms/afset007388 (foxes)

How to measure it?

It is possible to automatise the check for metadata availability and format validation. Content quality requires manual evaluation.

What is a valid result?
For acceptable state, there should be metadata in an RDF-based format containing AFSET or related topics.

Benefits
This is not about FAIRness, but serves F2, I1, R1 and R1.3.

Related Resources
https://id.loc.gov/vocabulary/ethnographicTerms.html

3.3.8 Module B8: Dissemination of Oral History (meta)data via Europeana and libraries

Linguistic research data in the form of annotated audiovisual language resources containing biographical interview data are often valuable from a non-linguistic perspective. To ensure these resources are discovered by relevant disciplines beyond linguistics, more generic metadata and distribution channels can be used. This reuse scenario has been operationalized as the creation of additional metadata to make such resources findable in the Europeana portal and in library catalogues, including the WorldCat. For Germany, the Deutsche Digitale Bibliothek (DDB) is the national aggregator for Europeana, which means that Europeana metadata (EDM) provided to the DDB will be forwarded and integrated into Europeana, but also displayed in the DDB collections. Apart from simple format conversion, the transformation from common metadata formats for language resources into RDF-based EDM requires an enrichment of the (meta)data using standardised ontologies and vocabularies. The transcripts can be enriched with named entity recognition (NER) and additionally entity linking information, which is also applied to metadata elements. Subject indexing is performed using standards appropriate for the resource content.

Data Maturity Level 0
Data on this level is not relevant for this reuse scenario.

(B8.1) Relevant (meta)data can be made publicly available (Level I)

Description

For this third mission reuse scenario, (meta)data needs to be made publicly available, i.e. it must be clear which parts of the (meta)data can be made public. It is possible to only make metadata publicly available and provide links (persistent identifiers) to protected resources in archives and research data centres.

Possible Approach to Implementation

Informed consent that allows the public redistribution of the data has been collected from all participants and an open licence was chosen for the resource.

Example

The resource is available under the CC0 licence.

How to measure it?

An open licence can be automatically detected in structured metadata, if there is no clear licence covering all data allowing for public distribution, the metadata has to be manually assessed. If only metadata can be made publicly available, the resource and its parts have to be available via persistent identifiers.

What is a valid result?

All files to be used for this purpose have to come with an open licence or other explicit information allowing public distribution or they need to be available directly via persistent identifiers if only the metadata will be made publicly available.

Benefits

This is not about FAIRness, though the use of PIDs is required by F1.

Related Resources

Creative Commons Licenses: https://creativecommons.org/licenses/

(B8.2) Data covers the desired content and its authenticity and integrity is guaranteed (Level I)

Description

For data originally elicited for the purpose of linguistic research to be useful as a source in historical research, the content must not only match the current use case, but data also has to be reliable in the sense of authenticity and integrity. The status of (contemporary) witnesses has to be certain and provenance of the data clear.

Possible Approach to Implementation

The resource documentation contains detailed information on the verified identities of participants and on the provenance of data.

Example

The corpus guidelines distributed with the resource contains a project description including information on how subjects were recruited and data collected.

How to measure it?

A manual assessment of content suitability and data authenticity and integrity is necessary.

What is a valid result?

There is enough information to assure that the content is suitable and the data authenticity and integrity is sufficient.

Benefits

This is not about FAIRness, but R1 and R1.2 require rich metadata and provenance metadata, respectively.

(B8.3) DDB/Europeana metadata requirements are met (Level I)

Description

DDB/Europeana (EDM) metadata requires certain basic information in certain formats, in particular, as a Linked Data format (RDF) it requires information as resources, i.e. links (URIs) and not plain text values. Metadata requires a CC0 licence for distribution via Europeana.

Possible Approach to Implementation

If the original metadata format does not allow for the use of URIs, the relevant information can be provided within existing metadata using RDFa. Otherwise, the metadata must be enriched with links during customised format conversion from the original metadata format into EDM, or enriched after it has been converted into EDM.

Example

The CMDI metadata was enriched with RDFa for the relevant information and converted into EDM.

How to measure it?

The set of required metadata information for the Europeana EDM format is:

For edm:ProvidedCHO
dc:language (if edm:type = TEXT)
edm:type
dc:description | dc:title (one of these required)
dc:subject | dc:type | dcterms:spatial | dcterms:temporal (one of these required)

For ore:Aggregation
edm:aggregatedCHO
edm:dataProvider
edm:isShownAt | edm:isShownBy (one of these required)
edm:provider
edm:rights

For cc:License
odrl:inheritFrom

The elements required for the DDB are

- Data set identifier (mandatory)
  the identifier is provided in the attribute @rdf:about of the element <ore:Aggregation>
- Data provider identifier (mandatory)
  the ISIL of the providing institution is used in the element <edm:dataProvider> in the element <ore:Aggregation>
- Preview image (mandatory)
  a permanent link (URL) to the image (JPEG/PNG/TIFF, min. 800x600 px) to be displayed in search results
- Link to digital object (mandatory)
  the permanent link can be provided directly to the media file (in the attribute @rdf:resource of the

element <edm:isShownBy> or in @rdf:about of <edm:WebResource> as the child element of <edm:isShownBy>), to a designated viewer for the media file (in <ore:Aggregation> in <edm:isShownAt>), or to the digital object in its original context, i.e. a landing page (in @rdf:resource of <edm:isShownAt> or in @rdf:about of <edm:WebResource> as the child of <edm:isShownAt>)

- Legal status of the digital object (mandatory)
  the URI of the licence or rights information is provided in the element <edm:WebResource> of the element <edm:rights>
- Object title (mandatory)
  the title is provided in the element <edm:ProvidedCHO>, in the element <dc:title>
- Object type (mandatory)
  the object type must be from a controlled vocabulary, e.g. the Objects Facet of the Art & Architecture Thesaurus (AAT) or the Gemeinsame Normdatei (GND), a corresponding reference to Wikidata can be provided in addition to these. A German preferred label is required.
- Legal status of the metadata
  the URI for the licence or rights information of the metadata can be provided in <ore:Aggregation>, in the attribute @rdf:resource of the element <dcterms:rights>
- Funding
  information on funding can be provided in <edm:WebResource> using the element <marcrel:fnd>

The converted EDM can be validated including Schematron rules.

What is a valid result?

The converted EDM validates including Schematron rules.

Benefits

The EDM format is based on RDF and thus complies with the Interoperability principles in addition to the Reusability principles. The dissemination via Europeana increases Findability.

Related Resources

Europeana Data Model: https://pro.europeana.eu/page/edm-documentation

DDB requirements (German): https://wiki.deutsche-digitale-bibliothek.de/display/DFD/Anforderungen+an+die+Lieferdaten

ISIL: https://sigel.staatsbibliothek-berlin.de/de/startseite/

AAT:http://www.getty.edu/research/tools/vocabularies/aat/

GND:https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd.html

Wikidata: https://www.wikidata.org/wiki/Wikidata:Main_Page

(B8.4) Library catalogue metadata requirements are met (Level I)

Description

For integration of collections or corpora - not individual sessions or recordings - into library catalogues, only basic catalogue metadata, such as the main contributors and the title of the resource is required. Still, translating a great number of contributors with different roles into concise information on authors or editors is not trivial. Agents such as contributors or institutions also require identifiers from the GND. Additional subject indexing using controlled vocabularies such as the GND is highly recommended.

Possible Approach to Implementation

For the integration, metadata can be converted into the format used by the library that will be responsible for cataloguing, however as this process often includes complex adaptations, manual processing might be necessary and more efficient. As libraries favour different formats, a specific requirement regarding the format is not possible.

Example

For the integration into the library catalogue of the university library, basic information was provided as required by the cataloguing librarian.

How to measure it?

Manual assessment for completeness of basic catalogue metadata, in particular contributors' roles.

What is a valid result?

Complete basic catalogue metadata with clear contributor roles for the authors/editors to show in the record.

Benefits

Listing resources in library catalogues increases Findability.

Related Resources

GND: https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd.html

(B8.5) Transcripts include NE/EL annotation (Level II)

Description

Semantic enrichment of the data can be used as a basis for semantically enriched metadata but also for visualisation and analysis. Named entities such as persons, places, organisations etc. can be marked up as such and additionally linked to concepts in ontologies or controlled vocabularies. For specific use cases, specific ontologies can be used, however Wikidata is now established as a de-facto standard and many specific ontologies are being mapped or integrated.

Possible Approach to Implementation

Named entities can be annotated and linked manually or by using automatic methods, with these alternatives implying the familiar benefits and disadvantages in terms of speed, costs and precision.

Example

Named entities were annotated and linked manually using INCEpTION and Wikidata.

How to measure it?

Manual check for an annotation layer including NE and EL.

What is a valid result?

There is an annotation layer (or several layers) containing NE and EL annotation with appropriate references for the entity linking.

Benefits

The entity linking increases Interoperability.

Related Resources

Wikidata: https://www.wikidata.org/wiki/Wikidata:Main_Page

4. Implementation

Following the description of the quality criteria and evaluating if they can be checked automatically we can sketch the implementation of these automatic validation methods as well as potential workflows and technical requirements for deployment and maintenance. As stated before, the automatic checking procedures were implemented as part of the Corpus Services, initially developed at the HZSK.

4.1 Workflows

In the Objective section we sketch several general use cases, each using its own kind of workflow. These workflows also have to be reflected in automatic checking workflows. We developed three example workflows which have different characteristics and intended audiences. All workflows allow for continuous quality checking (Use case I) while the web application specifically aims at certification (Use case II and III).

4.2 Web application

The initial workflow was designed around a web application. The researcher is presented with a web frontend providing user-friendly access to the automatic checking procedures. Instead of having to decide themselves which checks are necessary for their data, the users are guided by answering a questionnaire. The answers to this questionnaire decide as a first step if data processing is possible and under which circumstances. As a second step the checkers are chosen depending on the kind of data and potentially discipline-specific requirements. Finally, the user can provide their data by uploading it and run the preselected checkers. The resulting report can be used to decide about a potential certification or guide improvement of the corpus data.

Because a web application needs to be hosted, this workflow works best in collaboration with an archive, sharing the work between the researcher and the archive. The archive hosts the web interface and the infrastructure for running the checks, the user provides all necessary information to evaluate the data quality. Only in case of an intended certification, the archive acts on the results on the checking procedures.

Implementation:

The frontend is a combination of client-side JavaScript using JQuery and the Bootstrap library and server-side Python code. In the backend a Java application provides an HTTP-API to the corpus services. For performance reasons the use of a reverse proxy such as Apache or Nginx is highly recommended.

4.2.1 Stand-alone application

A challenge with the web application is the requirement to upload all corpus data onto a server. Especially for audiovisual corpora containing large amounts of audio and video data this can be problematic. An alternative to the purely web-based application is a stand-alone application that can be downloaded and run locally making it unnecessary to upload the corpus data. Such an application can contain the same features as the web application or a reduced feature set to just run a set of preselected checks, e.g., selected using the web application. As a self-contained application it contains both the user interface and the whole set of checkers. The use of a stand-alone application is useful for continuous quality assurance but is not useful for a certification scenario.

Implementation:

A prototype of such a stand-alone has been developed. The user simply downloads a precompiled Java binary only requiring a Java installation.

4.3 Git workflow

The final workflow we want to describe is inspired by continuous integration methods used in software engineering. Using a version control system such as Git, automatic checks or processes can be executed whenever the stored information is updated. In the case of software engineering it could be checked if the code compiles and all test cases succeed. The combination of version control systems and automatic checking methods for research data have been deployed previously, e.g., in the Conquaire[30] project. Similarly, we can use this method to run our automatic checking methods to provide continuous quality checking when developing new corpora. The configuration can be created e.g., using the web application or manually.

Implementation:

The simplest way to implement this workflow is by integrating it in an existing Gitlab instance which already provides all means necessary. With a little extra effort it is also possible to use the workflow with any self-hosted Git repository using git hooks.

4.4 Development, Deployment, and Maintenance

The Corpus Services which form the foundation for the QUEST quality assurance methods have been developed since at least 2017 and there have been contributions by 12 people. Since June 2021, all the additional QUEST checkers have been added by one developer with less than 100% FTE. Adding additional checkers took, depending on the validation task, between one day and several weeks. Additional time was necessary for testing and documentation.

The effort for deployment depends on the intended workflow, ranging from basically no effort using the stand-alone application or including the checks in a Gitlab CI workflow to moderate effort for deploying the web application.

To guarantee for long-term use of the checking procedure some maintenance will be required, specifically when the Java version changes, dependencies become deprecated, and to handle security issues either in the Java version or in one of the dependencies. The maintenance is low effort on average but requires staff available with the necessary skills.

4.5 Automatic Checks

A wide range of automatic checkers have been developed in the context of quest, covering most of the work packages. The generic criteria in WP 1.1 and WP 1.2 are more thoroughly covered while it was more difficult to implement automatic quality assurance procedures for the more discipline-specific usage scenarios. A comprehensive documentation of the automatic QUEST quality assurance methods is available in the Corpus Services wiki.[31]

[30] https://www.uni-bielefeld.de/ub/digital/forschungsdaten/kompetenzzentrum/projekte/conquaire/

[31] https://gitlab.rrz.uni-hamburg.de/corpus-services/corpus-services/-/wikis/QUEST

5 Related Documents

Arestau, Elena (2022): Curation of Learner Corpora. https://www.slm.uni-hamburg.de/ifuu/forschung/forschungsprojekte/quest/ueber-das-projekt/projektergebnisse/arestau-learnercorpora.pdf.

Arestau, Elena (2022): Curation of Interpreted Corpora Using The Example of ComInDat. https://www.slm.uni-hamburg.de/ifuu/forschung/forschungsprojekte/quest/ueber-das-projekt/projektergebnisse/arestaucomindat.pdf

Arkhangelskiy, Timofey, Hedeland, Hanna & Riaposov, Aleksandr (2021). Evaluating and assuring research data quality for audiovisual annotated language data: In: Navarreta, Constanza & Eskevich, Maria (Hg.): *Selected Papers from the CLARIN Annual Conference 2020*: Virtual Event, 2020, 5-7 October. (Linköping Electronic Conference Proceedings 180). Linköping: Linköping University Electronic Press. S. 1-7.

Aznar, Jocelyn and Lange, Herbert (2022) RefCo and its Checker: Improving Language Documentation Corpora's Reusability Through a Semi-Automatic Review Proces. In: Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022). 20-25 June 2022, Marseille. S. 2721–2729.

Aznar, Jocelyn and Seifart, Frank (2022) The RefCo Toolkit. https://zenodo.org/record/7380448.

Hedeland, Hanna (2021) Towards comprehensive definitions of data quality for audiovisual annotated language resources. In: Navarreta, Constanza & Eskevich, Maria (Hg.): *Selected Papers from the CLARIN Annual Conference 2020*: Virtual Event, 2020, 5-7 October. (Linköping Electronic Conference Proceddings 180). Linköping: Linköping University Electronic Press. S. 93-103.

Hedeland, Hanna (2022) FAIR-Prinzipien und Qualitätskriterien für Transkriptionsdaten: Empfehlungen und offene Fragen. In: Schwarze, C. & Grawunder, S. (Hrsg.) *Transkription und Annotation gesprochener Sprache und multimodaler Interaktion. Konzepte, Probleme, Lösungen*. Narr Francke Attempto.

Isard, Amy (2022) Approaches to the Anonymisation of Sign Language Corpora. 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources. 25 June 2022. Marseille.

Isard, Amy & Arestau, Elena (2022) Curation Criteria for Multimodal and Multilingual Data: A Mixed Study within the QUEST project. In: Navarreta, Constanza & Eskevich, Maria (Hg.): *Selected Papers from the CLARIN Annual Conference 2021*: Virtual Event, 2021, 27 - 29 September. (Linköping Electronic Conference Proceedings 189). Linköping: Linköping University Electronic Press. S. 56-68.

Rau, Felix, Majka, Nicole & Schwiertz, Gabriele (2022): Metadata Recommendations for Audio-Visual Language Data. DOI: 10.5281/zenodo.7346840.

References

Angermeyer, Philipp S, Meyer B and Schmidt, T (2012) “Sharing community interpreting corpora. A pilot study”. In: Multilingual Corpora and Multilingual Corpus Analysis. Vol. 14. Amsterdam/Philadelphia: John Benjamins, pp. 275–94.

Bell, P. and Payant, C. (2020) Designing Learner Corpora. Collection, Transcription, and Annotation. In: Paquot, M. and Tracy-Ventura, N. (eds.): The Routledge Handbook of Second Language Acquisition and Corpora. Routledge.

Bernardini, S., Ferraresi A., Lefer M. A., and Miličević, M. (2016a) Simplification in translation and interpreting: Using a tri-directional intermodal corpus to shed light on commonalities and differences. Paper presented at Translation and Interpreting. Convergence, Contact, Interaction (TransInt). Trieste, Italy 26–28 May 2016.

Bleicken, Julian, Hanke, Thomas, Salden, Uta and Wagner, Sven (2016) “Using a Language Technology Infrastructure for German in Order to Anonymize German Sign Language Corpus Data”. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Portoroˇz, Slovenia, pp. 3303–3306.

Crasborn, O. (2010) What Does “Informed Consent” Mean in the Internet Age? Publishing Sign Language Corpora as Open Content. Sign Language Studies, 10(2):276–290.

Deutsche Forschungsgemeinschaft. (2020) Digitaler Wandel in den Wissenschaften. Zenodo. https://doi.org/10.5281/zenodo.4191345

Diemer, Stefan, Brunner M. and Schmidt, S. (2016) “Compiling Computer-Mediated Spoken Language Corpora: Key Issues and Recommendations”. In: International Journal of Corpus Linguistics 21.3, pp. 348–371.

Gilquin, G. (2015) From design to collection of learner corpora. In: Granger S., Gilquin G., & Meunier F. (eds.) The Cambridge Handbook of Learner Corpus Research (Cambridge Handbooks in Language and Linguistics). Cambridge: Cambridge University Press, pp. 9–34.

Granger, Sylviane, Gilquin, Gaetanelle, Meunier, Fanny (2016): The Cambridge handbook of learner corpus research. Cambridge: University Press

Granger, S. & Paqout, M. (2017) Towards standardization of metadata for L2 corpora. CLARIN workshop on Interoperability of Second Language Resources and Tools (Gothenburg, Sweden, du 06/12/2017 au 08/12/2017).

Hedeland, Hanna (2021) Towards comprehensive definitions of data quality for audiovisual annotated language resources. In: Navaretta, Constanza and Eskevich, Maria (Hg.): Selected Papers from the CLARIN Annual Conference 2020: Virtual Event, 2020, 5-7 October. (Linköping Electronic Conference Proceedings 180). Linköping: Linköping University Electronic Press, 93.103.

Hedeland, H. (2022) FAIR-Prinzipien und Qualitätskriterien für Transkriptionsdaten: Empfehlungen und offenen Fragen. In: Schwarze, C. & Grawunder, S (Hrsg.) Transkription und Annotation gesprochener Sprache und multimodaler Interaktion. Konzepte, Probleme, Lösungen. Narr Francke Attempto.

RfII – Rat für Informationsstrukturen (2019) Herausforderung Datenqualität – Empfehlungen zur Zukunftsfähigkeit von Forschung im digitalen Wandel, zweite Auflage.

Kisler, T., Reichel, U., and Schiel, F. (2017) “Multilingual Processing of Speech via Web Services”. In: Computer Speech & Language 45.C, pp. 326 – 347. https://www.sciencedirect.com/science/article/abs/pii/S0885230816302418

König, Alexander; Frey, Jennifer-Carmen; Stemle & Egon W. (2021) "Exploring Reusability and Reproducibility for a Research Infrastructure for L1 and L2 Learner Corpora" Information 12, no. 5: 199. HTTPS://DOI.ORG/10.3390 /info12050199

Le Bruyn, B. & Paquot, M. (2021) Learner Corpus Research Meets Second Language Acquisition. Cambridge: University Press.

LIPPS Group, Barnett, R., Codó, E., Eppler, E., Forcadell, M., Gardner-Chloros, P., van Hout, R., Moyer, M., Torras, M. C., Turell, M. T., Sebba, M., Starren, M. & Wensing, S. (2000) The LIDES Coding Manual. A document for preparing and analyzing language interaction data, Version 1.1– July, 1999. International Journal of Bilingualism 4, pp. 131–270 .

McEnery, T. and Hardie, A. (2011) Corpus Linguistics: Method, Theory and Practice. Cambridge Textbooks in Linguistics. Cambridge University Press.

Megyesi, B., Granstedt, L., Johansson, S., Prentice, J., Rosén, D., Schenström, C-J., Sundberg, G., Wirén, M., Volodina, E. (2018) Learner Corpus Anonymization in the Age of GDPR: Insights from the Creation of a Learner Corpus of Swedish. In Proceedings of the 7th NLP4CALL, SLTC workshop, Stockholm, Sweden.

Meyer, Bernd (2012) “Ad hoc Interpreting for Partially Language-proficient Patients: Participation in Multilingual Constellations”. In: Coordinating Participation in Dialogue Interpreting, Claudio Baraldi and Laura Gavioli (eds), Amsterdam/Philadelphia, John Benjamins, pp. 99–113.

Paquot, M. and Tracy-Ventura, N. (2020) The Routledge Handbook of Second Language Acquisition and Corpora.

Russo, Mariachiara, Bendazzoli, C., Sandrelli. A., and Spinolo, N. (2012) The European Parliament Interpreting Corpus (EPIC): Implementation and developments. In Breaking ground in corpus-based interpreting studies, ed. F. Straniero Sergio, and C. Falbo, 35–90. Bern: Peter Lang.

Rock, F. (2001) Policy and Practice in the Anonymisation of Linguistic Data. International Journal of Corpus Linguistics, 6(1), pp. 1–26.

Schembri, A., Fenlon, J., Rentelis, R., Reynolds, S., and Cormier, K. (2013) Building the British Sign Language Corpus. Language Documentation & Conservation, 7, pp. 136–154.

Schiel, Florian and Kisler, Thomas (2019) “BAS Web Services for Automatic Subtitle Creation and Anonymization”. In: Proceedings of Interspeech 2019. Graz, pp. 3671–3672.

Singleton, J. L., Jones, G., and Hanumantha, S. (2014) Toward Ethical Research Practice With Deaf Participants. Journal of Empirical Research on Human Research Ethics, 9(3), pp. 59–66.

Stemle, E., Boyd, A., Janssen, M., Lindström Tiedemann, T., Mikelić Preradović, N., Rosen, A., Rosén, D. & Volodina, E. (2019) Working together towards an ideal infrastructure for language learner corpora. In: Andrea Abel, Aivars Glaznieks, Verena Lyding & Lionel Nicolas (eds.) Widening the Scope of Learner Corpus Research. Selected papers from the fourth Learner Corpus Research Conference. Corpora and Language in Use – Proceedings 5, Louvain: Presses universitaires de Louvain, pp. 427-468.

Volodina, E., Megyesi, B., Wirén, M., Granstedt, L., Prentice, J., Reichenberg, M., & Sundberg, G. (2016) A friend in need? Research agenda for electronic second language infrastructure. In: Proceedings of the Sixth Swedish Language Technology Conference (SLTC-2016), Umeå, 17-18 November, 2016, pp. 1–4.

Xia, Zhaoyang, Chen, Yuxia, Zhangli, Qilong, Huenerfauth, Matt, Neidle, Carol and Metaxas, Dimitris (2022) “Sign Language Video Anonymization”. In: Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources, pp. 201-211. Marseille, France. https://www.sign-lang.uni-hamburg.de/lrec/pub/22038.html

Zanettin F. (2012) Translation-Driven Corpora: Corpus Resources for Descriptive and Applied Translation Studies, Manchester, St. Jerome Publishing.