UDKAST:
Anbefalede standarder for sprogressourcer

1. november 2021

DIGITALISERINGSSTYRELSEN

Baggrund - User stories

1

Som dataudgiver vil jeg vide, hvilket format jeg skal anvende således, at mine sprogdata kan omdannnes til eller indgå i et større korpus (fx Gigaword).

2

Som dataanvender/udvikler kan jeg mere effektivt anvende data, hvis formaterne er åbne, simple, og veldokumenterede - og at samme format eller sammenlignelige formater anvendes.

3

Som redaktør for sprogteknologi.dk vil jeg gerne kunne oplyse potentielle og eksisterende dataudgivere om hvilke formater, der er mest hensigtsmæssige for dataanvenderne/udviklerne

4

Som dataindsamler vil jeg gerne være sikker på, at de data, jeg indsamler, bliver egnede til anvendelse til sprogteknologiske formål.

5

Som dataanvender/udvikler vil jeg undgå, at tekniske barrierer eller manglende kvalitet gør, at jeg ikke kan anvende data.

6

Som redaktør for sprogteknologi.dk vil jeg sikre, at de data, vi udstiller, har formater og kvalitet, der gør, at de kan anvendes af så mange som muligt.

DIGITALISERINGSSTYRELSEN

Generelle anbefalinger

Til sprogdata anbefales generelt

- Åbne formater frem for proprietære formater
- Tekstnære formater (plain text) frem for formater med binære data

Til intern strukturering af sprogdata anbefales generelt

- Simple mappestrukturer
- Meningsbærende filnavne der er uafhængige af mappestrukturer

Til intern dokumentation / metadata

- Angivelse af id, udgivelse, sted, indsamlet/bygget for hvert dokument
  Format=JSON: https://www.json.org/json-en.html
- Angivelse af sprog:
  - IETF BCP 47: https://tools.ietf.org/search/bcp47
  - ISO 639: https://www.iso.org/iso-639-language-codes.html
- METASHARE-standard til udvidet dokumentation: http://www.meta-net.eu/meta-share/metadata-schema
- Data Statements for NLP til at synliggøre bias: https://aclanthology.org/Q18-1041/

Anbefalinger: Lyd / tale

Anbefalede formater:

1. .wav [uncompressed, non-lossy]:
   http://www-mmsp.ece.mcgill.ca/Documents/AudioFormats/WAVE/WAVE.html
2. .flac [compressed, non-lossy]
   https://xiph.org/flac/documentation.html

Lydoptagelser bør altid være i .wav (evt. gemmes i .flac).

Anbefalede sample-rates:

Uanset format bør sample-rate altid vælges fra denne liste:
8kHz / 16kHz / 20k / 22.050kHz / 44.1kHz / 48kHz;
og bit-rates = 16 / 24

Organisering af lydfiler:

Lydfiler og tilhørende transskriptioner bør være opdelt på samme måde således, at indholdet af én lydfil svarer til indholdet af én transskriptionsfil, og disse to filer bør nemt kunne relateres, fx via navngivningen.
Opdel gerne lyd- og transskriptionsfiler på sætningsniveau.

Anbefalinger: Video

Anbefaling vedrørende videooptagelser

- MJPEG-2000 lossless som backendformat
  (ISO/IEC 15444-3:2002 Information technology — JPEG 2000 image coding system
  — Part 3: Motion JPEG 2000:
  https://www.iso.org/standard/33875.html)
- MPEG-2 eller H.264 (typisk inkluderet i MPEG4) til behandling af videoindhold
  (MPEG-2-standarder udgives som dele af ISO/IEC 13818-1:2019
  (Information technology — Generic coding of moving pictures and associated audio information):
  https://www.iso.org/standard/75928.html
  https://www.itu.int/rec/T-REC-H.264

Anbefalinger: Tekst

Anbefalede tekstformater:

- .txt (til simple ikke-formaterede tekstdokumenter)
- .csv, XML, JSON, JSONL (til måledata, annotationsdata m.m. fx vektordata, statistiske data, tidskoder, etc.)*
- *Andre formater til strukturerede tekstdokumenter?*

Kodning

Foretrukken:

- UTF8 (Unicode Transformation Format 8-bit): https://www.unicode.org/

Andre:

- UTF16 (Unicode Transformation Format 16-bit); https://www.unicode.org/
- ISO-8859-1 (Information technology — 8-bit single-byte coded graphic character sets — Part 1: Latin alphabet No. 1): https://www.iso.org/standard/28245.html

*Ikke fastlagt anbefaling endnu

Anbefalinger: Leksikalske ressourcer

Anbefalinger vedr. maskinlæsbare leksika eller termbaser

- Lexical Markup Framework (LMF), ISO 24613:2008: https://www.iso.org/standard/37327.html
- TermBaseExchange format (TBX), ISO 30042:2019: https://www.iso.org/standard/62510.html

Anbefalinger vedr. Knowledge Engineering

- Resource Description Framework (RDF): https://www.w3.org/TR/rdf11-primer/
- RDF Schema 1.1 (RDF-S): https://www.w3.org/TR/rdf-schema/
- Ontology Web Language (OWL): https://www.w3.org/TR/owl2-overview/
- Simple Knowledge Organization System (SKOS): https://www.w3.org/TR/2009/REC-skos-reference-20090818/

Øvrige identificerede annoteringsmuligheder?

Korpusannotering

.tei -Text Encoding Initiative (TEI) (til annotation, opmærkning og repræsentation af korpora): https://tei-c.org/release/doc/tei-p5-doc/en/html/

Syntaktisk annotering

- ISO/DIS 24611 Morpho-syntactic Annotation Framework (MAF): https://www.iso.org/standard/51934.html
- ISO/CD 24615:2010 Syntactic Annotation Framework (SynAF): https://www.iso.org/standard/37329.html

Semantisk annotering

- Language resource management — Semantic annotation framework (SemAF) — Part 1: Time and events (SemAF-Time, ISO-TimeML): https://www.iso.org/standard/37331.html

Dialogannotering

- *ISO 24617-2 Language resource management — Semantic annotation framework (SemAF) — Part 2: Dialogue acts: https://www.iso.org/standard/76443.html*

Multimodal annotering (Gestus i multimodal kommunikation)

- The MUMIN annotation framework: https://www.cst.dk/mumin/resources/MUMIN-coding-scheme-V3.3.doc

Følelsesannotering (herunder sentimentannotering)

- Emotion Markup Language (EmotionML): https://www.w3.org/TR/emotionml/

Bemærk: Forskellige CLARIN K-centre anbefaler også standarder for deres respektive ekspertiseområder. Du finder relevant information på denne CLARIN-webside.