Dokumentasjon og metadata innanfor språkvitskap

RDA-kurs i FAIR forskingsdatakuratering
UiT, 20.-21. januar 2020

Philipp Conzett, UiT

Fleire relevante standardar

- Ulike typar av språkdata/språkressursar
- Ulike teoretiske og metodiske tilnærmingar til beskriving og analyse av data

- *De jure*- vs. *de facto*-standardar
  - De jure: Standard som er utvikla eller i alle fall godkjend av ein organisasjon som har som oppgåve å utarbeida standardar.
  - De facto: Retningsline eller tilråding som er teken i bruk av (store) delar av brukarar innanfor det aktuelle området, og som derfor kan seiast å ha utvikla seg til ein standard.
- Nokre døme:

ISO-639-3

- Kanskje den mest kjende de jure-standarden (innanfor språkvitskap)
- Sett med forkortingar/**kodar for namn på språk**
- I stor grad basert på språkodar i *Ethnologue*, ein årleg oversikt over verdsspråk, utgjeven sidan 1951 av SIL International, ein kristen non-profit-organisasjon (døme på utvikling frå de facto- til de jure-standard)
- Også brukt på andre fagområde, men særleg viktig for språkdata
- Døme: språkkodar for offisielle språk i Noreg:

Finnish, Kven [fkv]
Norwegian [nor]
Norwegian Sign Language [nsl]
Norwegian, Traveller [rmg]
Romani, Vlax [rmy]
Saami, Lule [smj]
Saami, North [sme]
Saami, Pite [sje]
Saami, South [sma]
Saami, Ume [sju]

Open Language Archive Metadata (OLAC Metadata)

- Utvikla av Open Archives Initiative (OAI)
- Basert på m.a. Dublin Core og OAI Protocol for Metadata Harvesting (OAI-PMH)
- Brukt på språkressursar og språkteknologiverktøy
- Særleg for hausting av metadata frå arkiv
- Men (delar) kan også brukast meir generelt, t.d.
- OLAC Role Vocabulary: http://www.language-archives.org/REC/role.html
- Sett med termar for ulike roller deltakarar i eit språkvitskapleg prosjekt kan ha, og definisjonar av desse rollene:

annotator, author, compiler, consultant, data_inputter, depositor, developer, editor, illustrator, interpreter, interviewer, participant, performer, photographer, recorder, researcher, research_participant, responder, **signer,** singer, **speaker**, sponsor, transcriber, translator

Utfordring: Fleire standardar for det same

- Misnøye med eksisterande standard fører ofte til utvikling av endå ein ny standard.
- Resultatet er/kan bli eit **mylder av standardar**.
- Treng **harmonisering og koordinering**
- Det finst forsøk på dette, t.d. i **CLARIN**

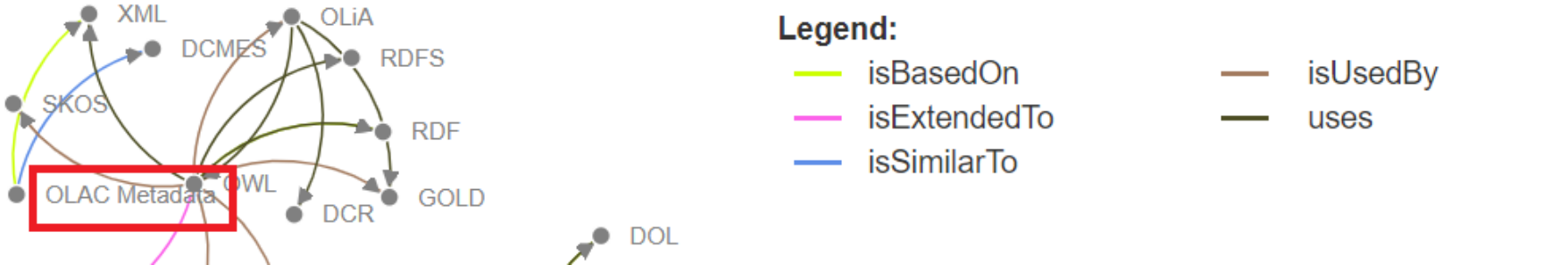
CLARIN

- **C**ommon **La**nguage **R**esources and Technology **In**frastructure = ERIC (European Research Infrastructure Consortium); jf. ELIXIR?
- Mål: “create and maintain an infrastructure to support the **sharing, use and sustainability of language data and tools** for research in the humanities and social sciences.” (https://www.clarin.eu/content/clarin-in-a-nutshell)
- **Retningsliner for føretrekte standardar** for språkdata og -
  - **Open** standards are preferred over proprietary standards
  - Formats and protocols should be:
    - **well-documented**
    - **verifiable**
    - **proven** (being used in practice)
  - ...

CLARIN - oversikt over standardar

- Nyttig oversikt over standardar og spesifikasjonar, og korleis dei er relaterte til kvarandre: https://clarin.ids-mannheim.de/standards/:

Abbreviation/Name	Topic(s)	Responsibility	CLARIN Center(s)
OLAC Metadata	Metadata	OLAC	CLARIN-PL, TLA, UC
OLiA	Ontology | Terminology		
OntoIOp	Knowledge Representation | Ontology	ISO	
OpenCyc	Ontology	Other	
OWL	Knowledge Representation | Ontology	W3C	UC



Component MetaData Infrastructure (CMDI)

- https://www.clarin.eu/content/component-metadata
- Ikkje ein metadatastandard, men eit **rammeverk** for å beskriva metadata**profilar** som inkluderer **element frå fleire metadatastandardar**
- Ein profil består av **komponentar**, t.d.:
  - Generelle metadata henta frå Dublin Core, t.d. dc.title, dc.creator, ...
  - Språkkode henta frå ISO-639-3
  - Contributor henta frå OLAC Role Vocabulary
  - …
- Profilar og komponentar er gjorde tilgjengelege i CLARIN sitt Component Registry (https://catalog.clarin.eu/ds/ComponentRegistry/#/)
- Fare: Fleire komponent og profilar som handlar om det same...

RDA Linguistics Data Interest Group

- CLARIN er mest aktiv/relevant i **Europa**.
- Vi treng også eit **globalt nettverk**/plattform for å ta opp spørsmål rundt handtering av språkdata.
- Dette er målet med RDA Linguistics Data Interest Group (https://www.rd-alliance.org/groups/linguistics-data-ig).
- Har først konsentrert oss om tilrådingar for **sitering** av språkdata. “The Tromsø recommendations for citation of research data in linguistics” kjem snart!
- **Metadata** er eit anna område som vil vera relevant for gruppa.

Neste post på program

Gruppearbeid

>> Sjå delt Google-mappe.