Provenance practices at the IISG / CLARIAH WP4

Richard Zijdeman
(HuC-IISG, University of Stirling)
September 3rd, 2018
State of the Art

• datasets (excel files and what not)
• Linked Data
• software
Depositing datasets - Status quo

Situation:
- local (institute) instance of Dataverse
- basic provenance questions + metadata (e.g. author, period, region)
- many other options
Depositing datasets - Results

We now have provenance! Yaj! But...

- heterogeneous input of provenance
- provenance not provided in particular vocabulary
Moving forward:
• Align questions with items from Collections (EAD / MARC)
• Monitor provenance output
• (Start to think about provenance of versions)
Linked Data (importance)

- reproducible cleaning of data (Linked Edit Rules)
- reproducible linking of datasets (SPARQL-query)
- getting data out there: 'RAW DATA NOW' (Sir Tim Berners-Lee)
Linked Data – Converting datasets

• conversion tool 'CoW' (CSV on the WEB (CSVW))
• auto-generates basic provenance (in dc-terms and prov-o)
• and allows for specific using Jinja templating
Linked Data – Querying datasets

• grlc (SPARQL queries as API)
• store (github) and share SPARQL queries over the WEB
• ... so storage, versioning but not preservation of queries
Linked Data – Current Issues

• "Is my:sameAs the same as your:sameAs?" (Idrissou et al. 2017)
• Querying: which vocabularies to query for? (e.g. dc-terms and prov-o)
• Updates?
  • original data (rows vs. columns)
  • updates in LD (augmentation)
Software

• git! (or other repo)
• docker (helps -current- platform dependency issues)
• is preservation of code not enough?
Lack of provenance is as much a technical issue as it is a social issue. In addition to algorithms we need social mechanisms to generate provenance.