Provenance in WP5’s Media Suite

CLARIAH Provenance workshop
DANS
September 4, 2018
The CLARIAH Media Suite: http://mediasuite.clariah.nl/

- A high-level tool (user-friendly research environment) built in a modular/sustainable approach
- Integrates data and tools for both novice users and advanced users (via Jupyter notebooks)
- We are not yet actively implementing provenance solutions based on “formal” models/solutions, but rather respecting the principles of provenance information (i.e., “source critique” in scholarly terms): transparency and traceability (location of the sources, and of changes made to the sources on the process of providing access).
Media Suite’s building blocks

1. Data sources
2. APIs
3. Components / Tools
4. Enrichment services
5. Web-based federated login

Challenges in respecting provenance and integrating provenance information

1. Data sources:
   - CKAN registration
   - Human-readable descriptions of collections/data and the content providers
   - Human-readable descriptions and conversion files explaining data processes
1. Data sources

Currently: “user-friendly” way of respecting provenance information:

- Respecting/offering original metadata (by content provider)
- Following the principles of data transparency: visualization of metadata completeness
- Providing metadata dictionaries
Challenges in respecting provenance and integrating provenance information

3. Components/Tools:

Currently: Making easier for users to trace metadata back

- Links to original sources
- Metadata presentation in a transparent way

Future work: encode provenance information formally (depending on CLARIAH common approach)
Challenges in respecting provenance and integrating provenance information

4. Enrichment services:
   • Currently: we provide live statistics/ASR work on a separate web site,
   • Future work: analyze how to integrate provenance information in the ASR metadata
Challenges in respecting provenance and integrating provenance information

5. Login/Authentication:
   • Working together with WP2 on an “authenticatie route”

6. Workspace
   • Currently: storing user annotations and saved queries
   • Future work: deciding on what data from the users should be stored for keeping provenance information in user annotations (and also for log analysis)