CLARIN
Common Language Resources and Technology Infrastructure

Oxford Text Archive

Jens Stegmann

Oxford Text Archive, IDS Mannheim, IMS Stuttgart

jens.stegmann@ims.uni-stuttgart.de

Workshop on Federated Search Infrastructures
Nijmegen, Netherlands
2011-05-09

Structure of the Talk

- Background on the Oxford Text Archive
- Where we are
- OAI-PMH and Resources
- Metadata Crosswalks
- Plans for the Future

Background on the OTA

- Head of the OTA: Martin Wynne
- Hosted by the Oxford University Computing Services (OUCS), close ties with the Oxford e-Research Centre
- Production Server: http://ota.oucs.ox.ac.uk/
- The OTA contains a heterogeneous collection of (mostly) textual data that differ w.r.t. to, e.g.:
  - Genres
  - Languages
  - Time of origin
  - Modality
  - Annotation
  - Encoding

CLARIN-related: Where we are

- Authentication and authorisation: *Shibboleth Service Provider* protects a test set of data on an Apache running on a dedicated virtual server provided by the OeRC

  https://ota.oerc.ox.ac.uk/secure
- Registering *persistent identifiers* for resources → realised via a client that talks to the RESTful PID web service hosted by GWDG for the EPIC consortium
- *OAI-PMH 2.0 compliant data provider* component based on the XMLFile Perl framework (makes heavy use of XSLT)

  http://ota.oerc.ox.ac.uk/oai2/XMLFile/ota/oai.pl
- *XSLT-based metadata crosswalks*: mapping the original TEI Headers to DC and OLAC in order to provide them via OAI-PMH; corrections and revisions, where necessary

OAI-PMH and the Resources

- OAI-PMH: Perl-based implementation configured and adapted to our needs. It makes use of XSLT in order to generate the respective formats dynamically (TEI → oai_dc, TEI → olac)
- Two *sets of* metadata resources ready to be harvested
- Set “newheaders”
  - 17 resources
  - Carefully crafted corpora of special interest to linguists
  - Only the metadata are TEI-compliant = TEI Headers
- Set “nuOTA”
  - 297 resources
  - Works which are in the public domain; humanities
  - Fully TEI-compliant resources (header and body)

Metadata Crosswalks

- TEI → DC and TEI → OLAC
- Mapping from a finer-grained to a coarser scheme, hence: many-to-one situations involved, OLAC is more elaborate than DC with xsi:type information and the likes
- Some properties are determined dynamically if necessary, e.g. file size information w.r.t. dc:format and dcterms:extent
- Note: PIDs (dc:identifier) have been registered to the production server URLs for the time being
- Crosswalks proved the need for revisions and updates w.r.t. certain resources and the underlying nuOTA TEI format
- XSLT stylesheets make use of push and pull techniques, where appropriate, but it's mostly pull at the moment
- Due to performance issues we resorted to using XSLT 1.0

Plans for the Future

- Become a member of the *CLARIN SPF*
- Crosswalks in order to provide the original *TEI Headers* (with slight omissions) and *CMDI* metadata via OAI-PMH.
- Bringing the *British National Corpus* into the CLARIN realm
- Apply the lessons learnt to the *rest of the OTA resources*?
- As the web services/tools “do not seem to come to the data” in the near future, we might want to provide at least some limited *search/query functionality on OTA content* as a web service. (Talks at this workshop may be especially valuable in this respect.)
- This might also allow for *integration* into platforms like WebLicht, for example

Thank you for your attention

CLARIN has received funding from
the European Community's Seventh Framework Programme
under grant agreement n° 212230