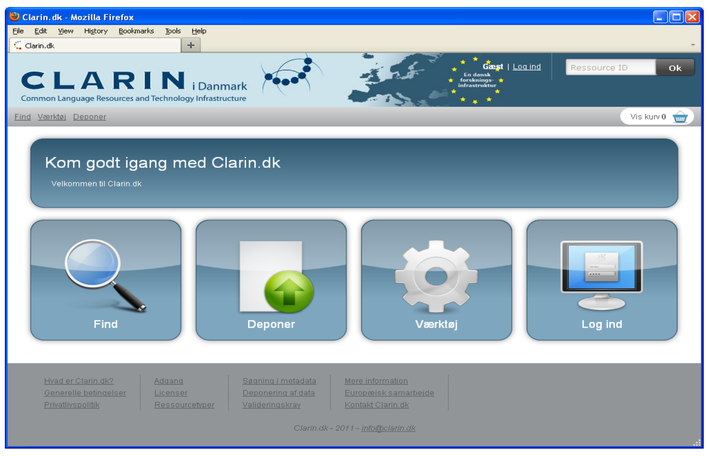
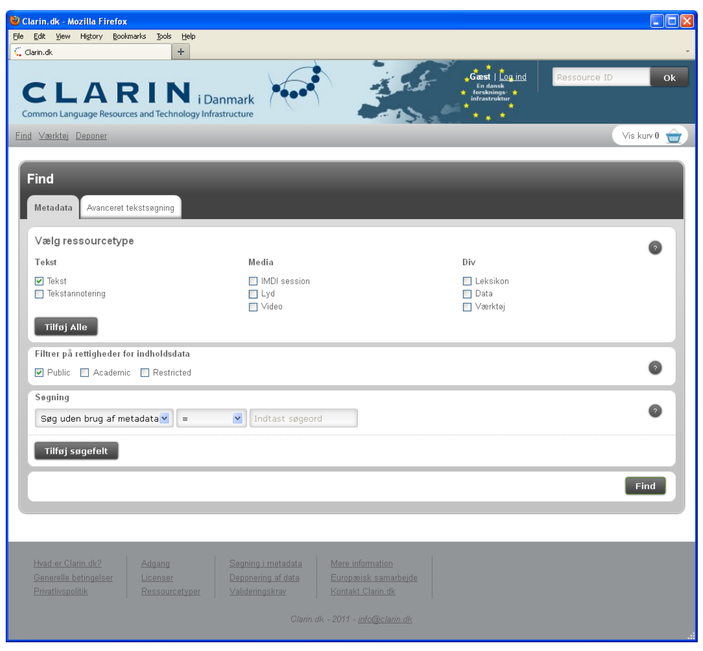
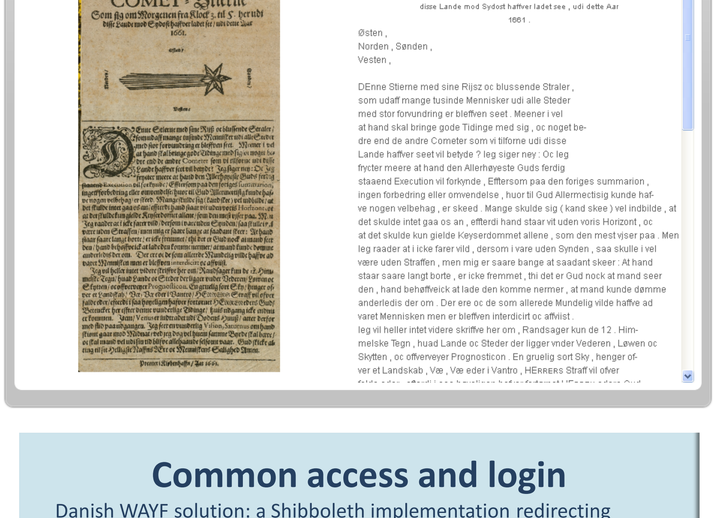
How Danish users tried to answer the unaskable during implementation of clarin.dk

Lene Offersgaard, Bart Jongejan, Bente Maegaard
University of Copenhagen, Centre for Language Technology



Search and viewing





What is clarin.dk

A research infrastructure for the humanities in Denmark. Focus is on written and spoken language resources, multimodal resources and tools.

Researchers' needs

- Repository for sharing and preserving resources and tools from project to project.
- Standardized ways to specify formats and metadata about resources, without losing diversity needed by research
- Access to the repository without having to use yet another account
- Easy inclusion of new researchers, students and institutions
- Search features for resources from all institutions even if access rights are restricted
- Combined search in metadata and content for text resources
- Easy access to and use of tools

The resources

The diversity of resources included:

- Contemporary and old, general language and specialised sublanguage texts, as well as parallel corpora with Danish as one of the languages.
- Annotations of these texts
- Audio and video recordings of spoken language and gestures
- Media annotations of these in XML and non-XML-formats
- Lexicon resources covering computational dictionaries and dialect dictionary
- Tools, both to be integrated in repository and tools to be stored for user download
- A few other resources of various types: tree banks and grammars

User involvement

- Focus groups for design of web interface, iterative meetings
- Focus groups for other specific design issues
- Iterative implementation
- Standardising resources, users choosing file formats, metadata formats and metadata elements
- All user wishes for optional metadata were accepted
- We value the involvement of the users. The dialogue and the iterative process will be continued where relevant, in the follow-up of the project

Standardising resources

TEI P5 is used for simple text, text in a specific TEI P5 DK-CLARIN format, text annotations and lexicon metadata.

IMDI is used for metadata for audio, video and media annotation metadata.

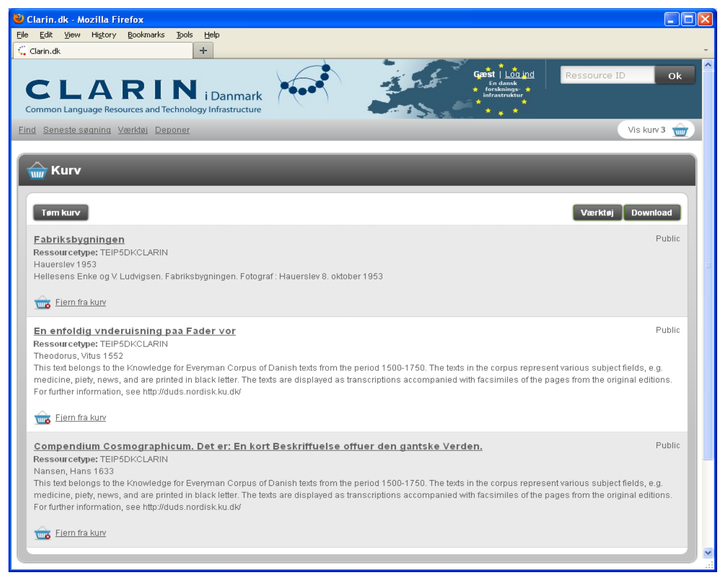
CMD framework for the resource types "data" and "tools"
http://www.clarin.eu/cmdi

Common metadata is in OLAC, for easier search
Dublin Core used for OAI-PMH metadata harvesting

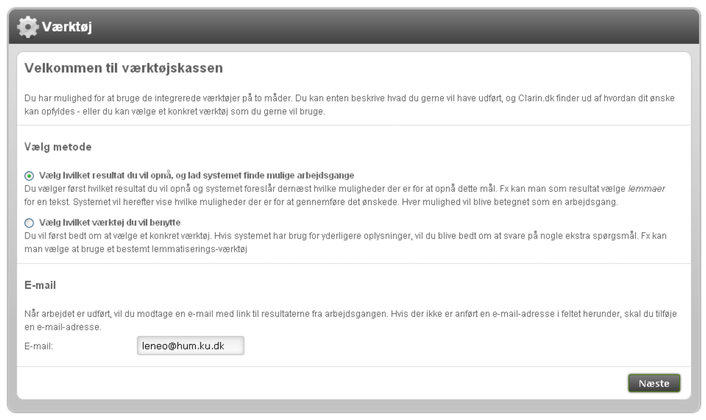
Common access and login

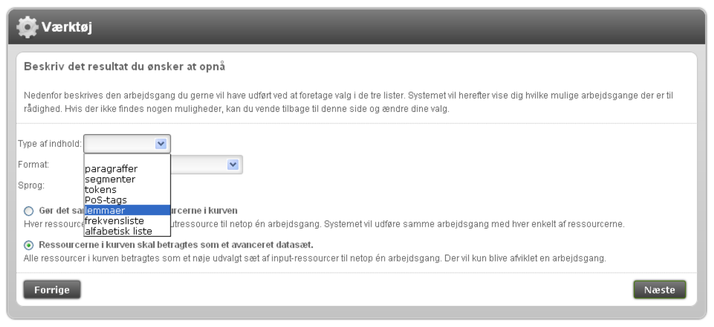
Danish WAYF solution: a Shibboleth implementation redirecting authorization back to the users' home institution, thereby letting these institutions handle the authentication and authorization of the user.

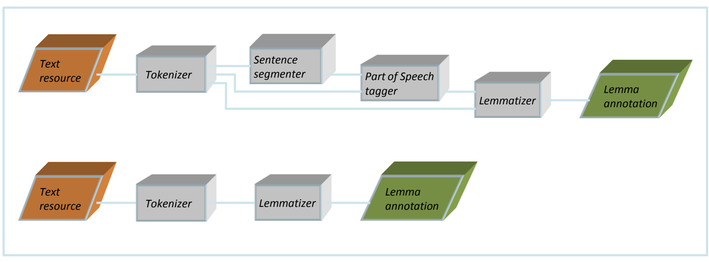
Basket



Tools and workflow planner







Future

From 2012, clarin.dk is participating in CLARIN ERIC – an European collaboration of research infrastructures. On national level clarin.dk is part of the upcoming Danish national research infrastructure for the humanities, Digital Humanities Laboratory, and the repository will continuously broaden use and facilities.

DET HUMANISTISKE FAKULTET
KØBENHAVNS UNIVERSITET

Technical implementation

clarin.dk uses a Service-Oriented Architecture (SOA). The implementation is based on eSciDoc (The Open Source e-Research Environment) and The Fedora Commons repository system. All XML-files are stored in a separate database, MarkLogic, which also provides xml search facilities.

Funding

The DK-CLARIN project is supported by the Danish Agency for Science, Technology and Innovation. Project members are eight Danish universities and cultural institutions.

Background information: http://DK-CLARIN.ku.dk/english
Infrastructure: http://clarin.dk

Acknowledgements

We thank users and developers from:
University of Copenhagen
University of Southern Denmark
University of Aarhus
Copenhagen Business School
Society for Danish Language and Literature
Danish Language Council
The Royal Library
The National Museum of Denmark.