CLARIN-DK – status and challenges

Lene Offersgaard, Bart Jongejan, Mitchell Seaton,
Dorte Haltrup Hansen

University of Copenhagen
Njalsgade 140, DK-2300 Copenhagen S
Denmark

leneo@hum.ku.dk, bartj@hum.ku.dk, seaton@hum.ku.dk, dorte@hum.ku.dk

ABSTRACT

The initiative CLARIN-DK (starting as a Danish preparatory DK-CLARIN project) is a part of the Danish research infrastructure initiative, DIGHUMLAB. In this paper the aims, status, and the current challenges for CLARIN-DK are presented. CLARIN-DK focuses on written and spoken language resources, multimodal resources and tools, and involving users is a core issue. Users involved in a preparatory project gave input that led to the current user interface of the resource repository website, clarin.dk. Clarin.dk is now in the transition phase from a repository to a research infrastructure, where researchers and students can be supported in their research, education and studies. Clarin.dk works with a Service-Oriented Architecture (SOA), uses eSciDoc and Fedora Commons, and is primarily based on open source solutions. A key issue in CLARIN-DK is using standards such as TEIP5, IMDI, OLAC, and CMDI for resource metadata. Optional metadata fields suggested by users have been included when it could comply with the standards, allowing for the diversity needed when describing the research material. Current work includes normalising metadata naming in the search pages, and making search more user-friendly by adding selectable pick-lists for query values. Also a consolidation of metadata quality is currently performed by changing some metadata values to a more harmonized set of values. All deposited metadata are maintained. Clarin.dk will apply for assessment as a CLARIN ERIC B centre in 2013 enforcing the sustainability and persistency of the infrastructure. Clarin.dk has already joined the national identity federation WAYF, implemented SSL-certificates, and offers harvesting of metadata via OAI-PMH as part of the CLARIN centre requirements.

KEYWORDS: Infrastructure, Language Resources, Repository, metadata, CLARIN.

Lene Offersgaard, Bart Jongejan, Mitchell Seaton and Dorte Haltrup Hansen 2013. CLARIN DK – status and challenges. *Proceedings of the workshop on Nordic language research infrastructure at NODALIDA 2013.* NEALT Proceedings Series 20 / Linköping Electronic Conference Proceedings 89: 21–32.

1 CLARIN-DK - a part of the Danish Infrastructure DIGHUMLAB

1.1 DIGHUMLAB

The Danish national, distributed research infrastructure initiative DIGHUMLAB[1] was launched in 2012 and has as goal to integrate and promote digital resources, communities, tools and opportunities to Danish researchers in the humanities and social sciences, and also at European and international levels. The DIGHUMLAB infrastructure initiative funds three focus areas, one being *Language based resources and tools*. CLARIN-DK covers current activities in this area, in close collaboration with the European CLARIN ERIC. One of the other focus areas of DIGHUMLAB is the Danish contribution to the ESFRI project DARIAH.

1.2 From preparatory phase to infrastructure

The initiative CLARIN-DK started as the preparatory Danish DK-CLARIN project (http://dkclarin.ku.dk/english) with the aim of creating a research infrastructure for the humanities, focusing on written and spoken language resources, multimodal resources, and tools. The project was a joint effort of eight leading Danish humanities institutions: four universities and four cultural institutions; at the same time it was a joint effort of researchers and developers. The project specified and implemented the first version of the clarin.dk research repository in the same time frame as applied for the European CLARIN preparatory phase project. This timing issue made it difficult to take full advantage of the findings and solutions of the European CLARIN project. As part of the DIGHUMLAB infrastructure, the opportunity to collaborate with the CLARIN ERIC partners is now available and will be addressed in section 7.

Clarin.dk is now in the transition phase from a repository integrating some tools to a research infrastructure, where researches and students can be supported in their research, education and studies. A part of the work in CLARIN-DK is therefore to carry out a user survey, where researchers on all Danish universities with faculties of humanities are invited to dialogue meetings on their own premises. Involving users is a core issue for CLARIN-DK and the challenge is to rethink the user's plans and wishes into technical feasible solutions given the available resources with respect to data, tools and time.

1.3 Objectives

During the implementation phase until 2017 CLARIN-DK will work on:

- Collect digital material and make it available – specifying metadata and using standards
- Collect and create tools and make them available – focusing on interoperability
- Implement user interface facilities to make it easier for researchers and students to benefit from the current infrastructure

[1] http://dighumlab.dk

- Disseminate knowledge about language resources and tools - facilitating knowledge sharing at both national and European level.
- Provide a CLARIN technical centre – improving the current technical infrastructure with persistency as an important aim.

We anticipate that the research methodology of the researchers using clarin.dk will be influenced by the new opportunities to include data and tools in their research leading to quite new experiments and experiences for the benefit of research.

1.4 Content of repository

The current version of clarin.dk can be found on http://clarin.dk. The available content is a number of resources collected in collaboration with researchers involved in the preparatory DK-CLARIN project. More information can be found in (Fersøe and Maegaard, 2009), (Asmussen, 2011a) and (Asmussen, 2011b).

The diversity of resources can be seen in table 1, also stating the numbers of current resources included. For all resources metadata are publicly available, but the resources are either available for public or academic use. More details about the metadata can be found in section 4.

Resource type	Description	Count
Text	Contemporary and old, general language and specialised sublanguage texts, as well as parallel corpora with Danish as one of the languages.	45.156
Text Annotation	Annotations of the texts above	140.921
Audio	Audio recordings of spoken language	37
Video	Video recordings of spoken language and gestures	80
Media annotations	Media annotations of audio and video recordings in XML and non-XML-formats	45
Lexicon and knowledge resources	Lexicon resources covering computational dictionaries and dialect dictionary	3
Tools and services	Tools integrated in the infrastructure as services, and tools stored for user download	11
Other Data	A few other resources of various types: zipped corpora and datasets.	9

TABLE 1: Resources available in clarin.dk, Mar2013. Resource types will be added when needed.

2 Researchers' needs

One concern is to build a repository; to build an infrastructure is in our view quite another assignment. A repository mainly focuses on preservation of data letting users

reuse resources. An infrastructure on the other hand, should enable the researchers to work with the resources and to extract new knowledge. We believe that involvement of the potential users when creating a research infrastructure is very important and must play a role during all the phases from specification to test of implementation.

2.1 User dialogue

Already in the interviews with potential users in the preparatory DK-CLARIN project, we realised that it is very difficult for users to imagine their requirements to repository facilities enabling a new digital or data-driven angle on their research. The baseline sketched was to make existing tools and available data integrated in the same platform, thus providing the opportunity to experiment with tools and data. Especially a streamlined common format for as many resources as possible and the possibility to access all available Danish data sources from one single repository was seen as a great benefit. However, the researchers also agreed that for resources that already are available in other databases or through other user-interfaces it should in each case be considered whether it would be beneficial to make the resources and tools available through clarin.dk, freeing the researcher from administering their data.

In collaboration with the researchers a list of issues were prioritized during implementation. The researchers wanted a repository to handle easy storage, sharing and use of resources:

- A repository to deposit data material and tools, to preserve these resources from project to project and to share them with others.
- 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 metadata 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

An on-going series of dialogue meetings has already invited researchers from all the Danish faculties of humanities to meetings on their own campuses. These meetings have been supplemented with a series of meetings with potential data providers, focussing on making new resources available. Currently results from the on-going dialogue meetings, that cover wishes for functionality and the availability of new resources, are being evaluated. New issues may be added to the list above as a result of this process.

A key desire from the text researchers was an advanced text search facility combining metadata and content search and the possibility for extracting the resulting list of resources, including both texts and annotations. On the basis of this extract the researcher could then create a tailored annotated corpus search application that could be available for research and teaching as long as it was needed. From a user's point of view this seems simple, but for the developers the varieties of text and annotation

formats and the possibility of an undefined number of diverse annotations for each text pose a problem. This task will be one of the focus areas in the upcoming work.

3 User functionality of clarin.dk

Currently the resource types mentioned in table 1 can be uploaded in clarin.dk and stored with metadata. An example of the view of a resource, “En nøttelig Legebog”, a book about health from 1533 can be seen in figure 1. As the available version only covers a part of the original book the resource title include “Del A”, indication that the whole book is not available. All resources are before upload described by type specific metadata by the data provider, and the metadata are deposited together with the resource itself. During depositing the platform checks that the resource and its metadata must validate against the CLARIN-DK specific TEIP5, IMDI and RNG schemas that are applicable for the resource type. There is no top down control of the metadata used as long as the content of the metadata comply with the decided standards and schemas.

Clarin.dk offers a toolbox with a number of tools that can be used to process the resources including a workflow planner that allows the user to focus on their goals, letting the workflow planner find the appropriate tools given the requested result, e.g. a frequency list. The current tools mainly offer text annotation and conversion, but also an OCR tool and a speech to text tool are integrated. More details about tools and workflows can be found in (Jongejan, 2013) and (Offersgaard, 2011).

The user can also search for resources based on metadata information, inspect all metadata, and according to permissions: inspect, use or download the resources.

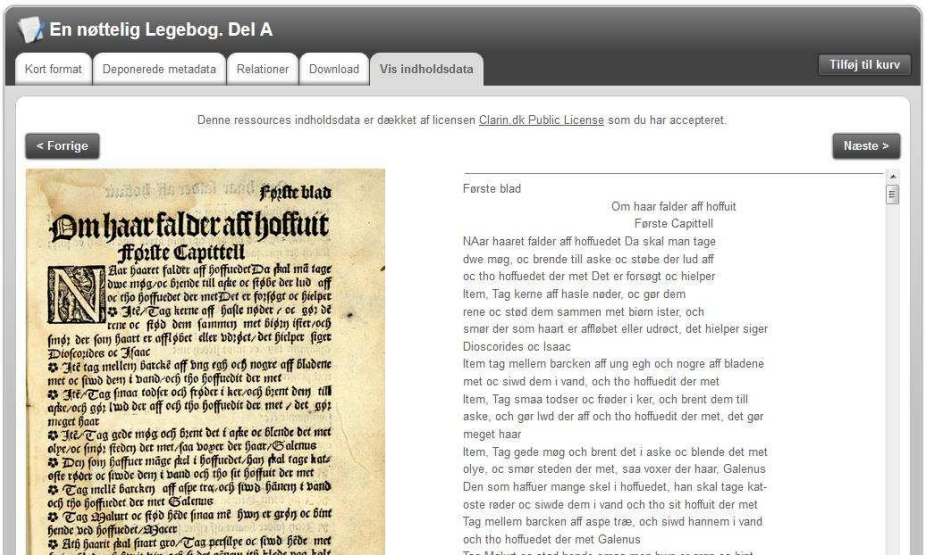

FIGURE 1: A resource containing part A of the old book “En nøttelig legebog” on the subject health from 1533. Url: https://clarin.dk/clarindk/item.jsp?id=dkclarin:597250#show_content-4

We will now go into detail with the work on metadata for resources and certain changes that are currently carried out in CLARIN-DK.

4 Standardising metadata for resources

It is important for infrastructure initiatives to give high priority to the use of current standards and already known and used formats. When collecting materials the existing resources will appear in a number of standards and formats, and it can be a challenging task to agree on these standards and formats. So streamlining of the resource metadata according to the CLARIN recommendations (Hinrichs, 2009 and Broeder, 2012), creates opportunities when sharing. In CLARIN-DK, the CLARIN recommendations were taken into account by letting clarin.dk enable all resources - independently of resource type – to get metadata in both CMDI and OLAC format; CMDI to fulfil the CLARIN recommendation and OLAC to make the metadata harvestable through the OAI-PMH protocol.

4.1 Resource specific metadata

For each resource type the relevant users were involved in selecting both the relevant metadata and relevant formats. As an effect of the user involvement in the metadata specifications, all user wishes for optional metadata were accepted, i.e. the developers accepted the wish for diversity in ways to describe the research material.

The users have chosen to use different standards for expressing the resource specific metadata. TEI P5[2] is used for simple text, text in a specific TEIP5 DKCLARIN format, text annotations and lexicon metadata. However, TEI P5 is not suitable for all tasks, and IMDI is therefore used for metadata for audio, video and media annotation metadata. The CMDI framework[3] provided from the European CLARIN project was much appreciated for specifying metadata for the resource types “data” and “tools” as no other current standard fulfilled these metadata requirements in a concise manner. The researchers are then able to specify metadata in widely used and well-known metadata formats, and the provided metadata are processed after upload and converted to metadata records in OLAC and CMDI to fulfil the CLARIN recommendations.

Resources in clarin.dk share a set of 15 common metadata elements, a subset of which is obligatory for all resources. The benefit of using a common core set of metadata elements is that it forms a common basis for the metadata search in the user interface. These metadata elements are supplemented with 72 metadata elements that are resource specific, e.g. *Translator* for texts, *DataFormatIn* for tools, *ActorRole* for multimedia and *Dialect* for lexica.

[2] http://www.tei-c.org/Guidelines/P5/

[3] http://www.clarin.eu/cmdi

The obligatory metadata set is basically the core OLAC metadata set. The elements *Coverage* and *Relation* have been omitted and instead the elements *Conforms to Standard* and *Size* have been added. Obligatory metadata are: *Title*, *Language* (not obligatory for data and tools), *Resource type*, *Creator*, *Creation date*, *Subject* (not obligatory for data and tools), *Description*, *Source title* (only for text), *Publisher*, *Publication date*, *Sponsor*, *Data provider*, *Format*, *Conforms to Standard*, and *Size*. In addition all resources get *Id* and *Rights* metadata when they are uploaded to the repository.

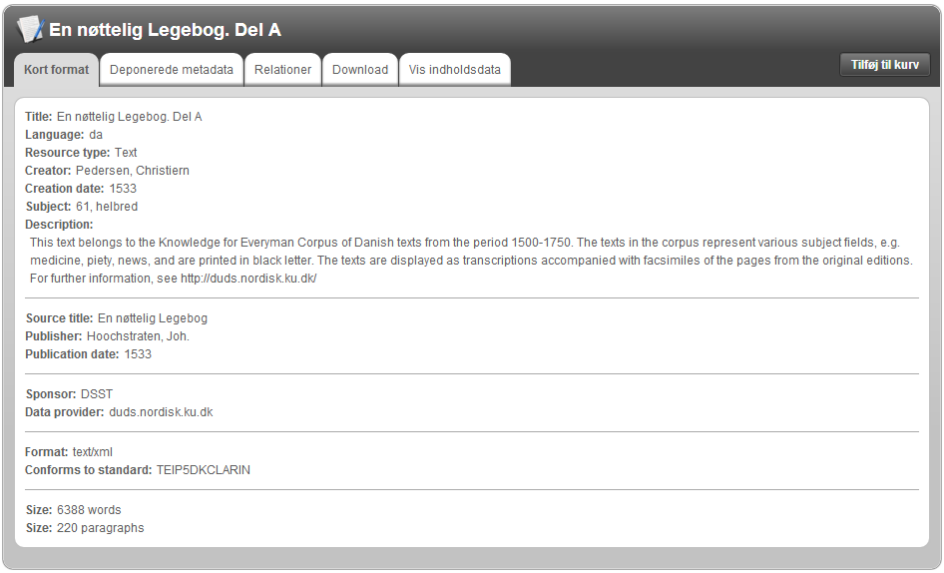

FIGURE 2: Screen dump of general metadata for the text document in figure 1. Url: https://clarin.dk/clarindk/item.jsp?id=dkclarin:597250

4.2 Normalising metadata making search more user-friendly

One challenge in clarin.dk is to display the metadata for very different resource types in a uniform way. There is a huge diversity in the contents of metadata fields, because many fields can hold free text and because the validation schemas are non-strict. The benefit is obviously great flexibility when uploading resources, but the cost can be great complexity when searching for resources. Our task is to balance this trade-off which we do by mapping the original metadata to two new set of metadata, currently focussing on OLAC but later also using CMDI. As an example we can look at the rather simple metadata field *Subject*, which can also be named subject domain. In TEIP5 it can be found in both the field *Domain* and to some extent also in the field *CatRef* referring to a *CatRef* Scheme. So first of all we need to decide which metadata field to map to the OLAC metadata *Subject*. Currently we have decided to map *Domain* in TEIP5 to *Subject* in OLAC, but reviewing this with users might lead to a join of both *Domain* and *CatRef* to the same searchable Subject field.

The text providers agreed to a large extent on how to use the special metadata format TEIP5DKCLARIN, see (Asmussen, 2011a) for more details. When users are applying the decisions to real metadata, a diversity not earlier recognized can turn out to be needed. This can be seen for the values for *Subject*. As values we find *9999999999, n/a, general, Denmark in recent times, Health and medicine, Law, Nanotechnology, Agriculture, 20-religion, 61-health, 641-food, 641-food 631-agriculture* and more - a great variety in both naming, structuring, coverage, granularity and point of departure. A fixed list of values would clearly not be satisfactory for neither the data provider nor the future user of the data. Taxonomies are very difficult to use across disciplines.

In clarin.dk we are currently making the metadata search more user-friendly by applying two major changes. In the search interface we are implementing updateable pick lists for the values used in the selected metadata fields *Language, Resource type, Subject, Data provider, Sponsor, Format* and *Conforms to Standard*.

The other change is to rename some values that cover the same to one common value e.g. changing *Health and medicine* and *61-health* to the common term *Health (DK5-61)*, as the numbers used refer to the Danish Library classification system DK5[4]. These changes are not done in the deposited metadata, but in the special common set of metadata that are available for all resources and that are searchable, so the original, deposited metadata will always be available and preserved unchanged. To give an impression on the amount of changes, we will for 21855 resources change the subject from *9999999999* to *Unspecified* and 71 resources subject will be changed from *61-health* to *Health (DK5-61)*. But we will keep the diversity that is specified by data providers by allowing the addition of new subject areas when needed, as we think users from different research areas will need different classification systems.

5 Other challenges

Currently we are also working on making the search interface more intuitively easy to use, a difficult task when the end-users potentially come from all research disciplines in humanities.

When the user searches for relevant resources, it is currently not easy to get a quick overview of the available resources. Therefore re-development of the clarin.dk search pages are carried out during 2013 and more web-pages informing the user what is available and how to use the resources will be implemented.

Implementing a metadata-editor is also planned for this year, as this will make it easier for new users to upload resources. Now the metadata has to be specified in advance. For a text the user will specify metadata in a TEIP5 header which is validating against a special CLARIN-DK TEIP5 schema, before the text is ready to be deposited. In the future users will be able to specify the metadata in a special editor that guides the user to select the right resource type to deposit and the possible metadata fields to fill out.

Currently the resources in clarin.dk can have relations to other resources, but this will be extended by implementing the collection resource type. Collections can be used to

[4] A Danish variant of the Dewey Decimal Classification

bundle homogeneous text resources in a text corpus or to relate all resources from the 1950's to a heterogeneous collection that can hold not only texts but also e.g. images and audio from this period.

Users of texts have strongly recommended that clarin.dk provide a combined view of texts and annotations with search facilities included. As mentioned before this task has some challenges as texts and annotations can be from a number of different data providers and is still under specification.

Important tasks will also be to help researchers deposit their data in clarin.dk, both data that can easily be converted and deposited using the current resource types, but also allowing for new resource types to be added when research work and resource preservation has a need for it.

More tools will be available, but focus will mainly be on tools broadly usable for research tasks. Clarin.dk has planned to apply for assessment as a CLARIN B-centre during 2013, see section 7.

6 Implementation

Clarin.dk works with a Service-Oriented Architecture (SOA), and uses the eSciDoc (The Open Source e-Research Environment [5]) Infrastructure as stable middleware for authentication, access, submission, search and modification of the clarin.dk repository. The clarin.dk architecture is built on top of many open source products and technologies, such as Apache HTTP server, Apache Lucene, JBoss Application Server (AS), PostgreSQL database, and FedoraCommons[6]. MarkLogic[7] is enterprise software, a XML database for performance, reliability and scalability of the large data store required for our XML data storage needs. eSciDoc provides essential middleware services, defines our content model for the repository, and assists with indexing the repository collection so we are able to provide efficient search features and usability to the user functions.

To allow easy login administration the Danish WAYF solution was chosen: a Shibboleth implementation redirecting authentication to the users' home institution. This fits with the recommended solution from CLARIN, and is a flexible and easy solution for user administration.

6.1 Technical details

The backend infrastructure implementation is based on eSciDoc and the FedoraCommons repository system. FedoraCommons is a common component used by two thirds of the current CLARIN centres[8]. All associated XML content files of an

[5] https://www.escidoc.org/

[6] Flexible Extensible Digital Object Repository Architecture http://www.fedora-commons.org/software

[7] http://www.marklogic.com/

[8] Overview CLARIN centres: http://clarin.eu/node/2971 and https://centerregistry-clarin.esc.rzg.mpg.de/

eSciDoc item are referenced and stored in the separate MarkLogic database, which also provides efficient search facilities. More details can be found in (Conrad 2010)

Clarin.dk services are built on top of eSciDoc, to provide four integrated components – *Deposit, Search, Deliver*, and *Tools*. The *Deposit* service enables the submission of packaged resources for processing and then inclusion into the repository. Submitting users are notified via email of the newly published resources once the depositing process is completed. The *Search* service provides an interface to eSciDoc's core SRW/U search service. Clarin.dk utilises SRU/CQL, and Lucene indexes, to query specific, indexed metadata fields. SRU is recognised as an OASIS standard, and CQL is a common formal query language. An offline copy of a resource can be obtained via the *Deliver* service, with the possible inclusion of annotations, and choice of offline format for the selected resources. The *Tools* service enables the ability to run tools services or programs, on a collection of resources to obtain a desired offline output. Backend services are programmed in Java, and deployed in a WAR package through JBoss AS. The PostgreSQL database is used with eSciDoc/FedoraCommons, as well as keeping a record of Clarin.dk deposits and their status.

The front-end environment of clarin.dk is formed in a HTML (HTML5) web application, with cross-browser compatibility. AJAX capabilities of the web browser, event handling and DOM manipulation are used via the jQuery JavaScript library. The JSON data format is used for passing defined static data and configuration to the front-end web-pages. JSP (Java Server Pages) are used primarily to push and display dynamic content to the browser, from server-side HTTP requests, and Session (JavaBean) stored user data. Typically, CSS (Cascading Style Sheets) are used to control and format the web-page aesthetics and layout. Some additional jQuery plugin libraries are used, such as Validation, Cookie, ColorBox (overlay iframe) and a custom UI library. Documentation web-pages are presented on the website, where they provide usable and detailed information to assist the user.

7 Assessment to become a CLARIN ERIC B Center

To ensure interoperability among the CLARIN centres in Europe, national centres can apply for assessment as a CLARIN B centre[9]. The assessment criteria[10] for becoming a B centre includes a number of requirements concerning e.g. having a repository system that can pass a quality assessment procedure as the *Data Seal of Approval*[11], joining a national identity federation such as WAYF, servers having SSL-certificates, offering harvesting of metadata in CMDI format with the OAI-PMH-protocol[12], and associating persistent identifiers to resources and metadata.

Clarin.dk will apply for assessment in 2013 enforcing the sustainability and persistency of the infrastructure. Clarin.dk has already joined WAYF, implemented SSL-certificates, and offers harvesting of metadata via OAI-PMH. Persistent identifiers, PID's, will have

[9] CLARIN center types: http://www.clarin.eu/page/3542

[10] Checklist for CLARIN B Centres: http://www.clarin.eu/page/3577

[11] http://datasealofapproval.org/

[12] Open Archives Initiative Protocol for Metadata Harvesting http://www.openarchives.org/pmh

to substitute the current resource id's in clarin.dk as these are not defined as persistent; metadata harvesting will need to be extended from DC and OLAC formats to also include CMDI, and the *Data Seal of Approval* must be achieved.

8 Conclusion and Outlook

Clarin.dk is in the transition from a repository to a research infrastructure, building on the results in the preparatory project DK-CLARIN. Clarin.dk is currently a repository with already over 186000 resources. By a user dialogue processing user requirements from the Danish faculties of humanities are now collected. These requirements will be used as guidelines when prioritising the coming extensions to clarin.dk.

Use of standards are widely implemented in clarin.dk, but as clarin.dk is meant to handle language resources for many disciplines a number of standards have to be brought into use. The data providers and researchers will have different ways to work with the resources and will need different facilities to support their research, so clarin.dk is now going into a development phase where researchers that volunteer to help specifying and using clarin.dk in research and teaching have a chance to influence the future facilities. The focus will still be resources containing language.

Currently some changes to content and display of metadata are carried out, and extensions with a metadata editor and better search facilities are planned for the rest of 2013. Deposited metadata has shown that even given a detailed metadata specification as for the text resources in clarin.dk, data providers fill out the metadata fields with a broad variation. Some variations are simple errors, other are caused by different ways to interpret specifications. In the future the metadata editor should guide the user in specifying metadata but still permitting a large variety of values if needed.

Important tasks are to help researchers deposit their data in clarin.dk and to make available more tools that can extend usability of clarin.dk. Working for assessment as a CLARIN B Centre during 2013 will enforce the sustainability and persistency of the infrastructure.

Acknowledgments

DIGHUMLAB Digital Humanities Lab Denmark is supported by Danish Agency for Science, Technology and Innovation, Ministry of Science, Innovation and Higher Education. The preparatory project DK-CLARIN was also supported by the Danish Agency for Science, Technology and Innovation, as well as by all eight partner institutions. We want to thank all partners for their contribution.

References

Asmussen, J. (2011) Text metadata: What the header of a text item looks like, *DK-CLARIN WP2.1 Technical Report*, http://korpus.dsl.dk/clarin/corpus-doc/text-header.pdf

Asmussen, J. (2011) Text formatting: Bringing corpus texts into good shape and enabling flexible annotation of them. *DK-CLARIN WP2.1 Technical Report.*

Asmussen, J. & Halskov, J. (2009) Compiling and annotating corpora in DK-CLARIN. Interpreting and tweaking TEI P5. In *Proceedings of the Corpus Linguistics Conference CL2009*. University of Liverpool, UK 2009. http://ucrel.lancs.ac.uk/publications/cl2009/

Conrad, A. (2010). The use of eSciDoc in Clarin.dk. *eSciDoc Days* Copenhagen, 2010. https://www.escidoc.org/pdf/day1-conrad-clarindk.pdf

Broeder, D. (2012) CMDI: a Component Metadata Infrastructure. CMDI (Component Metadata Infrastructure) workshop, September 13, 2012 MPI for Psycholinguistics, http://www.clarin.eu/sites/default/files/cmdi-daan.pdf

Fersøe, H & Maegaard, B. (2009). CLARIN in Denmark – European and Nordic Perspectives. In: *Nordic Perspectives on the CLARIN Infrastructure on Common Language Resources*, NEALT Proceedings Series, Vol. 5, pp. 6-11. Electronically published at Tartu University Library (Estonia) http://hdl.handle.net/10062/9944.

Halskov, J., Hansen, D. H., Braasch, A., & Olsen, S. (2010). Quality indicators of LSP texts – selection and measurements: Measuring the terminological usefulness of documents for an LSP corpus. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation: LREC 2010* (s. 2614-2620). Valletta, Malta: European language resources distribution agency.

Hinrichs, E. W. (2009). CLARIN Short Guide Standards for Text Encoding. http://www.clarin.eu/files/standards-text-CLARIN-ShortGuide.pdf

Jongejan, B. *Workflow Management in CLARIN-DK*. In *Proceedings of the Nordic Language Research Infrastructure Workshop at NoDaLiDa, Oslo, May 22, 2013*

Offersgaard, L. Jongejan, B. and Maegaard, B. (2011). How Danish users tried to answer the unaskable during implementation of clarin.dk. In *SDH 2011 – Supporting Digital Humanities*, Copenhagen.