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

Lene Offersgaard, Bart Jongejan, Bente Maegaard

University of Copenhagen
Njalsgade 140, DK-2300 Copenhagen S
Denmark
leneo@hum.ku.dk, bartj@hum.ku.dk, bmaegaard@hum.ku.dk

Abstract

In this paper we describe the aims of the Danish national research infrastructure preparatory project, DK-CLARIN, 2008-2011. In the description we focus on user perspectives and the efforts to involve users, in particular for those aspects that are close to the user, i.e. resources and tools. We also describe some important aspects of the resulting implementation. Users were involved in various ways, most importantly through focus groups, in particular a focus group for the design of the web interface, and a focus group for specific design issues such as metadata search and other types of search mechanisms. clarin.dk accepts a wide variety of resources, e.g. text, text annotation, audio, video, media annotation, lexicon and tools. Those tools that are integrated in clarin.dk can be activated through a workflow planner. DK-CLARIN is now through the preparatory phase, and is ready to participate in CLARIN ERIC when it is established. The participation will be through the upcoming Danish national research infrastructure for the humanities, Digital Humanities Laboratory, planned to start early 2012.

The DK-CLARIN project

The aim of the Danish CLARIN project (DK-CLARIN; infrastructure: http://clarin.dk and background information: http://DK-CLARIN.ku.dk/english) is to create 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 clarin.dk research infrastructure 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.

The challenge

Even though Steve Jobs[1] said - *It's really hard to design products by focus groups. A lot of times, people don't know what they want until you show it to them* - we want to stress that DK-CLARIN developers heavily challenged the researchers to stretch their imagination specifying what basic repository functions they can use and/or will need in future research. The variety of resources from eight different research environments called for pretty general solutions for the repository, and therefore the result is a basic infrastructure facility, with search and viewing possibilities and a selection of resource annotation tools. Ever since the rough outline of the project the researchers contributed to the project description, and the main objective was always to create an infrastructure responding to the researchers' needs.

[1] BusinessWeek (25 May 1998)

Researchers' needs

The initial interviews of the users showed that it is very difficult for them 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 data available 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 a number of resources that already now are available in other databases or through other user-interfaces it should in each case be considered whether only metadata for these resources should be deposited or whether it would be beneficial to make the resources and tools available through clarin.dk freeing the researcher for administering the data.

A keen desire from the text researchers was an advanced text search facility combining metadata and content search and the possibility for extracts of such a search result list of resources, including both texts and annotations. On the basis of this extract one could then create a tailored annotated corpus search application that could be available for research and teaching as long 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 availability of an undefined number of annotations for each text containing different types of information, this task was only feasible with limitations. The current solution is described below in the section "Search and viewing".

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

- A repository to deposit data material and tools in order to preserve resources from project to project and in order to share resources.
- Standardized ways to specify formats and metadata about resources, without loosing 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

In the following sections we will go into more detail with the user needs. Before that we will give a brief overview of the resources in focus by the involved researchers.

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
- And a few other resources of various types: tree banks and grammars.

An overview of the different types of resources can be found in [Fersøe and Maegaard, 2009], while the research work is described in a number of publications, see References.

Standardising resources - Metadata[2] and formats

DK-CLARIN gave high priority to the use of current standards and already known and used formats. The users of course arrived with their already existing resources and it took some time to arrive at commonly agreed standards.

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.

[2] In DK-CLARIN, information about the author and publisher of the text is considered metadata, while linguistic annotations of the text, for example the annotation of lemmas, will not be considered as metadata, but as an annotation.

The users have chosen to use different standards for expressing the resource specific metadata. TEI P5 is used for simple text, text in a specific TEI P5 DK-CLARIN format, text annotations and lexicon metadata. However, TEI P5 is not suited for all tasks, and IMDI is therefore used for audio, video and media annotation metadata. The CMD framework[3] provided from the CLARIN project was much appreciated for specifying metadata for the resource types "data" and "tools" as no other current standard fulfilled the metadata requirements in a simple manner.

All resources in DK-CLARIN share a set of common metadata elements, a subset of these elements are obligatory for all resources, while others only are obligatory for one or more resource types. Besides these common metadata elements a number of resource type specific metadata elements have been specified in DK-CLARIN. 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.

Obligatory metadata:

- title
- creator
- creation date
- publication date(issued)
- format
- publisher
- description
- subject
- resource type
- language (not obligatory for data and tools)

To comply with the CLARIN project requirements metadata in clarin.dk are harvested by CLARIN with the OAI-PMH-protocol[4].

Easy access and login

The access to the DK-CLARIN repository should be easy: No special software requirements for the users, no registration procedure for neither researchers nor students. This means that a minimal requirement is a repository structure where researchers and students and also the public can access the repository and see what is available through metadata search. At the same time this fulfils the user requirement to be able to get an overview of available resources. Dependent on the access rights for the individual resources they may only be available for researchers and students or a restricted group of users, or even only the specific researcher who provided the resource.

To allow easy login administration it was chosen to use the Danish WAYF solution: a Shibboleth implementation redirecting authorization back to the users' home institution, and thereby letting these institutions handle the authentication and authorization of the user. This fits with the recommended solution from CLARIN, and is a very flexible and easy solution for user administration.

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

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

Resource types

The types of resources which can be deposited and used in clarin.dk are currently:

1. text, both simple text format and TEI P5 DK-CLARIN text format
2. text annotation
3. audio
4. video
5. media annotation, annotations in either XML, Praat or CLAN format
6. lexicon
7. tool
8. data

The list of resource types can be extended when needed.

Search and viewing

From clarin.dk's front page the search page is one click away.

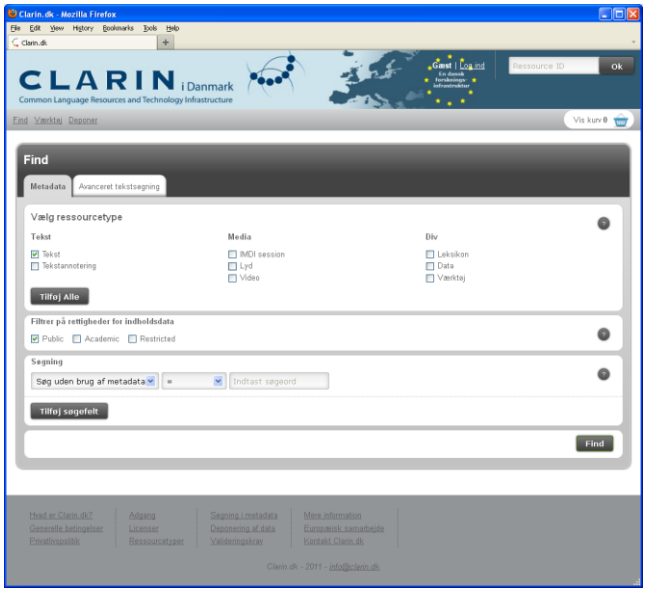

Figure 1: The search page

The search page allows the user to search in the stored resources' metadata. The search page has two tabs. The first tab is for meta-data search, while the second tab allows advanced users who master XPath to search in text resources themselves. The metadata search tab (see Fig. 1) allows the user to find resources of specified types (selected by ticking off one or more of the eight check boxes in the upper part of the window), of specified access types (public, academic and restricted) and with specified content in zero or more specified metadata fields (lower part of the window).
Clicking the 'Find' button starts the search and brings up the result page, part of which is shown in Fig. 2. Each result shows a few metadata, such as the document title. Also shown is a clickable basket-icon, a metaphor for a collection of items that the user wants to carry with her during a visit to the clarin.dk website, with the option to download the items or to use tools to annotate the collected items.

To see the full resource, clicking the title brings the user to a page where the user can see the metadata in full length and in a somewhat condensed form, and where the user also can download and view the resource. The latter is currently not implemented for all resource types, but for example resources of type 'text with images' can be viewed already, as illustrated in Fig. 3.

Figure 2: Search results

When searching using metadata the user gets a dynamic drop down list with the metadata available for the chosen resource types. As an example the user can search for resources created in 1720'ies by choosing the metadata field "CreationDate" and specifying "= 172*". User experience with metadata search have revealed that even when the metadata formats and expected content are thoroughly described through technical reports e.g. [Asmussen 2011], large diversity can be seen in the deposited metadata. Further more it can be difficult for users unfamiliar with metadata creation to choose the right metadata fields when searching. The users have therefore reported that the metadata search has to be extended with a more user-friendly interface, including drop-down lists for closed value list and help functionality for each metadata field.

To provide the user with a tailored annotated corpus search application, currently the repository must rely on the user collecting texts with the same type of annotations in the basket. With a collection of similarly annotated texts the user can go to the basket and choose to download the collection in a number of ways. One of the options is an intertwined CQP[5]-ready file of the selected texts and their annotations. This file can be downloaded and imported in CQP

[5] http://cwb.sourceforge.net/

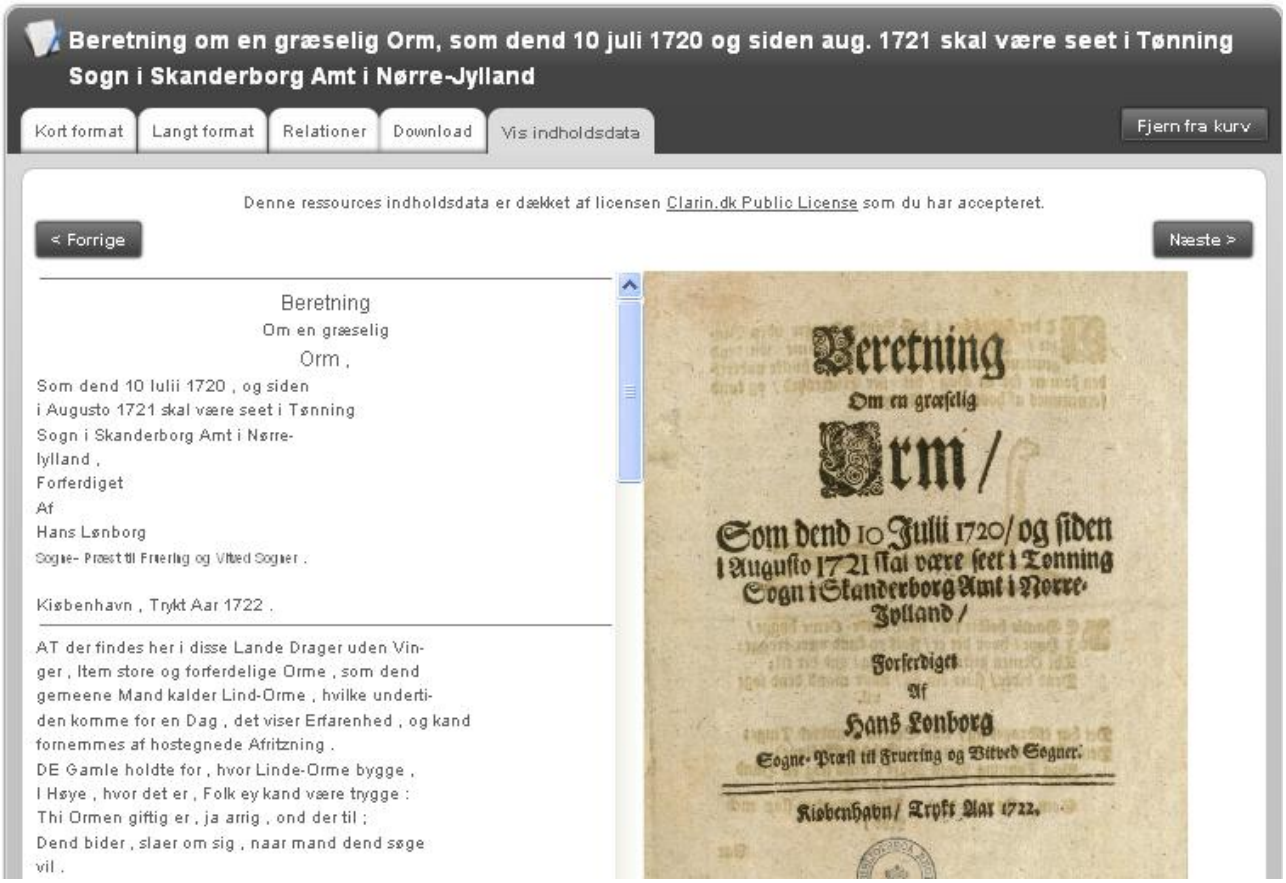

Figure 3: Viewer for resource of type 'text + images'

Integration of tools

To select the right tool or the right series of tools for e.g. creation an annotation of a text can be difficult for a user who has no detailed knowledge of the accessible tool(s) and their requirements as to formats, tag sets and so on. Especially in the clarin.dk infrastructure, where the user encounters various new tools, she probably wants to avoid the tedious specification of tools and their correct settings, but rather focus on the results. The infrastructure's Tools module therefore includes a workflow planner, where a user only needs to specify the kind of annotation she wants. The workflow planner will then list the possible ways to get to this result. (See Fig. 4.) The user interface also allows the user to select a tool from a list and to run that tool with selected input, without going through a workflow planning stage. This method can be used by experts or e.g. during courses when the students are focusing on one specific tool.

Activation of tools

To work with tools, the user must first have selected one or more resources and put them in the basket. If there is more than one resource in the basket, the user must tell the Tools module whether these resources should be acted upon one by one or whether they should be regarded as a complex set of input data to a single workflow traversal. Also, the user must choose whether she wants to select a tool and send the input to that specific tool, or whether she wants to specify her goal instead and leave the choice of appropriate tools to the Tools module to decide. If the user selects a tool, there is the possibility that the chosen resource(s) and the chosen tool don't match. In some cases that is a risk that is worth to take. For example it might be possible to tokenize a Danish text with a tokenizer that is designed for English. In other cases the tool may return an error message when it discovers that e.g. the format of the resource is not understood.

Workflow planner

The workflow planner is an advanced piece of software. It can be compared with an trip planner for journeys on the surface of our fair planet, with a few enhancements: the user doesn't need to state where she starts from and the user can state her destination in terms of city name, type of accommodation, or degree of tanning after one week near the pool, or a combination of these. Instead of city, type of accommodation and tanning, the Tools workflow planner has *language*, *format* and *facet.* The *language* feature doesn't need explanation. The *format* feature describes the packaging, for example whether the output should be cast in some XML-format, in a comma separated format, or perhaps in a JPEG format. The *facet* feature is the most important feature and the least standardized one. It is the feature that defines what "added

Figure 4: The user can choose between the workflows that lead to the goal *lemmas* specified in the preceding step. (Not shown here). The first workflow involves four tools, the second only two. The flow of data is quite complex in the first workflow, as indicated by the references ① back to the tokenizer's output.

value" a tool produces other than translations from one format to another format or from one language to another language. So, a lemmatizer's added value is the list of lemmas of the words in the input. In the same way, an OCR-reader's added value is the text hidden in a picture. From these two examples we can learn that for example "image", "text" and "lemmas" are facets. They offer different ways of looking at the same resource, a newspaper for example.

The workflow planner digs even one level deeper by allowing further specification of each feature. In the trip planner example, the extra level of detail corresponds to adding "Ibis" or "Hilton" if the accommodation is "hotel", or "City centre" or "airport" if the city is "Copenhagen", or "tan lines" or "all body" if the tanning is "Dark intermediate". Thus, for each feature *value* there is a possibility to list minor variations that may apply. The workflow planner in the Tools module uses these sub-specifications of features in a forgiving way: if a tool requires input characterised by the facet "tokenized" and the further specification that the tokens must follow the rules as stated for the Penn Treebank corpus (splitting *can't* in *ca* and *n't*), then an input said to contain tokens, without any sub-specification, is regarded as acceptable. If it also is said that the tokens in the input are created by merely surrounding all punctuation with white space, then the input is not considered acceptable anymore, because the sub-specifications don't match. Conversely, an input consisting of Penn Treebank tokens can be accepted by a tool that accepts tokens, without further specification. One has to bear in mind that the acceptance of an input also depends on other features, if they are specified. If languages or formats don't match, it is not of much help that the facets match. The extra level of specification is normally hidden from the users of the infrastructure and first and foremost introduced to allow 'fuzzy fits' between tools that can usefully cooperate without being perfectly tuned to each other. The user is only confronted with the extra level if there is a choice between workflows that are identical at the tool and feature levels and only differ at the feature sub-specification level.

Integration of a tool

For the special group of infrastructure users who can provide new tools to the infrastructure, a special web service is made to make integration a smooth process. In many cases, new tools can be integrated without any involvement of a clarin.dk administrator. If the new tool handles facets, formats or languages that are not yet defined in the infrastructure, involvement from a moderator is needed to extend the tables of facets, formats and languages as needed. The moderator's role is to make an educated assessment of whether additions really are needed or whether sub-specification of existing facets, formats and languages would be a better decision, allowing extensive cooperation with already registered tools. The web service for tool integration is a wizard guiding the tool provider through a number of choices from fixed sets of values. Only the 'boilerplate'-section of the registration form (tool name, description, version, etc.) requires some typing at the keyboard – all other actions are point and click actions. After filling out the registration form the tool provider can (and should) deposit the tool in the repository. The deposited information comprises the boilerplate data and a condensed version of the language and format information. Only deposited tools are searchable as tool resources in the clarin.dk infrastructure. Also tools that are not integrated can be registered and deposited in the infrastructure, optionally together with a downloadable installation file.

Tool resources are the only type of resources for which metadata can be created using a facility in the clarin.dk infrastructure. On the downside, the tool provider wanting to integrate a tool in the infrastructure must ensure that the tool can communicate with the infrastructure using the defined protocol. The protocol is well documented and help can be provided by the maintainers of the infrastructure.

Example of integrated tool

Here is an example of the data that is stored for an integrated tool. The examples are based on the real data for UCPH-CST's lemmatizer. This lemmatizer supports ten languages. English (`en`) and Danish (`da`) are special, because the lemmatizer can utilize extra part-of-speech information (`pos`) on the input tokens for these two languages only. The lemmatizer can handle flat text (`flat`), DK-CLARIN's XML-format for text annotations (`txtann`) and also unspecified XML (`xm`). If the language is English and if extra PoS-information is provided, then the

PoS-tags must be selected from the Penn Treebank tag set (`PT`). Other tag names are not understood. If the language is Danish and if extra PoS-information is provided, then the PoS-tags must be selected from the Parole tag set for Danish (`Par`). As a true Swiss Army knife, the lemmatizer not only can output lemmatized text (`lem`), it can alternatively output a list of lemmas, either sorted alphabetically (`alf`) or according to frequency (`frq`). If the output is one of these lists, the format of the output is flat text. All this information is succinctly stored in the following data structure:

```
  ( CST-Lem
  .    (facet,(tok.lem))
       (format,(flat.flat)+(txtann.txtann)+(xm.xm))
       (lang,(de.de)+(el.el)+(fr.fr)+(is.is)+(la.la)
             +(nl.nl)+(pl.pl)+(ru.ru))
  )
+ ( CST-Lem
  .    (facet,(tok.alf+frq))
       (format,(flat+txtann+xm.flat))
       (lang,(de.de)+(el.el)+(fr.fr)+(is.is)+(la.la)
             +(nl.nl)+(pl.pl)+(ru.ru))
  )
+ ( CST-Lem
  .    (facet,(tok pos^PT.lem))
       (format,(flat.flat)+(txtann.txtann)+(xm.xm))
       (lang,(en.en))
  )
+ ( CST-Lem
  .    (facet,(tok pos^PT.alf+frq))
       (format,(flat+txtann+xm.flat))
       (lang,(en.en))
  )
+ ( CST-Lem
  .    (facet,(tok pos^Par.lem))
       (format,(flat.flat)+(txtann.txtann)+(xm.xm))
       (lang,(da.da))
  )
+ ( CST-Lem
  .    (facet,(tok pos^Par.alf+frq))
       (format,(flat+txtann+xm.flat))
       (lang,(da.da))
  )
```

In this notation, the `+` always indicates alternation (OR). In expressions like `(flat+txtann+xm.flat)` and `(ru.ru)` the left hand side of the dot specifies input feature(s) and the right hand side the corresponding output feature(s). In an expression like `tok pos^Par` the elements after the white space are optional. So it says that the input must contain tokens and that it optionally may contain PoS-tags. The `^` indicates that a specialization follows. In this case it says that the Parole tag set must be utilized.

Metadata creation limitations in workflows

A philosophically interesting limitation of the workflow planner is that tools that create metadata from data cannot be included in workflows, except perhaps as the last step. Examples of such tools are language guessers and format guessers. It is easy to see why such programs are problematic: suppose that the user wants to create lemmas from a text, but that the language of the text is unknown. The lemmatizer surely needs to know what language a text is written in, so we might tentatively precede the lemmatizer step with a language guesser step in the workflow. Now, the output from the language guesser is not known when the workflow is created and might be a language that the lemmatizer does not support. That means that the workflow cannot be guaranteed to succeed when executed! Worse, the parameters that must be sent to the lemmatizer may depend on the outcome of the language guesser. It is easy to see that computing workflows quickly becomes as complex as forecasting the weather if parameters cannot be set on beforehand.

Related work

The German WebLicht project serves the same needs as clarin.dk's Tools module, but the approaches are in many ways very different. Here are some differences between WebLicht and clarin.dk's Tools module.

- WebLicht is bigger in terms of involved developers, institutions, users and tools, and it has been in service for a much longer time than the Tools module in clarin.dk. Although a straight comparison between numbers of tools is misleading, it is clear that WebLicht currently can give access to the most diverse set of tools, at least for German.
- WebLicht can only handle text resources. Clarin.dk's Tools module is agnostic as to the type and format of the resources it handles.
- WebLicht does not compute workflows. Instead, the user must assemble workflows in steps. For each step, the WebLicht User Interface proposes a list of tools that, given the output from the previous step, do apply. The clarin.dk Tools module does all the assembling fully automatic and only presents the user for a list of complete workflows, from which the user can chose one.
- WebLicht's user experience is tool-oriented. Clarin.dk's Tools module is goal-oriented by default, but allows a tool-oriented experience for the daring.
- To make the use of WebLicht more convenient to the end user, there will be predefined processing chains [Hinrichs 2010]. In clarin.dk, the need for predefined processing chains is absent. Clarin.dk's Tools module already presents processing chains to the user. It has to be noted though that as the number of tools grows in clarin.dk, the need may arise to spare the user for very long lists of processing chains. On the other hand, WebLicht's envisioned list of predefined processing chains may also become unwieldy.
- Clarin.dk's Tools module is accessible for everybody. Access to the WebLicht web application is currently restricted, either by password or by affiliation.
- WebLicht can immediately handle input that is uploaded by the user. Coupling of WebLicht and an eSciDoc based repository is ongoing. The Tools module in clarin.dk only processes resources that are deposited in clarin.dk's eSciDoc repository.

There has not been much user experience with clarin.dk's Tools module yet, but as it was conceived as the logical next step after UCPH-CST's online tools website, which has been well received by teachers at UCPH and by other users all over the world, we expect that clarin.dk's user interface to integrated tools will be well received.

The closeness to the bare metal that WebLicht offers is certainly a positive aspect for expert users. Also, WebLicht's TCF (Text Corpus Format) files seem to be prettier and easier to handle for the casual user who wants

her text annotated than TEI P5 files, the text representation standard adopted by the DK-CLARIN project. On the other hand, the TEI P5 files are made for storage in a searchable repository, whereas TCF-files currently lack the needed metadata that makes search feasible. It seems logical to work towards bridging technology that combines the best of both worlds.

Currently stored resources and tools

Users are constantly uploading resources, and at the time of writing this report this is the statistics:

Resource type	Deposited resources
Text[6]	21956
Text annotation	86553
Audio	6
Video	6
Media annotation	16
Lexicon	3
Tools	5
Data	0

Table 1: The number of resources currently deposited in clarin.dk

User involvement

Throughout the paper we have described how user wishes have given directions for the requirement specification and implementation of metadata, of standard formats etc. In this section we describe the use of focus groups for additional design issues.

Focus group for design of web interface

The data provider project partners covered a broad spectrum of points of view and background experiences: Researchers from universities, from The Danish National Museum, from the areas of speech, text and multimodality, just to mention a few different types.

A focus group for the 'look and feel' and functionality of the clarin.dk web was therefore created and in a number of iterations developers presented their suggestions and got valuable feedback on functionality. This approach was chosen because it is not possible to ask users a priori what they want – but when you present them with a proposal, you can get their feedback[7].

However, time was an important challenge, and only a core part of the user wishes could be implemented in the current version. The process assured that the functionality implemented was widely motivated by the users and that the users' acquaintance with the infrastructure was increased stepwise during the implementation phase.

Focus groups for other specific design issues

In a number of meetings specific design issues such as metadata search, advanced search, viewing and delivering functionality were discussed with users with special interest in each topic, and in the same way as for the web design these meetings gave very valuable input to the developers.

Summing up on focus groups, we can state, contrary to Steve Jobs, that we value the involvement of the users, and that the dialogue and the iterative process will be continued where relevant, in the follow-up of the project.

Technical implementation

Clarin.dk uses a Service-Oriented Architecture (SOA). The implementation is based on eSciDoc (The Open Source e-Research Environment[8]) and The Fedora Commons repository system[9]. All XML-files are stored in a separate database, MarkLogic[10], which also provides xml search facilities. More details can be found in [Conrad 2010].

Future plans

There are two aspects of future plans: The organisational aspect and the technical/content aspect.

Organisationally, DK-CLARIN is now through the preparatory phase, and is ready to participate in CLARIN ERIC when it is established, hopefully by January, 2012. The participation will be through the upcoming Danish national research infrastructure for the humanities, Digital Humanities Laboratory, planned to start early 2012. This means that currently we are in an interim period with limited resources.

On the technical and/or content side we are improving the functionality, first of all focussing on inadequate aspects of the current implementation. A list has been made with priority additions to the current system. These include more user-friendly metadata search facility and better user guidance including a number of use cases illustrating different researchers' use of clarin.dk. Especially users mentioned that ways to explore tools calls for both for simple and more advanced use cases. For texts and text annotations focus is on extending the combined metadata and content search for texts and extending of the viewing possibilities for text annotation resources. Extensions that explore linking of resources are also considered in the scope of dictionaries and texts.

As users' needs and ideas develop, keeping clarin.dk a valuable infrastructure for the humanities will demand for continuous support and extension work and will be carried out in the scope of the upcoming Digital Humanities Laboratory.

[6] The Danish National Museum has deposited almost 7000 resources, consisting of originally scanned images of archive cards, which contain text and a photo. These are categorized as text.

[7] This is actually another way of interpreting the word 'unaskable' in the title of this conference.

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

[9] http://www.fedora-commons.org/software

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

Acknowledgements

As mentioned above, the DK-CLARIN project was performed by a consortium with the following members:

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

We want to thank all of the consortium members for their contribution, not least The Royal Library and the Society for Danish Language and Literature who were also part of the implementation team.

This project was supported by the Danish Agency for Science, Technology and Innovation, as well as by all partner institutions.

References

Asmussen, J (2011) Text metadata: What the header of a text item looks like, DK-CLARINWP 2.1 Technical Report, http://korpus.dsl.dk/clarin/corpus-doc/text-header.pdf

Asmussen, J. & Halskov, J. (2009) Compiling and annotating corpora in DK-CLARIN. Interpreting and tweaking TEI P5. In CATHERINE SMITH MICHAELA MAHLBERG, VICTORINA GON - , editor: Proceedings of the Corpus Linguistics Conference CL2009. University of Liverpool, UK 2009 〈URL: http://ucrel.lancs.ac.uk/publications/cl2009/〉

Braasch, A. & B.S. Pedersen (2010). Encoding Attitude and Connotation in Wordnets, In: The 14th EURALEX International Congress, Leeuwarden , The Netherlands.

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

Christiansen, T.U. & Henrichsen, P.J. (2011) "Objective Evaluation of Consonant-Vowel pairs produced by Native Speakers of Danish", *Forum Acousticum* 2011

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. Braasch, A. Haltrup Hansen, D. & Olsen, S. (2010) Quality indicators of LSP texts - selection and measurements. Measuring the terminological usefulness of documents for an LSP corpus, Proceedings of LREC 2010, pp. 2614--2620. Malta

Henrichsen, P.J. (2008) "One for all and all for one - Recycling Scandinavian phonetics", in E. Ahlsen et al (eds) "Communication - Action - Meaning, a festschrift to Jens Allwood", Göteborg University Press, pp.191--205

Henrichsen, P.J. & Christiansen, T.U. (2011) "Information-based Speech Transduction*";* ISAAR-2011 (International Symposium on Auditory and Audiological Research)*;* 10pp

Henrichsen, P.J. (2011) "Fishing in a Speech Stream, Angling for a Lexicon*"; NODALIDA*, Riga

Henrichsen, P.J. (2010) "Ethical Intelligence in Social Recommender Systems", SRS-2010 (at IUI-2010, International Conference on Intelligent User Interfaces); Hong Kong; 4pp

Henrichsen, P.J. & T.U. Christiansen (2009) "Fishing for Meaningful Units in Connected Speech"; ISAAR-2009 (International Symposium on Auditory and Audiological Research); 10p

Henrichsen, P.J. (2008) "The CBS Text-to-Speech Workbench", CBS Theoretical Papers on Linguistics, CBS University Press, 26pp

Hinrichs, E.W., Hinrichs, M. & Zastrow, T. (2010). WebLicht: Web-Based LRT Services for German. In Proceedings of ACL (System Demonstrations), pp. 25--29.

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

Johannsen, A. & Pedersen, B.S. (2011). "Andre ord" — a wordnet browser for the Danish wordnet, DanNet . In: Proceedings from 18th Nordic Conference of Computational Linguistics, NODALIDA 2011, Riga, Latvia. Nothern Association for Language Technology, Vol. 11 pp. 295--298, University of Tartu.

Jokinen, K. Navarretta, C. & Paggio, P. (2008) Distinguishing the communicative functions of gestures. In A. Popescu-Belis and R. Stiefelhagen (eds.) Proceedings of 5th Joint Workshop on Machine Learning and Multimodal Interaction, Utrecht, September 2008, Springer, pp. 38--49.

Navarretta, C. Annotating Non-verbal Behaviours in Informal Interactions. To appear in A. Esposito, A. Vinciarelli, K. Vicsi, C. Pelachaud and A. Nijholt (Eds.) Analysis of Verbal and Nonverbal Communication and Enactment: The Processing Issues, LNCS 6800, Springer Verlag, pp. 317--324.

Navarretta, C. (2011) Anaphora and gestures in multimodal communication. To appear in "*Proceedings of the 8th* Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2011)", Faro Portugal, 6-7 October 2011 (11 pages)

Pedersen, B.S., Nimb, S. & Braasch, A.(2010). Merging specialist taxonomies and folk taxonomies in wordnets. - a case study of plants, animals and foods in the Danish wordnet. In: Proceedings from the Seventh International Conference on Language Resources and Evaluation, pp. 3181--3186. Malta.

Ruus, H. & Duncker, D. (2011)"Corpus-based variation studies – A methodology", in Language Variation – European Perspectives III, Selected papers from the 5th International Conference on Language Variation in Europe (ICLaVE 5*)*, Ed. by Gregersen, Frans, Parrott, Jeffrey K, Quist, Pia. John Benjamins, pp. 161--172.

Uneson, M.; P.J. Henrichsen (2011) "Expanding a Corpus of Closed-World Descriptions by Semantic Unit Selection", *CLA-11* (International Conference on Computational Linguistics Application), Jachranka