Why knowledge modeling is important for business and for a Danish terminology and knowledge bank

Bodil Nistrup Madsen
DANTERMcentre
Copenhagen Business School, Denmark
bnm.danterm@cbs.dk

Abstract

Businesses and organizations, including public authorities, have a growing need for organizing and handling large amounts of data. In order to manage complex knowledge, knowledge must be modeled and structured. One very powerful method used for structuring knowledge is the use of ontologies. Businesses and organizations need efficient tools for building domain-specific ontologies and systems for managing knowledge. Most medium-sized and large Danish businesses and organizations operate in a bilingual or multilingual environment, where knowledge is transferred and stored in Danish and/or English, and possibly other languages, so information must be freely retrievable and communicable in several languages. In this paper I will give examples of why knowledge modeling is important for businesses, and why knowledge modeling is a central part of the DanTermBank project, the aim of which is to lay the foundation for a national terminology and knowledge bank. Furthermore, I will briefly introduce our plans for teaching within the field of knowledge modeling at Copenhagen Business School, CBS.

Introduction

In order to manage vast quantities of digitized information, businesses and public authorities need to structure and manage their knowledge so that the right persons have access to the right information at the right time. This is posing an ever-increasing challenge in an age where the flow and complexity of information are growing rapidly, and where the need for quick, efficient and reliable communication is essential. The challenge can only be met if we are capable of structuring knowledge appropriately and if we can navigate freely in this knowledge.

In this paper, I will first give examples of activities in businesses and organizations where knowledge modeling is needed and examples of research which may contribute to the fulfillment of these needs. Then, I will introduce the concept of terminological ontologies and give examples of ontology work already carried out in the Danish authorities as well as examples of relevant education initiatives.

Finally, I will give an introduction to the plans and ongoing work in the DanTermBank project, which runs from 2011-2014 at the Department of International Language Studies and Computational Linguistics, ISV, CBS. The project is supported by the VELUX Foundation, www.veluxfonden.dk.

If a terminology and knowledge bank does not contain a sufficient number of terms, users will not feel encouraged to use it, and on the other hand, users will be frustrated if it contains a large amount of terms with only little or poor quality information. Therefore, it is necessary to use automatic procedures in order to extract and systematize information about terms. In this project, we will develop methods for automatic knowledge extraction and automatic construction and updating of ontologies. The project also aims at developing methods for automatic merging of terminological data from various existing sources as well as methods for target group-oriented knowledge dissemination. Terminological ontologies will play an important role in all the subprojects (see below).

Business activities

Figure 1 presents keywords related to activities of businesses and public authorities.

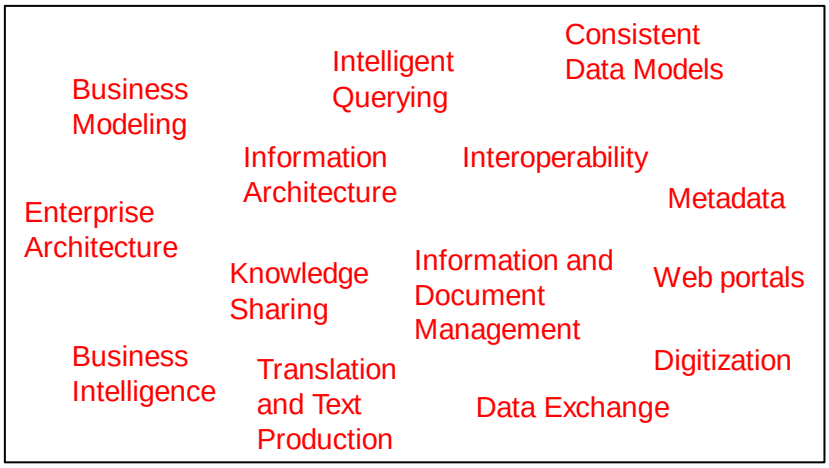

Figure 1. Examples of issues that are relevant for businesses and public authorities

- Descriptions of core aspects of a business: business modeling (Enterprise Architecture and Business Intelligence)
- IT-related issues: information architecture, consistent data models, interoperability of knowledge resources and data exchange
- Knowledge dissemination: digitization (web portals, metadata)
- Activities which support business processes: intelligent querying, knowledge sharing, information and document management, translation and text production.

As a basis for the successful conduct of central business activities, there is a need for concept clarification and knowledge modeling.

Humanistic research

Figure 2 presents examples related to humanistic research which can support business activities.

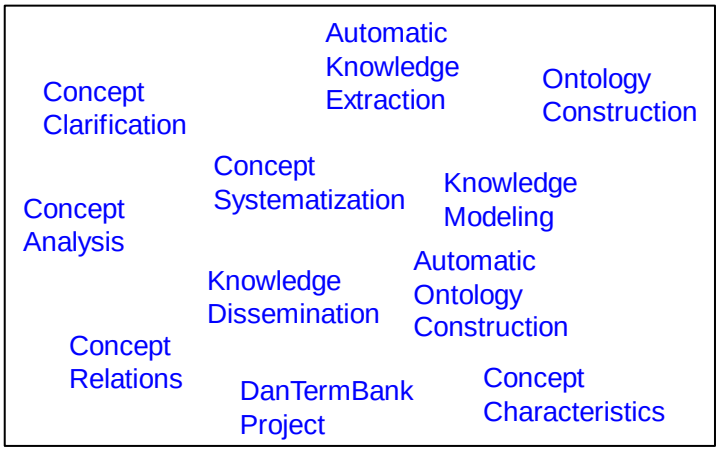

Figure 2. Examples of humanistic research which can support business activities

- Core research issues: concept clarification, concept analysis (concept relations and concept characteristics)
- Methods to organize knowledge: knowledge modeling, concept systematization, ontology construction
- Research issues of the DanTermBank project: automatic knowledge extraction, automatic ontology construction, knowledge dissemination.

What is an ontology?

Figure 3 presents a simplified example of a terminological ontology for taxes.

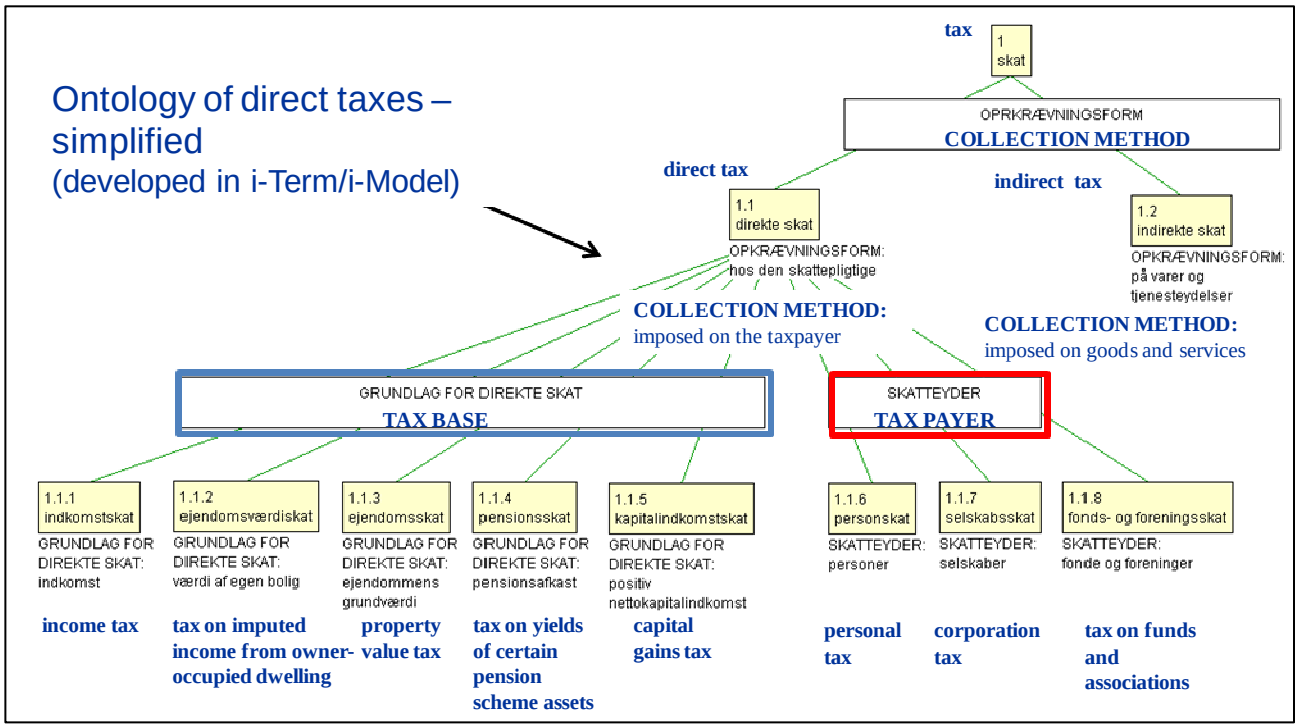

Figure 3. Example of a terminological ontology for Danish tax collection

The ontology in this figure has been created using the ontology modeling module i-Model of the terminology and knowledge management system i-Term ®, developed by the DANTERM Centre at CBS, www.i-Term.dk.

As an introduction, the elements of a terminological ontology will be briefly explained.

The ontology reflects the Danish taxation system. English glosses are added, but it should be emphasized that if we develop an ontology to reflect the English taxation system, it will differ from the Danish one. It is our intention to develop both a Danish and an English ontology In the DanTermBank project, and these will form the basis for clarification of equivalence (or non-equivalence) between the two languages, that will be reflected in the entries of the terminology and knowledge bank.

Concepts are represented by means of yellow boxes and type relations by means of green lines. The concepts *direkte skat* (direct tax) and *indirekte skat* (indirect tax) differ with respect to OPKRÆVNINGSFORM (COLLECTION METHOD). This is what we call a subdivision criterion. Subdivision criteria are shown in white boxes. Subdivision criteria group concepts according to the characteristics which are given in the form of feature specifications below the concepts – see *direkte skat* (direct tax) and *indirekte skat* (indirect tax): *OPKRÆVNINGSFORM: hos den skattepligtige* and *OPKRÆVNINGSFORM: på varer og tjenesteydelser* (COLLECTION METHOD: imposed on the taxpayer and COLLECTION METHOD: imposed on goods and services).

The subdivision criteria help the user to understand the meaning of the concepts, give a good overview and help the terminologist in writing consistent definitions. The definition of a concept is given by means of the position in the ontology and the characteristics. Definitions may be created on the basis of this information.

In Figure 4, the ontology from Figure 3 is compared with an extract of a classification system from a catalogue of public services, the class *Taxes on income and wealth* at the bottom of Figure 4.

The classification system is not based on the ontology. Some types of taxes are missing and the order of taxes in the classification system is not intuitive compared to the ontology: the classes corresponding to the sub-concepts in the ontology are not grouped according to the two aspects (subdivision criteria): *GRUNDLAG FOR DIREKTE SKAT* (TAX BASE) and *SKATTEYDER* (TAX PAYER). It is not clear whether *property taxation* covers both *ejendomsværdiskat* (tax on imputed income from owner-occupied dwelling) and *ejendomsskat* (property value tax), and *indkomstskat* (income tax) is missing.

Figure 5 shows how the ontology of taxes may grow if all types of tax are added, Nielsen (2009).

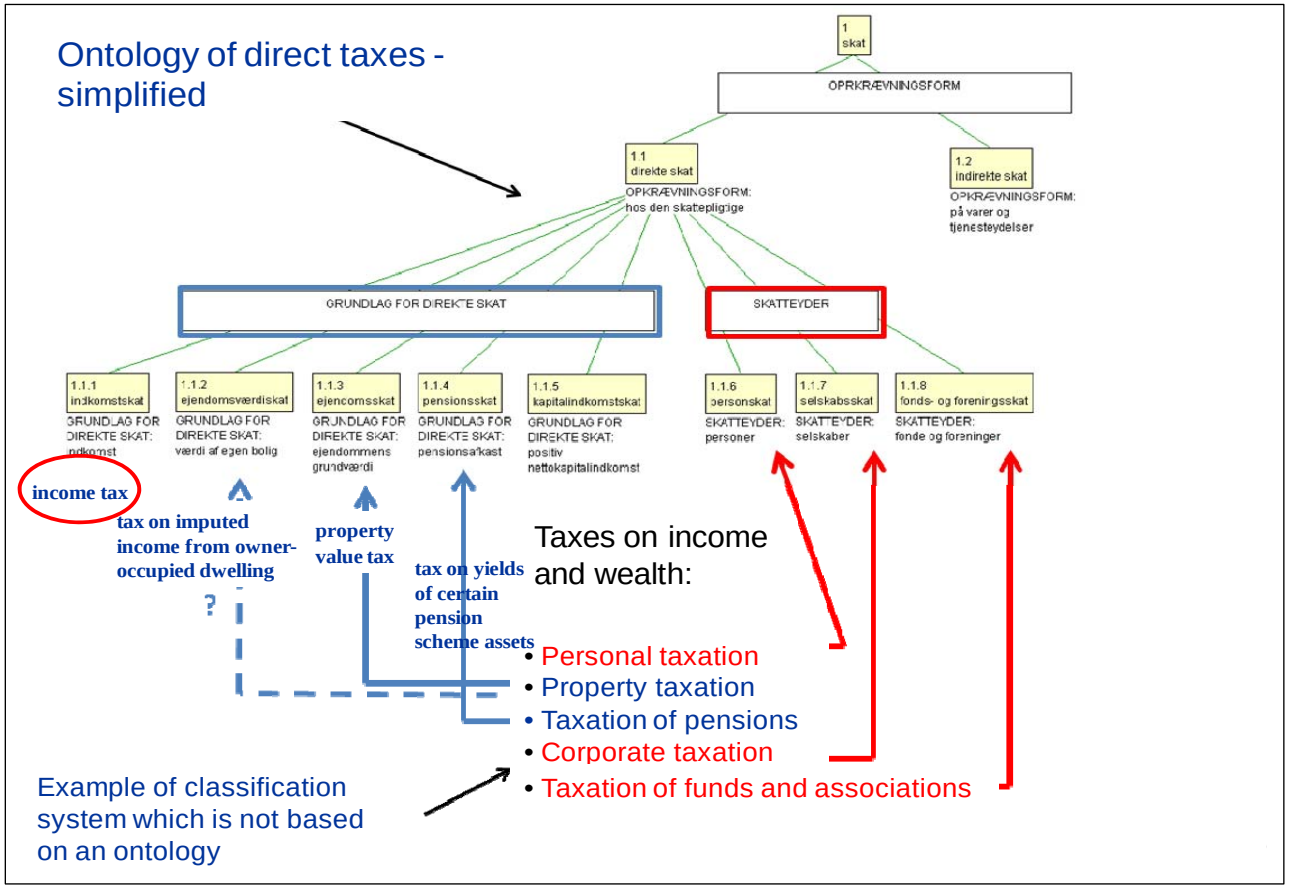

Figure 4. Comparison of the terminological ontology with an example of a classification system

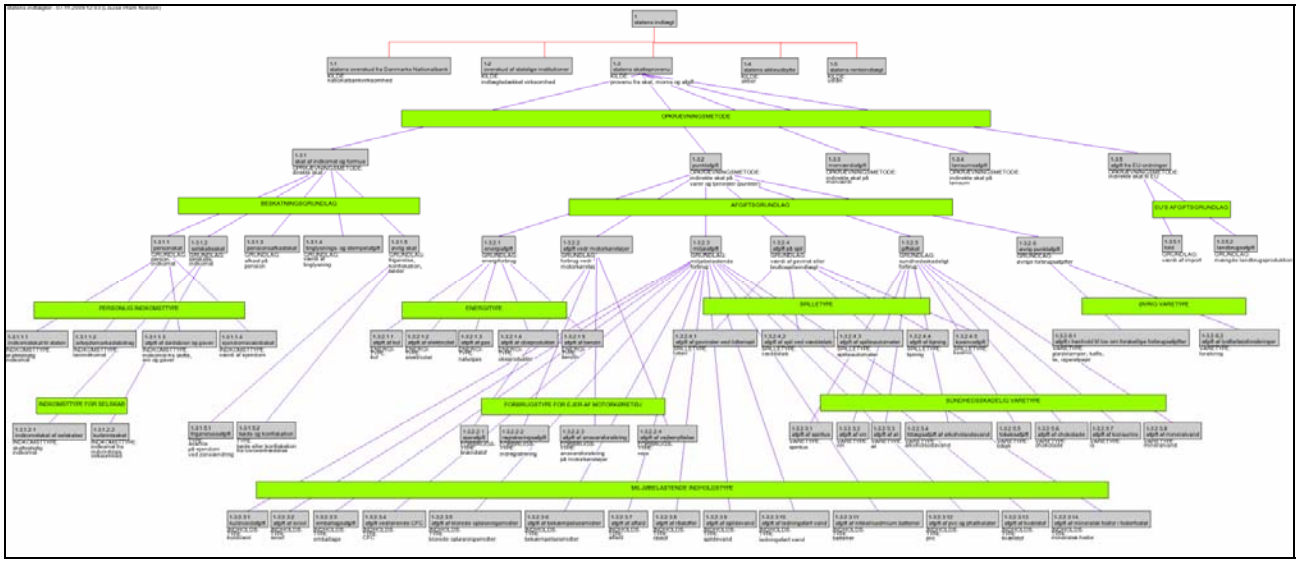

Figure 5. Example 'Statens skatteprovenu'

In the next section, I will go a little more into detail on the differences between ontologies and other knowledge structuring tools.

Ontologies vs. data models and classification systems

Figure 6 presents a small ontology illustrating the differences between ontologies, data models and classification systems (including taxonomies).

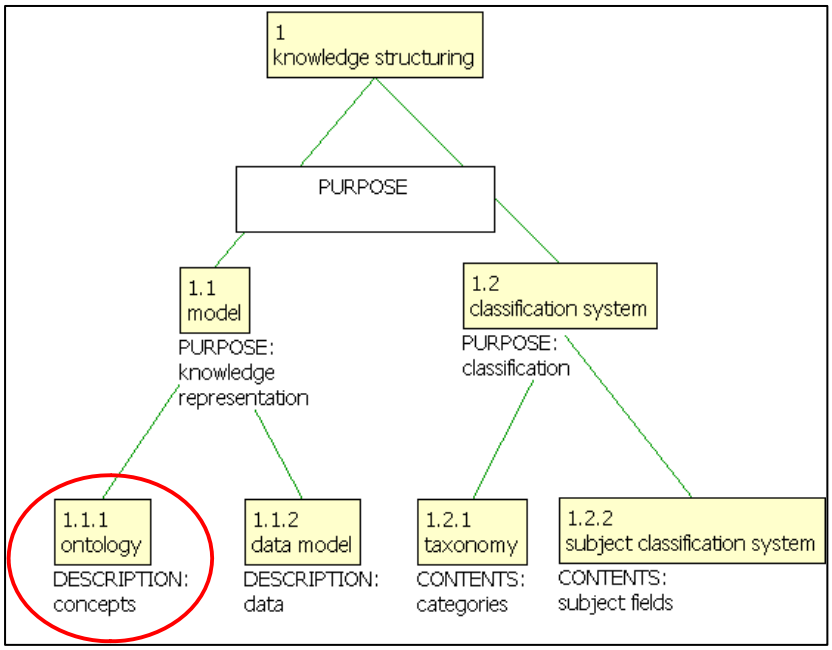

Figure 6. Ontology of central knowledge structuring concepts

On the basis of the characteristics, we suggest the following definitions of these concepts:

model:
simplified representation of knowledge about phenomena

ontology:
model for the description of knowledge about concepts

data model:
formal model for the description of types of data in an IT system

classification system:
system for the division of phenomena into classes

taxonomy:
classification system for the division of categories of a domain

Differences between ontologies and classification systems are also described in Madsen & Thomsen (2009). Madsen & Odgaard (2010) describe differences between ontologies and data models.

Examples of the use of ontologies

Here I will give some examples of the use of ontologies in businesses and organizations, cf. also the examples in Figure 1.

- Development of IT Architecture
- Ontologies as a basis for data models and exchange formats
- Ontologies as a basis for metadata taxonomies and meta models
- Intelligent, ontology-based querying and document management systems
- Software for semantic text control
- Knowledge structuring and knowledge sharing in companies and organisations
- Ontology-based translation systems.

The following figures visualize some of the above mentioned uses of ontologies.

Figure 7 illustrates a well-known situation: Customer and Developer do not speak the same 'language', and probably they do not have the same understanding of concepts that are relevant for development of an IT architecture or an IT system.

Concept clarification by means of knowledge modeling, i.e. formalization of the meaning of concepts, furthers understanding and success of the development task, cf. Figure 8.

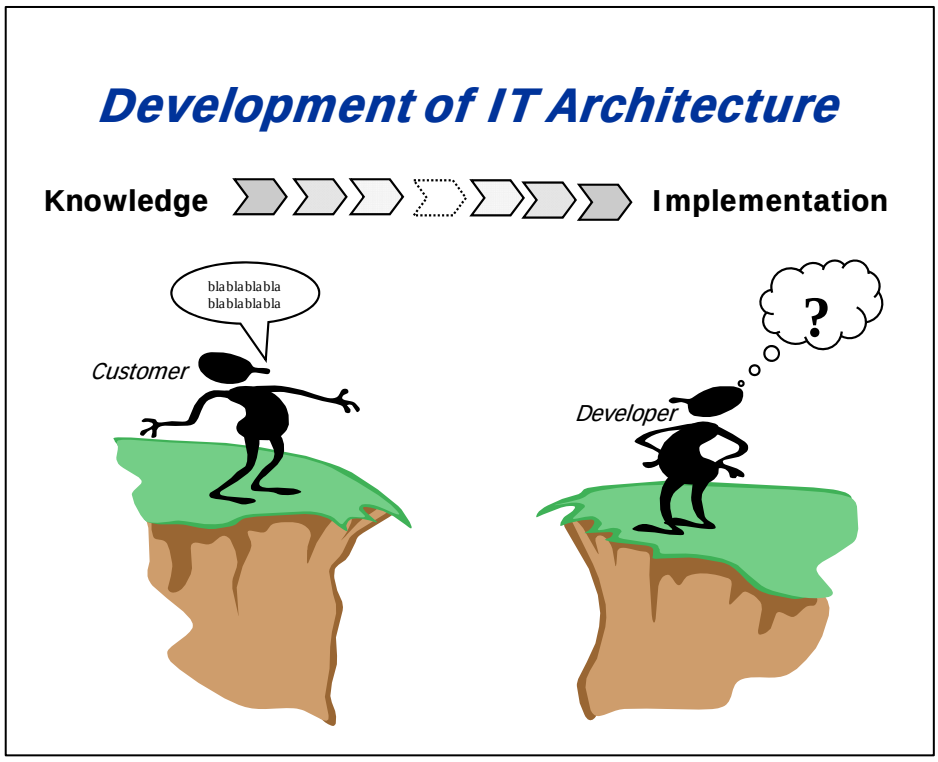

Figure 7. Knowledge gap in the development of IT architecture

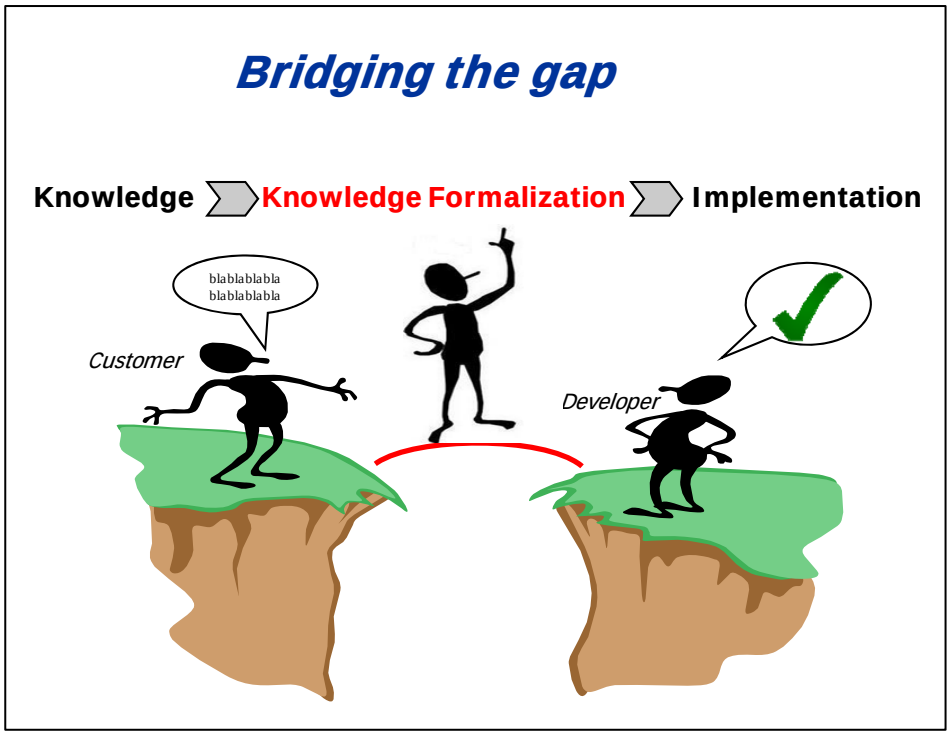

Figure 8. Bridging the gap by means of knowledge clarification

As seen from Figure 6, ontologies and data models have different aims:

- ontologies aim at concept clarification, mutual understanding of concepts and consistent use of terms
- data models aim at specifying the types of information of an IT system and their mutual relationships

Ontologies form a solid basis for the development of data models. Even domain specialists in the same organization may have different understandings of concepts, and in order to avoid errors and inconsistencies, concept clarification by means of ontologies should form the first step in the data modeling process, cf. Figure 9.

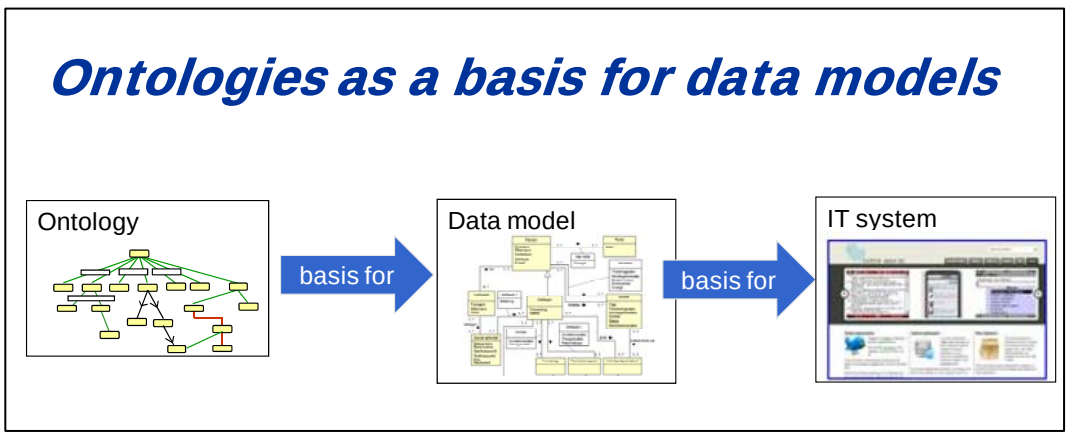

Figure 9. Ontologies as a basis for data models

When The Danish Prison & Probation Service (Kriminalforsorgen) started working on a new IT system, it was clear that the people involved did not speak the same language – they had different understandings of central concepts. Therefore, the Service initiated a cooperation with the DANTERM Centre for concept clarification by means of ontologies.

The problems increase when domain experts from different organisations are involved.

This was obvious in the cooperation between The Danish Prison & Probation Service and other institutions of the Ministry of Justice. The four institutions are going to exchange data, and therefore it is very important that they have the same understanding of the data they are exchanging, cf. Figure 10.

Cooperation between the DANTERM Centre and authorities under the Ministry of Justice:

- The Danish Prison & Probation Service
- The Danish Police
- The Prosecution Service in Denmark
- The Danish Court Administration.

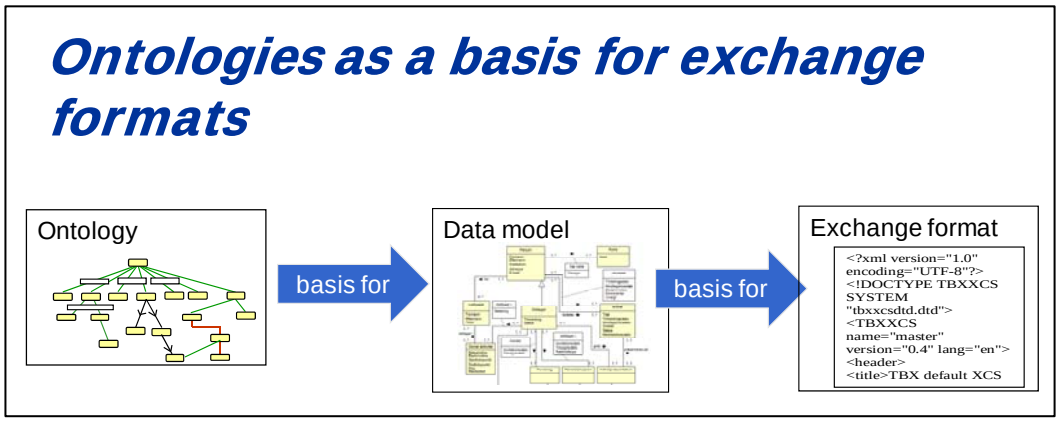

Figure 10. Ontologies as a basis for exchange formats

Also in this case, ontologies contribute to concept clarification and consensus on definitions of concepts. See also 'Modeling of concepts in the Ministry of Justice' below in this volume (Hoffmann, 2011).

The DANTERM Centre has also worked together with the former IT and Telecom Agency (IT- og Telestyrelsen) in order to set up principles for developing metadata taxonomies and to describe components of a meta model for Enterprise Architecture, cf. Figure 11.

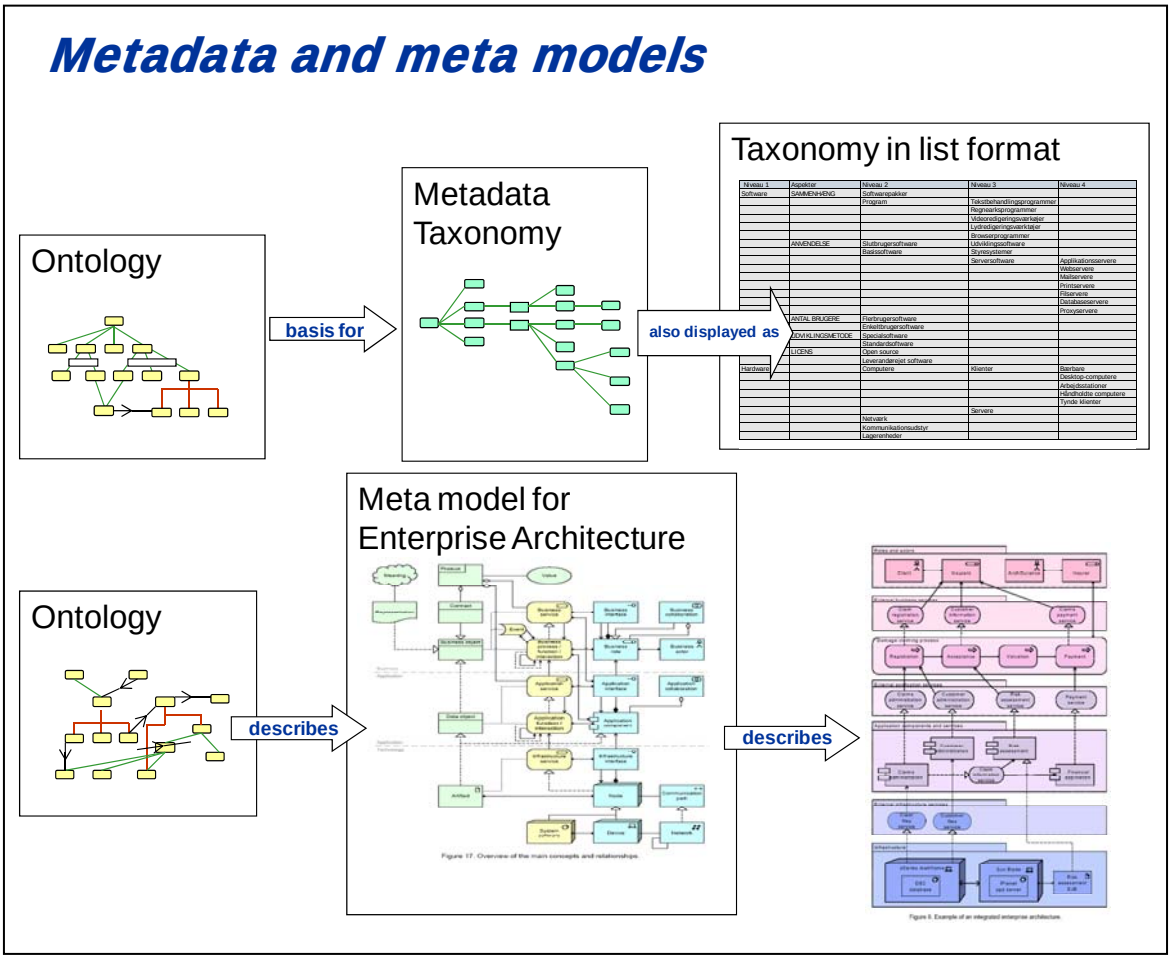

Figure 11. Ontologies as a basis for metadata taxonomies and meta models

Ontology work in Denmark

Many authorities develop ontologies as a basis for their digitization projects. Below some examples of public authorities who have been working or are working on the development of ontologies:

- The National Board of Health (Sundhedsstyrelsen)
  - Concept clarification for example as a basis for *electronic Health Care Records*
- The Capital Region of Denmark (Region Hovedstaden)
  - Concept clarification as a basis for *user interfaces in health care*
- The National Board of Social Services (Servicestyrelsen)
  - *Concept clarification* as a basis for putting social and welfare initiatives into practice

- The Digital Task Force (now under Økonomistyrelsen)
    - Concept clarification as a basis for a *catalogue of public services*
- The (former) IT and Telecom Agency (IT- og Telestyrelsen)
    - Ontologies and classifications (*methods and tools for IT architecture and enterprise architecture*)
- The Danish Prison & Probation Service (Kriminalforsorgen)
    - Concept clarification as a basis for *exchange formats and IT systems*

Huge potential for industrial partners

The potential for industrial partners is also huge. In particular, large businesses and organisations which face challenges due to the complexity of information are likely external partners. Partners in research projects are for example *Novo Nordisk*, *Novozymes*, *IBM* and *Ankiro*.

Most medium-sized and large Danish businesses and organizations have a need for easy access to structured knowledge in several languages. However, knowledge modeling is very time-consuming. Therefore, they need knowledge about and efficient tools for building domain-specific ontologies and systems for managing knowledge as well as a national terminology and knowledge bank, where they can find structured knowledge in Danish and other languages.

Education

The DANTERM Centre gives courses in concept clarification, terminology methods, and knowledge modeling. For example, in 2009: a knowledge modeling course of 5 full days for KL (Local Government Denmark), ITST (Danish IT and Telecom Agency) and KOMBIT (IT cooperation of the Danish municipalities). This course included: principles and tools for ontology work, ontology-based classification systems, thesauri, metadata taxonomies and data models.

It is the vision of the Department of International Language Studies and Computational Linguistics, to develop an executive master programme on Information and IT architecture in cooperation with other CBS departments. One central part of this programme would be knowledge modeling.

Electives will also be offered in other study programmes at CBS.

Need for automatic methods

Ontology construction is very time consuming, and therefore there is a need for automatic ontology construction based on various kinds of input: internet text and text files but also structured data, taxonomies, etc., cf. Figure 12.

On the basis of automatic extraction of information about concepts (concept relations and characteristics) it will be possible to produce draft ontologies, which may be validated against a set of rules for terminological ontologies in an automatic validation module.

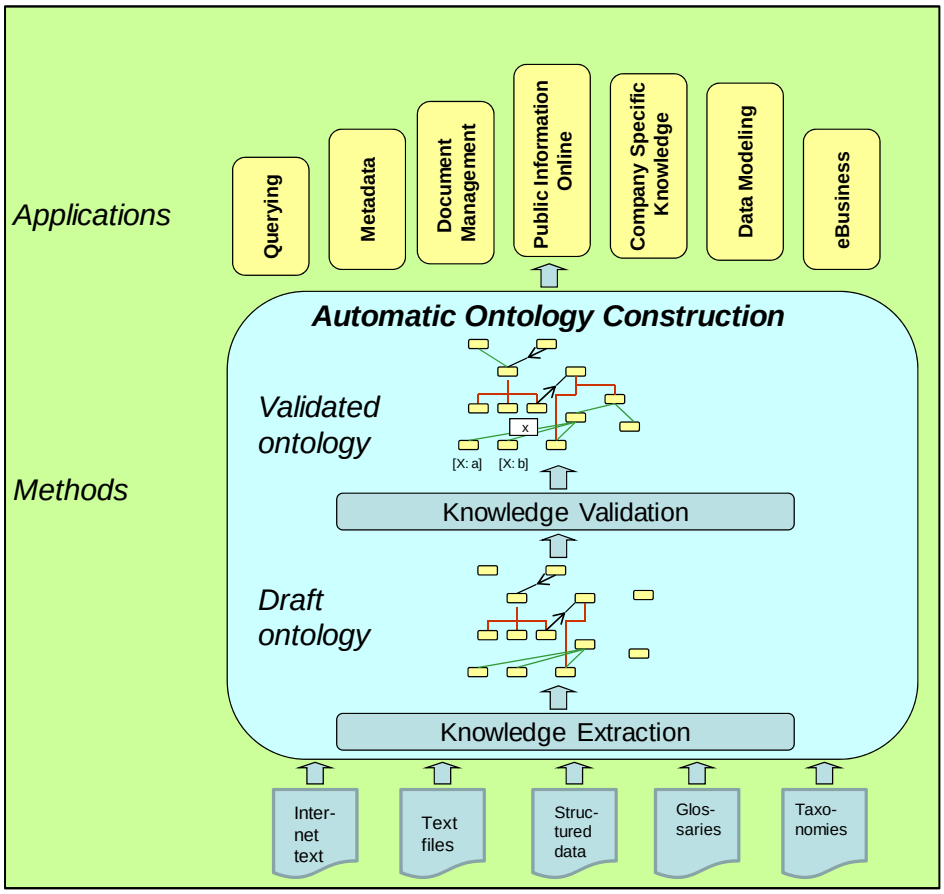

Figure 12. Automatic ontology construction – methods and applications

The DanTermBank Project

This leads us to the DanTermBank project in which we will develop automatic methods for extracting and structuring knowledge for a Danish Terminology and Knowledge Bank.

This Research project aims at establishing the foundation for a National Terminology and Knowledge Bank. As already mentioned, the project is funded by the VELUX Foundation for the years 2011-2014.

The project team comprises:

- Bodil Nistrup Madsen, Hanne Erdman Thomsen (senior researchers)
- Tine Lassen, Louise Pram Nielsen (Post.doc. and PhD-fellow)
- Anna Odgaard, Pia Hoffmann (project managers, the DANTERM Centre)

Background of the team members: terminology, computational linguistics, economy. The project team also includes a system developer.

Background

For many years, CBS has educated translators, and this is the reason why

- Concept clarification and terminology
- Terminology databases and knowledge dissemination
- Computer-aided ontology construction

play an important role in teaching and research.

Already back in the late 1970ies, we had an ambition to establish a term bank, but it was not possible to obtain political and financial support at that time.

In a research project with funding from the Danish Research Council for the Humanities from 1998-2001 we worked on

- Principles and methods for company specific terminology management systems and
- Principles and methods for computer-aided ontology structuring (CAOS).

As a result of this project, the DANTERM Centre started developing the terminology and knowledge management system i-Term, which is used by universities, companies, organizations and public authorities in Denmark and Europe. Ontology modeling in i-Model is based on user input, and has no consistency checking facilities.

Also, research on computer-aided ontology structuring was continued in the CAOS project, and a prototype was developed. This prototype is also based on user input, but it has some facilities for consistency checking and it warns the user, if concepts are placed incorrectly according to their characteristics. In the CAOS project we have formalized principles that have been used for many years in terminology work.

The principles used in CAOS are described in Madsen, Thomsen & Vikner (2004 and 2005). More information on the facilities of i-Model and CAOS may be found in Madsen (2006) and in the article below.

In 2008, the Language Committee of the Danish Government, published the report 'Sprog til tiden' ('Language on demand'), in which the importance of a freely accessible national terminology and knowledge bank was emphasized:

> *A prerequisite for continuous use and development of a national LSP in small countries like for example Denmark is free access to a terminology and knowledge bank comprising domain knowledge in Danish and foreign languages.*

Domain specific knowledge goes beyond traditional dictionary information. In order to clarify and distinguish the meanings of domain specific concepts, these must be described by means of characteristics and relations to other concepts, i.e. in the form of domain specific ontologies. On the basis of these, it is possible to develop consistent definitions that further understanding and the correct use of terms. As already mentioned, terminology work that includes development of ontologies is a very labour-intensive task, and therefore most companies cannot afford this kind of work.

So, when in September 2009 the VELUX Foundation invited researchers from CBS to apply

for funding, we concluded that this was our chance to combine results of previous research and development, and to develop innovative and advanced methods for dynamic and automatic extraction of knowledge about concepts from texts, for automatic construction of ontologies and for target group-oriented knowledge dissemination.

Problems in existing term banks

It is our experience that national term banks suffer from several problems:

1. Insufficient number of terms
2. Little or poor quality of information
3. Too many hits when looking up a term
4. Different needs of user groups not considered

Solutions to these problems may be:

1. Automatic extraction of terms and information about terms from texts
2. Automatic construction of ontologies
3. Use of ontologies for merging terminological entries from different sources
4. Target user-oriented knowledge dissemination

Re 1: If a term bank does not contain a sufficient number of terms, users will not feel encouraged to use it.

A solution to this problem is automatic extraction of terms and information about terms from texts.

Re 2: On the other hand, users will be frustrated if a term bank contains a large amount of terms with only little or poor quality information. Ontologies further understanding and the correct use of terms. Terminology work that includes development of ontologies is, however, a very labor-intensive task, and therefore most term banks do not include ontologies.

A solution to this problem is to use automatic procedures in order to develop ontologies.

Re 3: One way of increasing the amount of terms in a term bank is, as mentioned, to extract terms and information about terms automatically from texts.

Another method is to import terminology from different sources, such as other term banks or existing term lists. However, this approach will often lead to problems, since the term bank will then typically contain many entries comprising the same term, but with varying formulation of the definitions and/or different translations. Therefore, users will often get many hits when looking up a term, and they will find it difficult or impossible to choose the right entry.

A solution to this problem may be to use ontologies in the process of merging data from existing sources.

Re 4: Our project addresses the problems of the different needs of different user groups. We will use eye-tracking as one method for analyzing user needs.

In the next section, Figure 13, I present an overview of the DanTermBank project.

Overview of the DanTermBank project

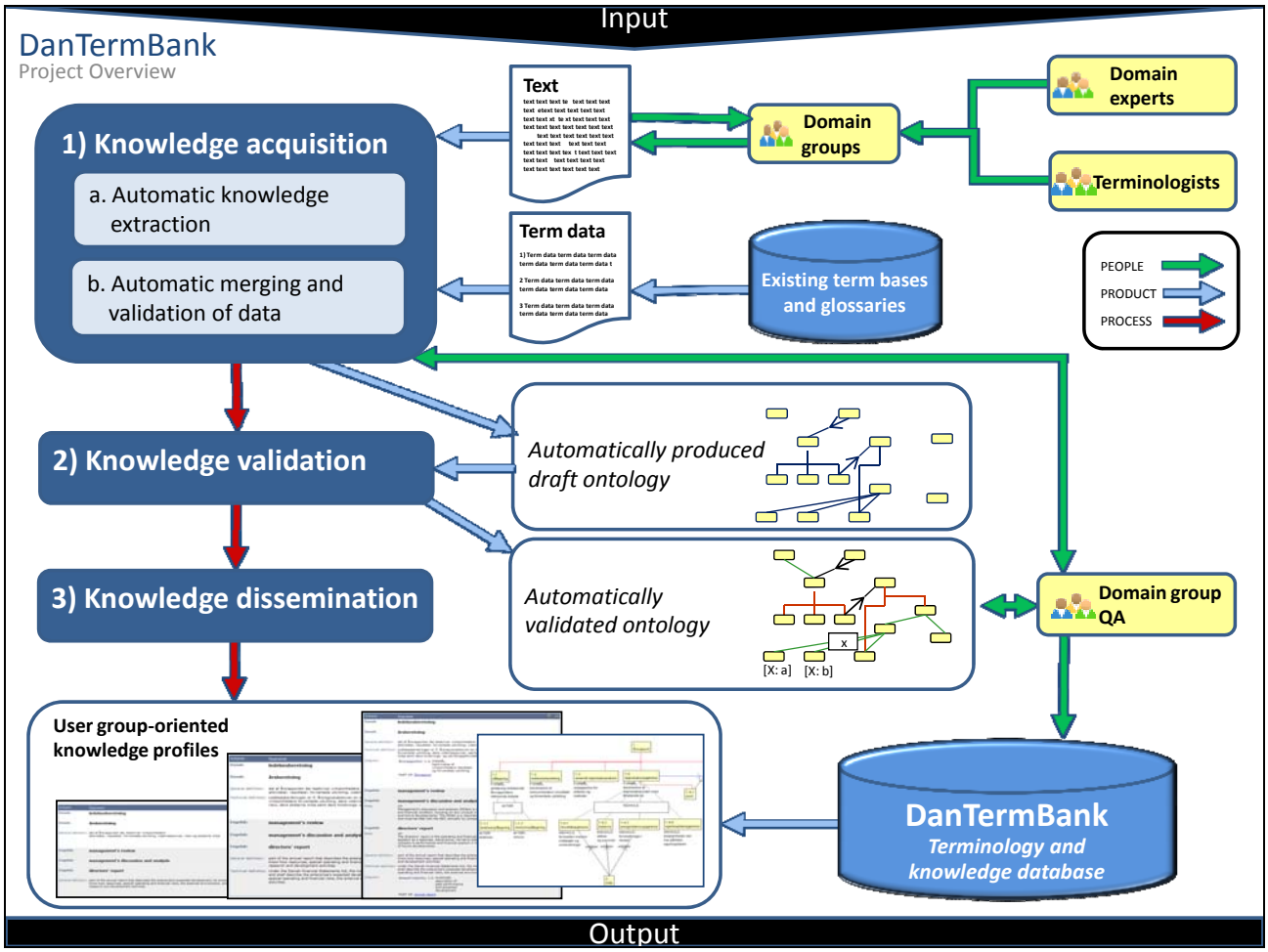

Figure 13: Overview of the DanTermBank project

In this project, we will develop ontologies and definitions within selected pilot domains, and this will result in terminological data within two important economic domains: taxes and auditing. Domain experts and users will be involved in various stages of the project.

The project consists of 3 subprojects, which I will present briefly below. More information may be found in Lassen, Madsen & Thomsen (2011) and Madsen, Thomsen, Halskov & Lassen (2010).

Knowledge acquisition

Subproject 1) a: Automatic knowledge extraction

The aim of this subproject is to develop new models of and methods for automatic extraction of concepts and information about concepts as well as a prototype which can automatically produce a draft version of a terminological ontology on the basis of an existing domain-specific text corpus, or on the basis of domain texts automatically collected from the Internet. Thus, the draft ontologies will contain subdivision criteria and characteristics as formal feature specifications on concepts.

Subproject 1) b: Automatic merging and quality assurance of data
The idea is that the contents of the terminology and knowledge bank should also be extended by means of import of the many term lists from the internet or from authorities, organisations and companies. By means of automatic ontology construction on the basis of the imported data, e.g. definitions, we aim at assuring the quality of the imported data.

Subproject 2): Knowledge validation
The aim is to develop methods and a prototype that may be used for automatic validation and dynamic expansion of the draft ontologies that result from the automatic knowledge extraction.

The Knowledge validation facilities will comprise:
- Automatic positioning of concepts in ontologies
- Automatic consistency checking
- Automatic changes to ontologies
- Dynamic updating of ontologies

Subproject 3): Knowledge dissemination
This subproject will focus on presentation of data in the term bank. Traditionally, terminology and lexicography have been separate research fields with different approaches to compilation and presentation of data. However, modern technology offers unlimited opportunities to meet the needs for several target groups in one database by offering the possibility of choosing between different presentations.

One important user group is businesses and organizations. Another important user group is schools and universities. Since English is often used in business communication and teaching, there is a need for at terminology and knowledge bank comprising structured knowledge about terms in both Danish and English. A third user group is foreigners, who, for example, need to find information in the web portals of the Danish authorities. Within these main user groups, however, a more detailed user group classification is needed.

The project will aim at developing prototypes for dynamic target-user oriented interfaces. Here, we will use eye-tracking methods for investigating the behavior of users while they use the terminology and knowledge bank for various purposes.

Conclusion
For a long period, many resources have been allocated to general language dictionaries, lexical databases and word nets in Denmark. There is, however, a big need for domain specific knowledge within scientific, technical, economical and legal domains which can be made accessible by means of a Danish terminology and knowledge bank.

In this paper, I have described the problems common to many existing term banks: insufficient quantity and quality of terms and information about terms, lack of dynamic, user-oriented information profiles.

In the DanTermBank project we will develop facilities for *automatic* methods for knowledge acquisition, knowledge structuring and validation and dynamic target-user oriented interfaces. To our knowledge, no other national term banks have developed similar facilities.

Acknowledgements
We thank the VELUX foundation for the financial support of the DanTermBank project.

Also thank you to my colleagues from the DanTermBank project team, Hanne Erdman Thomsen, Tine Lassen, Louise Pram Nielsen, Anna Odgaard and Pia Hoffmann, for input (cf. References), and to the head of Department Alex Klinge, ISV, for inspiration.

References
Hoffmann, Pia (2011). 'Modeling of concepts in the Ministry of Justice', this volume.

Lassen, Tine; Madsen, Bodil Nistrup; Thomsen, Hanne Erdman (2011). 'Automatic Knowledge Extraction and Knowledge Structuring for a National Term Bank'. In: *NODALIDA 2011 workshop: Creation, Harmonization and Application of Terminology Resources.* NEALT Proceedings Series Vol. 12. http://hdl.handle.net/10062/17274 (2011-05-09)

Madsen, Bodil Nistrup (2006).http://research.cbs.dk/research/terminological ontologies(120631) In: *Semantic Systems: From Visions to Applications*. Proceedings of the Semantics 2006, p. 271-282.

Madsen, Bodil Nistrup & Anna Elisabeth Odgaard (2010). 'From Concept Models to Conceptual Data Models'. In: Úna Bhreathnach & Fionnuala de Barra Cusack(eds.): *TKE 2010: Presenting Terminology and Knowledge Engineering Resources Online: Models and Challenges*, August 14 2010, Dublin.

Madsen, Bodil Nistrup & Hanne Erdman Thomsen (2009). 'Ontologies vs. classification systems'. In: *Proceedings of the NODALIDA 2009 workshop WordNets and other Lexical Semantic Resources — between Lexical Semantics, Lexicography, Terminology and Formal Ontologies*. Editors: Bolette Sandford Pedersen, Anna Braasch, Sanni Nimb and Ruth Vatvedt Fjeld. NEALT Proceedings Series, Vol. 7 (2009), 27-32.

Madsen, Bodil Nistrup, Hanne Erdman Thomsen, Jakob Halskov & Tine Lassen (2010). 'Automatic Ontology Construction for a National Term Bank', In: Úna Bhreathnach & Fionnuala de Barra Cusack(eds.): *TKE 2010: Presenting Terminology and Knowledge Engineering Resources Online: Models and Challenges*, August 14 2010, Dublin.

Madsen, Bodil Nistrup, Hanne Erdman Thomsen & Carl Vikner (2004). 'Principles of a system for terminological concept modelling'. *Proceedings of the 4th International Conference on Language Resources and Evaluation, Vol. I*, pp. 15-18. Lisbon.

Madsen, Bodil Nistrup, Hanne Erdman Thomsen & Carl Vikner (2005).'Multidimensionality in terminological concept modeling'. In: Bodil Nistrup Madsen, Hanne Erdman Thomsen (eds.): *Terminology and Content Development, TKE 2005*, 7th International Conference on Terminology and Knowledge Engineering, Copenhagen, 2005, s. 161-173.

Nielsen, Louise Pram (2009). Terminologisk begrebsafklaring på skatteområdet – en analyse af terminologilæren sammenlignet med fagleksikografi. ED-Afløsningsopgave afleveret på Åben Uddannelse, Copenhagen Business School den 16. november 2009.

***