# National Data Standardization: A Place for Topic Maps?

Lars Johnsen

University of Southern Denmark, Engstien 1, 6000 Kolding, Denmark
larsjo@sitkom.sdu.dk

**Abstract.** This article deals with national data standardization efforts in Denmark and discusses the role Topic Maps – and topic maps – may play in a new standardization strategy currently being considered by the Danish National IT and Telecom Agency. The strategy entails a paradigm shift from syntactic data standards based on XML schema to a more semantically based approach involving, among other things, the development, publishing and sharing of so-called definitions. The article gives an account of the historical, political and technical context of the strategy pointing out some of the opportunities and constraints this context poses for the introduction and application of Topic Maps as a recognized "data standardization standard" in Denmark.

## 1 Background

For some time now the use of open IT standards in local, national and international e-government has been promoted by the Danish National IT and Telecom Agency (ITST). This policy is in line with B103, a bill that was passed in 2006 by the Danish Parliament laying down a kind of "comply or explain" principle that authorities such as ministries and municipalities need to observe when implementing new IT solutions.

Seven sets of open standards considered to be particularly important have been made compulsory by the Danish state. One is OIOXML (= Open Information Online XML), a set of national guidelines for developing and reusing XML schemas for data exchange purposes. These OIOXML schemas comprise, among others, *core components*, XML schemas usable across the entire public sector, and *domain schemas* designed for data encoding in specific areas like education, environment or health [1].

To date, a fairly great number of OIOXML schemas have been developed by authorities and IT vendors and published and shared through a public repository (recently discontinued for reasons that will become apparent below). It is increasingly being felt, however, that many of these XML schemas lack a sufficiently clear semantic foundation. Basically, they only specify how data should be encoded and exchanged in XML and not what these data really signify or mean or in what contexts they are appropriate. Take one simple example like the XML schema specifying how the marital status of a person living in Denmark

Maicher, L.; Garshol, L. M. (Eds.): Linked Topic Maps. Fifth International Conference on Topic Maps Research and Applications, TMRA 2009 Leipzig, Germany, November 12–13, 2009 Revised Selected Papers. Leipziger Beiträge zur Informatik. ISBN 978-3-941608-06-1

should be represented in OIOXML. It states that the XML element to be used is <MaritalStatusCode> . . . </MaritalStatusCode> with one of the following data values:

1. married
2. divorced
3. widow
4. registered partnership
5. abolition of registered partnership
6. longest living partner
7. deceased
8. unmarried

Although these data values are based on concepts commonly known in Danish culture, they nevertheless give rise to a number of questions such as:

1. Does a "registered partnership" only involve persons of the same gender?
2. Widow is a recognized value. Why not widower?
3. Some data values denote "persons" (widow), some "relations" (registered partnership) and some "states" (divorced). Is this intentional or purely accidental?

Broadly speaking, the problem with existing OIOXML schemas is that they are not based on conceptual, or ontological, descriptions and are not linked to relevant documentation, i.e. guidelines, legislation, etc. In addition to such shortcomings, it has not always been easy to find the right schemas in the repository, grasp how individual schemas are interconnected, let alone obtain information on where and when individual schemas are actually used.

## 2 A New Paradigm

Therefore, the ITST is considering a new strategy for data standardization in country-regionplaceDenmark. The strategy entails a shift of emphasis from syntax to semantics that focuses on defining and describing concepts and processes relevant to e-government in Denmark rather than their formal encoding. Instead of XML schemas, data standards are to be based on several kinds of so-called *definitions* in the future: semantic definitions, data definitions, message and service definitions and so on.

Semantic definitions will play a pivotal role in the new setup. A semantic definition is the description of a concept deemed to be of importance in some domain. Its informational content is intended to be reused in ontologies, taxonomies and data models. Semantic definitions will be heavily based on ISO 1087, the international standard for terminology work.

Data definitions are specifications that define what information or data elements may be attached to certain concepts. For instance, a data definition may indicate that a "citizen" must have a civil registration number, an address and possibly a phone number. What data definitions will exactly look like is not yet totally clear but much inspiration seems to have been derived from the

Core Component Technical Specification (CCTS) with its emphasis on reusable information items and aggregating mechanisms [2]. Data definitions are organized in hierarchical structures to form messages which are the basic information units to be exchanged between software systems, typically web services. The precise structure of message and service definitions has not yet been finalized, either.

Information contained in definitions may be defined in terms of one or more *contexts*. The function of contexts is to indicate relevance or validity. A context may either be an organization or authority such as an institution, ministry or municipality, or some professional domain. It may also be a specific piece of legislation.

OIOXML formats will be developed for the various types of definitions in the pipeline. It is assumed that XML schema, and other "syntactical" resources like Web Service Description Language (WSDL) files, may be generated automatically, or semi-automatically, from these definitions thus reducing the technical burden of organizations and individuals keen to get involved in data harmonization and standardization. An open source desktop IT tool is being developed by the ITST to help achieve this goal.

The syntax-to-semantics strategy goes hand in hand with efforts to enhance public involvement and engagement. These efforts are most conspicuously reflected by the web site Digitaliser.dk (http://digitaliser.dk), which was launched quite recently. The web site has primarily been designed to be a meeting place for everybody interested in digitalization in the public sector. Its Web 2.0-ish functionality and features are evidence of this: its information architecture, for instance, is centered round user groups, or communities, and their possibilities for uploading, discussing and tagging resources. The site, however, is also meant to be the central repository of all interoperability assets relevant to the digitalization of the Danish public sector, including ontologies, taxonomies, XML schemas and WSDL files.

Although the strategy of the ITST will no doubt add a lot of value to data harmonization and standardization processes in Denmark, it arguably also raises some issues that have to do with the organization, integration, findability and navigation of resources. Put somewhat simplistically, the strategy seems to lack the glue that will make all the pieces fit together.

Thus, it is not difficult to see that Topic Maps may have a role to play in the scenario envisaged. With its emphasis on connecting concepts with content, Topic Maps could function as a superimposed (integration) model for data standardization that will allow resources such as semantic, data, message and service definitions, XML schema and WSDL files as well as metadata like tags to come together in meaningful, and accessible, structures. Some aspects of such a proposal are elaborated upon in the remainder of this paper.

## 3 The Semantic Foundation

As noted above, semantic definitions will form the basis of much, or most, data standardization work in placecountry-regionDenmark in the future. The minimal semantic definition will consist of a term, the verbal manifestation of a concept, and a definition of this concept. Following the ISO 1087 standard, however, semantic definitions will potentially be able to hold other conceptual and terminological information (see below).

There are several uses to which semantic definitions might be put in a Topic Maps-based approach. Most obviously, semantic definitions may function as PSI's. Since a semantic definition - by definition - denotes a concept, contains its definition and will have a permanent address at Digitaliser.dk, it naturally lends itself to this role. In this capacity semantic definitions are the meaningful anchor points to which all other resources can point.

But semantic definitions may also provide the stuff of which entire Topic Maps-based concept systems, or ontologies, are made. A semantic definition will normally contain information about one concept but will supposedly also include references to other concepts with which the concept is related. Using these references as subject identifiers, a Topic Maps system will be able to merge content from disparate semantic definition into one or more comprehensive, and hopefully coherent, concept systems.

Deploying semantic definitions as input for concept systems in topic maps of course invokes the question of how well information categories in ISO 1087 map onto Topic Maps constructs. A full-blown comparison between the two standards is, needless to say, not within the scope of this paper, but some general observations should make it clear that a mapping is in fact fairly straight-forward:

The key concept in ISO 1087 is that of "concept". Concepts are defined, somewhat vaguely though, as "units of knowledge". Concepts are divided into individual concepts and general concepts that "correspond" to one or more objects respectively. Objects are said to be "anything perceivable or conceivable".

The meaning of a concept is made explicit in a definition and through the assignment of characteristics, or properties, or more formally by means of feature specifications.

Concepts may be connected to each other through hierarchical or associative relations and be manifested in various ways, typically by terms, verbal designations. Labels may be attached to a term to signify its level of formality, applicability or status (preferred, obsolete, deprecated, etc.) or to indicate its role in relationships with other terms (synonym, antonym, etc.).

Data related to individual concepts and their designations are called terminological entries in ISO 1087 and normally constitute the basic information unit in a terminological data collection.

Although it is not entirely clear whether concepts are abstract or mental entities, or symbolic representations, it makes sense to interpret concepts in ISO 1087 as topics and objects as subjects. Likewise, individual and general concepts naturally map onto topic instances and topic types.

Terms are similar to names in the Topic Maps paradigm and may be typified through name types or scope. To indicate that one term/name is the synonym, antonym, homonym or equivalent of another, it needs to be reified.

Semantic content in the form of definitions or feature specifications are most naturally realized as occurrences and occurrence types. To add notes or similar additional information to an occurrence containing a definition, say, this occurrence must also be reified.

## 4 Topic Maps as a Resource Organization Model

The data standardization strategy will in due course yield a great deal of digital resources but so far no detailed scheme has been devised to ensure that these resources are properly linked, classified or organized. It is an implicit assumption that resources will be stored at Digitaliser.dk. The drawback of this solution is, as already mentioned, that this site is more of a Web 2.0 collaboration platform than a resource repository. One current problem is that resources must be placed with specific user groups or communities which in turn risk becoming a kind of silos within the site. Another is that the site today only provides rudimentary functionality for categorizing resources into classification systems like taxonomies or suchlike structures. The site does offer tagging as an organizational tool but as with most tagging systems, there is currently no way to link tags in hierarchical, or even associative structures.

What is expedient, though, is that the site has a REST API facilitating machine processing of some of its contents. This interface exposes site contents in XML via structured URL's and provides recognizable links between related sets of data. This means that all resources are given (more or less) transparent web addresses that may be used as subject locators.

This is especially useful in a scenario where Topic Maps might be used as an integrative technology for organizing and combining resources within the various "silos" of the site. Through the REST interface, data from, and about, resources in the site can be extracted and "mashed up" in one or more topic maps. One can think of several ways of fruitfully applying Topic Maps to organize and integrate resource data at Digitaliser.dk:

Firstly, it would be useful to be able to categorize resources in taxonomies across user groups. Such taxonomies would group interoperability assets according to, say, their function or genre and XML schemas according to their applicability (core component or domain schema) and/or the web services in which they are used. Secondly, it would make the site more user-friendly if users were able to visualize how resources are related, for instance how different XML schemas are embedded within each other. Thirdly, it would enhance the usability of the site if metadata like tags could be attached to resources in more flexible ways, for instance to resource structures, or if tags could be associated in typed relations to provide more adequate levels of description.

The attractive thing about Topic Maps is of course that all this can be done within one information model. A (merged) topic map may simultaneously classify a number of XML schemas, demonstrate how they are embedded within each other, indicate what tags or tag structures are attached to them, and link them

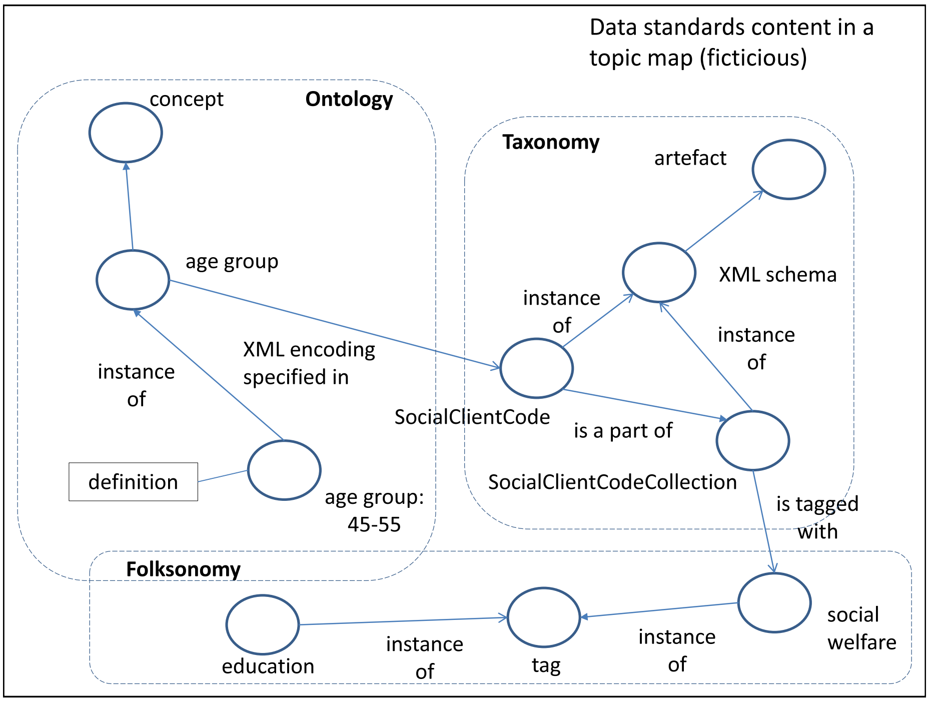

**Fig. 1.** Topic map organizing data standards content

to semantic definitions to expound the meaning of the concepts on which their element declarations are ultimately based (see figure 1).

## 5 Topic Maps as a National Data Standard?

Although it is evident that Topic Maps has a role to play in national data standardization in Denmark, and no doubt elsewhere, it is less evident just what status the standard should be given.

The most ambitious/interesting/daring approach in this context would be to use Topic Maps as the underlying data model for all resources that will make up data standards in the future. In essence, this would mean that semantic, data, message and service definitions should all be modeled according to TMDM and be represented and exchanged as topic maps. One unifying model for resources serving a common purpose!

This is probably not going to happen, at least not in the short run. Introducing Topic Maps as an all encompassing framework for building data standards is in itself not a major obstacle given the flexibility of the data model. The barrier is more of a political and a practical one. Topic Maps is, alas, not very well known in Denmark and by those who do know the standard it is more often than not simply perceived as a metadata standard along the lines of Dublin Core and similar standards.

Also, the whole conceptual basis of the Topic Maps paradigm with its notions of subject, topic, subject identity and merging and so on is likely to be seen as an additional complexity by the ITST in its endeavors to explain what data standards are and why they should be complied with.

Furthermore, the ITST would have to cope with the issue of translating all of the core terminology of Topic Maps into Danish if the standard was to be adopted at a national level. This in itself poses a bit of a challenge because many key terms in the model translate into Danish equivalents that for some reason or another seem less than perfect. Some examples:

In Danish there is really only one word that naturally translates “topic” and “subject” and that is “emne”. One could use the word “subjekt” for “subject”, since this is the term for the grammatical notion of a subject of a sentence. But “subjekt” sounds formal, almost nerdy, and when written it may be confused with its identical homograph “sub’jekt” (pronounced with a stress on the second syllable) meaning “seedy person”.

The equivalents of (subject) “identifier” and “indicator” are “identifikator” and “indikator”. These two words are so phonologically similar that they are likely to be confused. The proposal to use “descriptor” instead of “indicator” is, from a translation point of view, therefore recommendable. Subject “locator” also presents a problem since there is no direct translation of the term. One could try “adresse” (address) but as the word obviously also denotes locations of a physical kind, this is not optimal either.

The possible translations of “occurrence” are “forekomst”, “hændelse”, or “belæg”. “Forekomst” is used to indicate the place(s) where something actually occurs; “hændelse” means event; and “belæg” is a formal term connoting evidence of existence and validity. In this case the roots of Topic Maps in indexing theory and practice are more of a hindrance than help and one may wonder whether a term like “resource” (Danish: “ressource”) might not do the job better than “occurrence”.

“Association” is probably best translated as “relation” in Danish since “association” in Danish implies a kind of network of mental connotations that is triggered by a stimulus of a certain kind.

Last but not least, the notion of “merging” is tricky. In Danish equivalents like “sammenslutning”, “sammenlægning” or “fusion” are used about aggregation or integration processes that involve companies or organizations and not digital entities. “Sammenfletning” is perhaps more suggestive as it connotes things like (beautiful) wickerwork, the smooth operation of traffic flows on busy motorways and so on.

## 6 Topic Maps as an Ancillary Standard?

A more realistic scenario seems to be one in which Topic Maps is used as a kind of ancillary standard for organizing, integrating and enriching *data standards* *data* across organizational and technical boundaries. To help realize this goal, some essential resources need to be developed and put into place:

1. *A Topic Maps-based ontology for data standardization concepts and content.* This ontology must define, describe and relate topic types like concepts and terms and list what occurrence types may be attached to them. It should also declare topics representing "addressable" subjects relevant to the area of data standardization (definitions, XML schemas, WSDL files and so on) and indicate how metadata elements may be linked to these resources. Last but not least, the ontology should specify how ontological structures may be linked to classification schemes like taxonomies and folksonomies.
2. *Concrete topic maps containing relevant mergeable "public" information.* To ensure that they are applied properly and in the right settings, data standards may be enriched with information that indicates their context of use as this concept was defined above. These contexts might be lists of regions, municipalities, or public institutions, references to important legislation, or web service registries. In this connection, it would be appropriate, for instance, to create a Topic Maps-based version of FORM, the new reference model used to map services offered to Danish citizens and enterprises by the public sector. The reference model itself seems to map nicely onto a topic map as it not only constitutes a faceted taxonomy of services organized according to several parameters but also allows for links to relevant pieces of legislation, published workflows, organizational units, etc.
3. *Conversion tools.* Tools are needed to carry out the mapping of data in OIOXML, and other XML formats, to topic maps (in formats like XTM). Most obviously, tools must be developed that will allow users to extract REST API data directly from Digitaliser.dk and mash them up into topic maps. These tools may range from simple XSLT transforms to be used in a topic map editor like Wandora, or the desktop tool being developed by the ITST, to entire web applications. Some of these web applications may even be integrated with Digitaliser.dk itself. For instance, it would be useful to build a Topic Maps-based "portal" on top of Digitaliser.dk offering users more organized or holistic views of the resources contained in the site('s silos). Furthermore, tools to convert data from existing (proprietary) terminology systems in the public sector need to be developed.

This scenario may in fact accommodate a range of strategies for utilizing Topic Maps as a tool for organizing and integrating data standards content. Some strategies may be "product-oriented" in the sense that they explicitly seek to encourage the creation, publication and sharing of *OIO topic maps*, truly open and reusable information products about "the state of Denmark" (to use a somewhat lame literary allusion) while other approaches may aim at technological solutions based on Topic Maps operating "covertly behind the scenes" to create organizational and/or navigational overlays to existing data sources. In the latter case, emphasis is not so much on developing and sharing new, and hopefully better, information products but rather on integrating resources already in place in novel and interesting ways.

## 7 A Place for Topic Maps?

Since Digitaliser.dk is intended to be a platform for knowledge sharing and collaboration for organizations and individuals, there is no reason whatsoever why topic maps should not be uploaded, tagged and distributed via the site. Topic maps can be associated with certain user groups or uploaded to a section specifically set up for the purpose of disseminating and sharing topic maps relevant to e-government in Denmark. The vision of the latter solution might be the emergence of a national hub, or clearing house, for OIO topic maps.

XTM 2.0 seems to come in handy for this purpose. One important reason is that XTM 2.0 makes it possible to "package" conceptual information with OIOXML. Since an internal occurrence, or more precisely the <ResourceData> element, is permitted to contain arbitrary XML structures in XTM 2.0, OIOXML schema content, typically element declarations, can be embedded in a topic map thus allowing concepts to "travel together" with their valid OIOXML data values. Another is that the format allows topics to have several subject identifiers. This means that a topic representing an OIO concept may point to its PSI, i.e. semantic definition, through the normal "human readable" interface of Digitaliser.dk as well as its REST API. For instance:

```
<topicMap version="2.0" xmlns="http://www.topicmaps.org/xtm/">
 <topic id="enke">
  <subjectIdentifier href="http://digitaliser.dk/resource/123"/>
  <subjectIdentifier href="http://api.digitaliser.dk/rest/
                     resources/123/artefacts/enke.xml/content"/>
  <instanceOf>
   <topicRef href="#OIOconcept"/>
  </instanceOf>
  <name>
   <scope>
    <topicRef href="#DA"/>
   </scope>
   <value>enke</value>
  </name>
  <occurrence>
   <type>
    <topicRef href="#OIOXML"/>
   </type>
   <resourceData
    datatype="http://www.w3.org/2001/XMLSchema#anyType">
    <MaritalStatusCode>widow</MaritalStatusCode>
   </resourceData>
  </occurrence>
 </topic>
...
</topicMap>
```

This way of connecting conceptual information, semantic descriptions and OIOXML representation rules will no doubt alleviate the current problem of undocumented "stand-alone" OIOXML schemas mentioned above.

## 8 Topic Maps Throughout

The primary aim of this paper has been to demonstrate the relevance of Topic Maps within the national data standardization framework being proposed by ITST and the current technical infrastructure provided by Digitaliser.dk. But the application of sound Topic Maps principles, in particular the insistence on transparent, consistent and robust subject identification, may actually be taken a step further, i.e. to the framework and infrastructure as such. A simple example may illustrate this:

The REST API of Digitaliser.dk provides, as already noted, an interface to the site's contents and metadata. The XML structure of the data which may be extracted via the interface is specified in a number of XML schemas. But information on how these XML schemas themselves are to be understood is nowhere to be found (at the time of writing). And this can, and probably does, lead to ambiguity or even confusion sometimes. One may wonder, for instance, whether "TaggedItem" and "TaggedObject", the names of two XML elements in two separate schemas, are really labels denoting different subjects or just synonyms for the same thing, that is to say site resources that have been tagged.

In other words, to make the most of Topic Maps in a data standardization context like the Danish one, the principles of the paradigm should not only be applied to the data resources themselves but preferably also to the whole system underpinning their design, creation and dissemination.

## 9 Final Remark

I am indebted to three anonymous reviewers for their valuable comments on an earlier version of this article.

## References

1. IT- og Telestyrelsen (2008): *OIOXML som obligatorisk, åben standard – uddybende vejledning*. Accessible at:
http://www.itst.dk/arkitektur-og-standarder/Standardisering/Aabnestandarder/vejledninger-og-publikationer/OIOXML%20-%20uddybende%20vejledning.pdf
2. Core Components Working Group (CCWG) (2008): *Core Component Technical Specification (CCTS)*. Accessible at:
http://75.43.29.149:8080/display/public/Core+Components+Working+Group+%28CCWG%29