Towards Semantic Interoperability

Jeanine Lilleng
Brønnøysund Register Centre, Brønnøysund, Norway*
jel@brreg.no

* Currently also affiliated to
Department of Computer and Information Science,
The Norwegian University of Science and Technology
Trondheim, Norway
lilleng@idi.ntnu.no

Abstract: We explain how semantic interoperability is important for the Norwegian public sector. Important aspects for a technical solution are discussed. We choose UML to help us with our semantic interoperability efforts. Some of UML's strong point and week point are considered. Our resulting TOR system is also described.

1 Introduction:

Brønnøysund Register Centre began in 1997 the process of registering all information submitted by Norwegian companies. The motivation was to discover multiple reporting of identical data. Data definitions describing the information requested by agencies were gathered in a database. It was soon realized that one could create electronic forms based on these data definitions. Some additional functionality was added, and from 2001 XML Schema and X Forms specifications where made available. Today this database consists of well above 20 000 entries.

One of the main challenges for the government agencies is to manage their information. Most of the information is duplicated across the agencies. The different agencies are different agencies because they have different focus and different tasks. This also means that they handle their information in different ways. As long as there is no need for exchanging information this is not a problem, but when exchange of electronic information becomes an issue, it is not possible to exchange information between different databases directly.

Much of this information is submitted to the agencies by the citizen of Norway. Often they have to submit the same information several times. This should not be necessary, but with the current system for describing and saving information it is the only option. We have laws that say, that any agency can get information from any other, as long as they are allowed to request that information themselves. This law have potential for saving a lot of time for the Norwegian Citizens and the agencies, but we have not yet the technology to support this law.

The goal of the TOR project is to simplify information exchange between agencies and citizens. To make it possible to receive data in an electronic format it must be organized in some way. Most data retrieval by the Norwegian agencies is done through custom made web pages or through Altinn [1]. A lot of work is being put into the organization of this data, but still this is not very helpful when it comes to exchanging information. We believe that the

added work needed to support information exchange a lot better, is relative low compared to the effect.

We want to make use of well known technology and established standards to improve the Norwegian infrastructure. We are creating a family of tools; TORmodell, TORdesign and TORnett that makes it possible to model a domain and then use these models to request, retrieve and exchange information in a more efficient way. These tools are based on open standards and shareware / freeware. The end product is XML Schema and X Forms definitions. These definitions describe the information and the form designed for requesting the information.

These tools are to be freely available to the all Norwegian agencies and others interested. Everyone given access can model what they like. Reuse of modelled elements is important, but we do not want to have a strict control with the model created. In stead we expect self justice and system support to steer the overall model in a good direction. These principles are similar to the principles community development of software, like the principle Linux where built on.

The model describing information that will result from this work is expected to help computers identifying some of the semantics in the data being exchanged. This is also expected to increase reuse of already retrieved information and simplify information exchange inside and between agencies and / or other organisations. We hope that this will open other new possibilities as well, and be an important step towards semantic interoperability in Norway.

Currently the TOR project at the Brønnøysund Register Centre is working on these challenges. This paper describes some of the challenges we have met and sketches the system architecture.

2 Requirements

Since the standard ways to describe data cannot be directly applied to cover our needs in connection with semantic interoperability we have to take a closer look on other possible solutions, but first we se if there are other necessary considerations that must be made.

2.1 Local attachment

To many agencies their database and the data contained in it is one of their most important resource. When we move focus from data to information, this might be seen an attempt at removing resources, expertise and influence away from the agency. This will not help when it comes to cooperation in connection with the new system.

To reduce this effect it is very important to makes sure that the agencies still feel that the same connection with their information models as they do with their current databases. This means that it is important that they keep the ownership and responsibility for their domain models.

2.2 Local it-skills

Creating and maintaining a domain model require other knowledge than the knowledge required for doing database management. Another aspect is that some of the smaller agencies hardly know their current database schemas. Obscure techniques for describing the domain model like second order logic or frame based technology will require skills that very few possibly none of the agencies currently possesses.

The chosen technology and standards must be as close as possible to existing technologies and metaphors. This is important, both to make the introduction as smooth as possible and to avoid that costly consultant services must be bought to become operational with the new domain model.

It is also important that as much as possible of the existing infrastructure can be used together with the new semantic interoperable information exchange.

2.3 Local expert knowledge

To be able to create the domain model that must be the core of any semantic interoperable system. This domain model can only be created with close cooperation between it – people and the domain expert. The emerging model must be accessible to both groups. Hence it must be possible to visualize the domain models in a way also no it people can understand.

2.4 Many different actors involved

The fact that many agencies will be included in the effort of putting semantics into the information, also imply that many people will be involved. This makes it important to put as much of the organization and ideas form the modelling process into the model. Mental models found in the head of the modellers are difficult to get access to, and should not be necessary to create and / or change the models.

2.5 Forms and yearly revisions

The way the Norwegian agencies think about their information is closely related to their yearly revision of paper forms. Even though the amount of information submittet electronically is constantly rising, most agencies still have their main focus on paper. We wanted to create a system that was both based on the ideas of open source and community contributions like Wikipedia [7]. We wanted to make the work done by other modellers available and reusable. The system should also have multi user functionality.

3 Semantic interoperability

Semantic interoperability refers to a system's ability for exchanging information inside organizations with heterogeneous information systems and / or between organizations. These issues are known both for e-government and private businesses. Park and Ram [4] give a thorough introduction to the some of the challenges of semantic interoperability. Their main focus is on businesses and challenges in organizations with heterogeneous information systems, but their analysis also apply to government and inter-agency information systems.

Used in this paper semantic interoperability refers to the possibility for exchanging information between systems and organisations without having to do tailoring to make this

possible. Solutions allowing for semantic interoperability are more complex because they take into account what is needed to make seamless information exchange possible.

Semantic interoperability cannot exist before technical and syntactic interoperability are available. In Figure 1 - you can see how the different levels of interoperability relay on the level below them. What the different levels are considered to contain depends on the point of view of the paper. In this paper the focus is on information exchange and the sketch of the levels of interoperability reflect this.

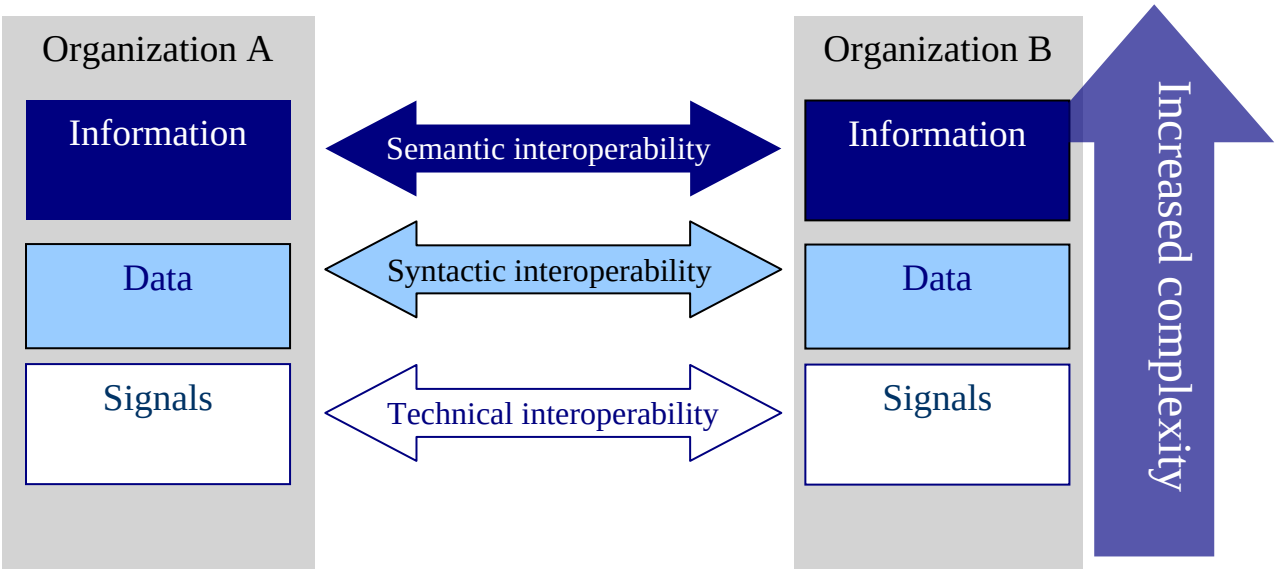

Figure 1 - The increased complexity and required infrastructure as we move towards more advanced interoperability

Technical interoperability is the first level, and makes it possible for two computers to exchange signals. Historically it has not been given that computers created by different manufactures could communicate. Today systems not being technical interoperable with other systems are very rare.

The fact that two systems are technical interoperable does not mean that they can interchange data. Before data can be exchanged the systems must agree on the format for data exchange. This is often solved by using XML today. As syntactic interoperability becomes common it becomes possible to focus on semantic interoperability. Our effort put into semantic interoperability is described in this paper.

4 From data to information

The first challenge we meet every time we talk about semantic interoperability is too much focus on data. The information in the different Norwegian public agencies is saved in databases and the database schemas are thought to be important in it own right. This makes sense if you take into account that the databases have been the spine of these agencies for a long time. Still this point of view can be contra productive when it comes to making semantic interoperability come true.

4.1 Identical information

Having a data centric point of view to information, means that identical information put into different database schemas appears different. To some extent this is true, but when you add interpretation of the data and get the resulting information, the information saved in the two databases are identical.

When you are working with semantic interoperability it is the similarity or dissimilarity of information the counts, not the way the data is formatted. This is an important point, because for many people working with databases and XML based information exchange, the information only match if the database / XML schemas are identical.

Person

first name	family name	address
Katrine	Olsen	Jonsvannsveien 34 7050 Trondheim
Ola	Normann	Ingensteds 42 4015 Stavanger

Figure 2 - One possible way to put the information about Katrine and Ola into a database.

Person

name	address road	adress number	postal code
Katrine Olsen	Jonsvannsveien	34	7050
Ola Nordmann	Ingensteds	42	4015

Zipcode - City

zipcode	city
7050	Trondheim
4015	Stavanger

Figure 3 - Another equivalent way to put the information into a database.

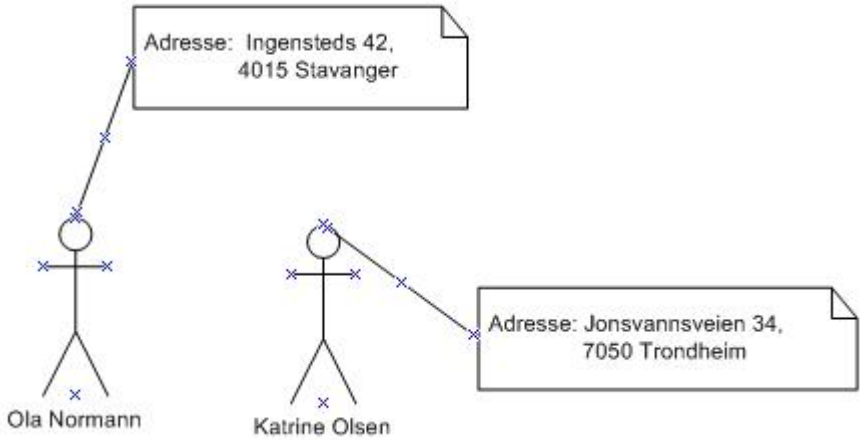

Figure 4 - Information to be represented in a database

Both Figure 2 - and Figure 3 - show databases containing the information found in Figure 4 - . It is easy to see that the data contained in the databases are different, in spite of these differences in data it is important to be able to recognize that the information contained is the same. Diagram like Figure 4 - do not exist for every database, so we have to find some other way to detect identical information independent of potential differences in data formatting.

4.2 Common information, separate data

The mental models that constitute your world view help you understand the world, but it sometimes also prevents you from seeing certain parts of the world accurately. This is especially so in the computer sciences, where everything we work with is models of the world. The previously mentioned database schemas are one way to see the world, but we

claim that having focus on information in stead might make the quest for semantic interoperability easier.

The sum of information that exists about a person or an object can be looked upon as one “unit” of information. This information can physically be situated in several different databases using different database schemas. This might not seem to be a very different point of view, but it leaves out some of the assumptions that can be troublesome when working with semantic interoperability.

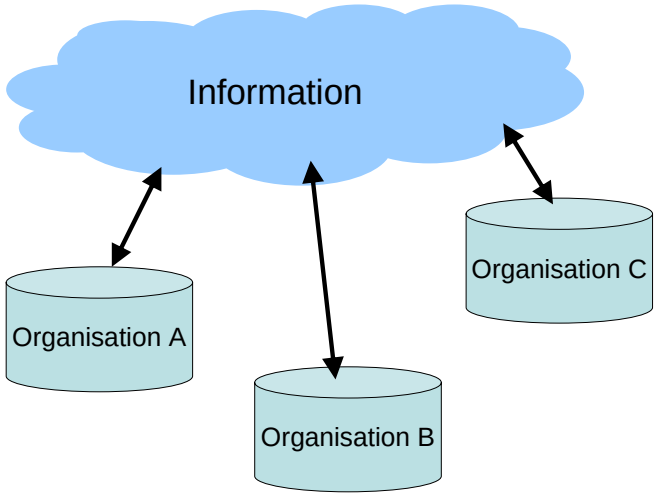

Figure 5 - Several locations or organizations can contribute to the overall amount of information.

Some implications of this point of view follow:

- The same information about the same item can be stored in different formats at different locations. The chosen format / database schema can be different, but that has nothing to do with the semantics of the data.
- When retrieving information about something it is probable that the needed information is spread to more than one location.
- For a system or a person retrieving information it is of no consequence how this information is stored and where. The local chosen organization of data should be transparent to the users.

So the conclusion seems to be that systems for retrieving and / or exchanging information should be based on something different that the local data storing format. This also fits with the idea that differences in data format do not necessarily mean that there are differences in the semantics.

What is needed is some way to describe, exchange and retrieve information that is independent of where and how the underlying data is saved.

5 UML

If any method of describing information is to be put into use in an organization, that have not got information description and Artificial Intelligence (AI) as their main focus, the method for describing must be relatively easy to learn and easy to understand. Some methods that might have been an alternative for others, like first order logic or KIF was not an alternative for us.

The challenge due to complex modelling where also brought to attention by Cranefield and Purvis [2].

Having done some research we realized that parts of the organisation already knew Unified Modelling Language (UML), and that they were familiar with using UML tools to describe software systems. Another important factor where the work done by KITH [3]: They use UML as a tool when they design messages for use in the health sector. As a part of this work they involve domain experts like physicians and nurses when they describe the domain. KITH uses UML class diagrams to describe the world. Their experience is that UML is a good tool for visualizing a domain model.

When you model you describe a chosen portion of the world. UML [6] was created to model object oriented software systems. UML has many different diagrams, which one you use depend on you chosen modelling focus. The most used diagram is probably the static structure diagram also know as a class diagram.

A UML class diagram can be used to create a conceptual description. It was created to describe software systems, but it can also be used to create an arbitrary ontology. What we need is an ontology that describes the information the agencies gather and maintain. This will make it easier both to keep tabs of existing information and exchange information between agencies.

5.1 UML and semantic interoperability

We needed a way to describe the information that was both easy to understand and easy to create and maintain. UML seems to cover these needs, but how does it support our needs in connection to exchange of information and semantic interoperability?

There is no formal semantic description as a part of the UML language. The semantics of the different elements are described in English text. This gives a known semantic challenge: Textual descriptions are often ambiguous.

Still the elements and constructs available in UML “imply” semantics that is relevant for us. KITH has used UML to some of the same ways we need to. Their experience indicates that some of the elements in class diagrams appear to introduce much ambiguity. These constructs are often used different by different users. Description of information is difficult and one model can be made in many different ways. By reducing the available mechanisms we reduce the possible ways to model the same item.

Most “languages” have several ways to say the same thing

It seems that we can manage without some of the elements available in UML. One of the first mechanisms we decided to get rid of was all forms of aggregation. Their semantics is not clear, and we can express this in other ways.

5.2 How we use UML

UML was created for and is mainly used in software engineering. When we want to use it for describing information, some adaptations must be made. The main adaptations and the reason for them are addressed in this section.

5.2.1 Elements and relations

We use classes, attributes, semantic types, associations, association classes and generalisation / specialisation. In Figure 6 - you can se an example of a diagram created using these mechanisms. We have done quite a lot of modelling already and the reduced set of UML elements seems to be sufficient for our needs.

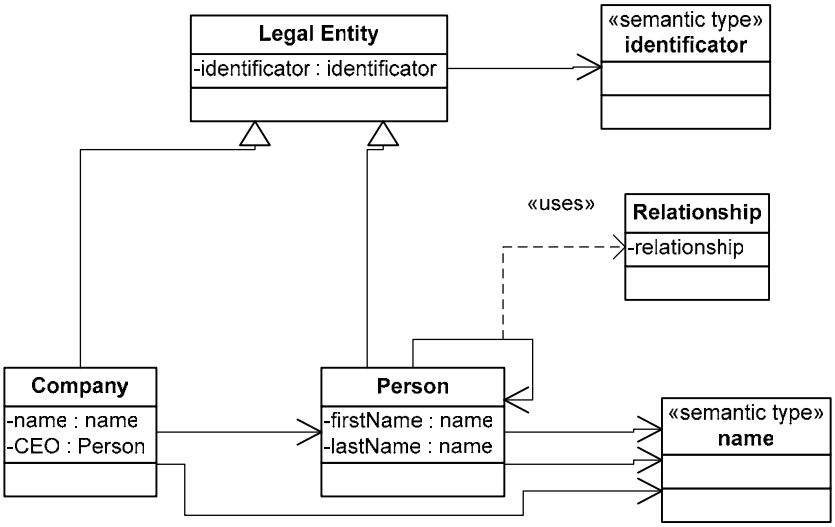

Figure 6 - An example diagram using the reduced UML set.

5.2.2 Semantic types

A central element in a UML class diagram is a data type. As you can see in 5.2.1 there are no data types in our system. There are two reasons for letting the data type go: Data types are used in these diagrams to indicate the type of the data to be saved. This is an implementational aspect that has nothing to do with semantics and the relations in information. The other challenge is that we want to emphasis the differences between data and information. For users familiar with UML, having data types present will probably make it more difficult to se the differences between modelling data and describing information.

To avoid these problems we introduced semantic types to substitute the data types. These semantic types have one semantic meaning. Seen from an information modelling point of view, it is much more interesting that a “Person” have a “firstName” that is of the semantic type name, than the fact that a “Person” has a “firstName” that is implemented as a string. Besides, to save the “firstName” as a string is only one of more possible approaches, and tells noting about the semantics of this “firstName”.

5.3 Beyond UML

For now our restricted UML is sufficient for our needs, but we might need to include more relations as time pass by. Even if we continue to move away from pure UML in that way, the help in visual model of the information and the familiarity for those who already know UML will help us. An interesting standard here might be the topic map technology [5].

6 The TOR architecture

The modelling done in TORmodell creates the basis for a semantically rich description of the domain. The form and message design done in TORdesign connects these descriptions with the needed formats and designs for making data exchange happen. The finished specifications are made available for the public and other agencies in TORnett.

Together these tools cover both the process of creating an electronic form or message and support for describing information semantically. As you work in TOR the described information will expand, and needed work before creating a new form / message will decrease.

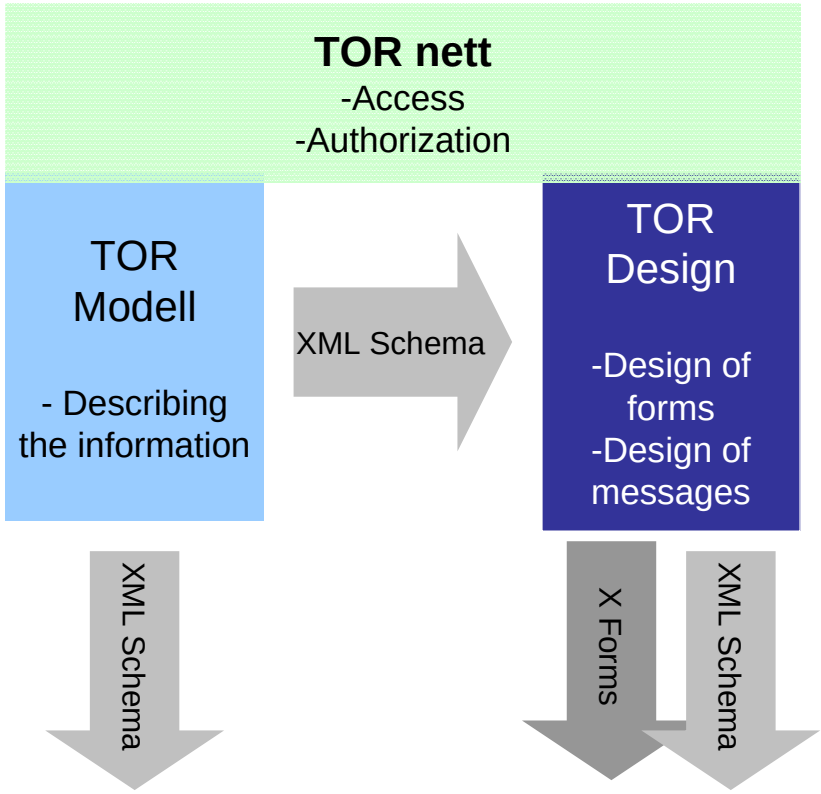

Figure 7 - The three main components of TOR.

6.1 TOR modell

All modelling in TOR is done in TOR model. This is a UML like tool where you can manipulate the model directly in graphical diagrams. The users of this tool must have some modelling experience. Subsets of the information model can also be defined and prepared for export to TOR design.

6.2 TOR design

This tool requires less knowledge of modelling than TOR model. We realize that we will have to do much of the modelling I TOR model ourselves in the start, but we belive that the needed knowledge for operating in TOR design already exist in the different agencies.

Both messages and forms are designed in TOR design.

6.3 TOR nett

All agencies will have access to TOR, at least a view access. On the other hand it is important that the correct person is given access to change the correct classes in the model. TOR nett makes sure that every one gets the correct access to the system.

When models are finished in TOR model and TOR design they can be made available to the public in TOR nett, so this is also the access point for the outside world.

6.4 Models

In different parts of TOR different tasks are done, based on different information. All this information is found or saved in on of the five models of TOR. The models are basically a collection of information.

6.4.1 Information Model

The Information Model is the part of the system where the semantic description of the information belong. All the modelling done in TOR model is saved in the Information Model. This model is common for every one that use the system, but what part you see of it depend on what you need.

The core of the Information Model is the modelling of the basic information; this is where domain models of different agencies meet. The basic information is identifying information like organizational number, information about persons and address information. Some parts of this core can be found in all agency systems.

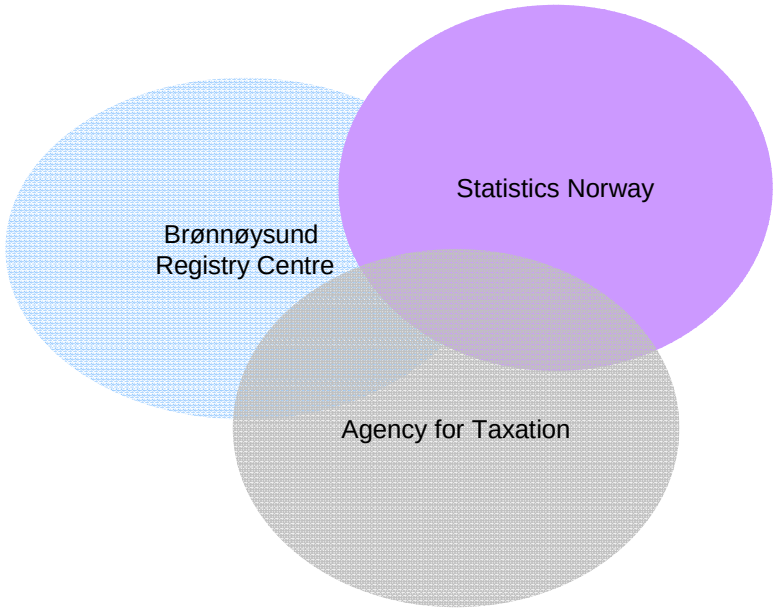

Figure 8 - How different domains can overlap in the information modell

The information in the different Norwegian agencies are not disjoint, and that is one of the reasons why we need to have a common model of the information being exchanged between different Norwegian agencies.

This model can be serialized into a XML Schema file.

6.4.2 Document Model

The information model will be huge when the system has been in use for some time. A document model is a subset of the classes in the information model. The document models are used as the basis for creating messages and forms. Associations and attributes that are irrelevant to the intended usage can be hidden in the model.

Document models are imported into TOR design and used to as the basis for creating messages and / or forms. The use of TOR design should require less understanding of modelling than TOR model, so it is essential to hide the full complexity of the information model.

This model can be serialized into a XML Schema file.

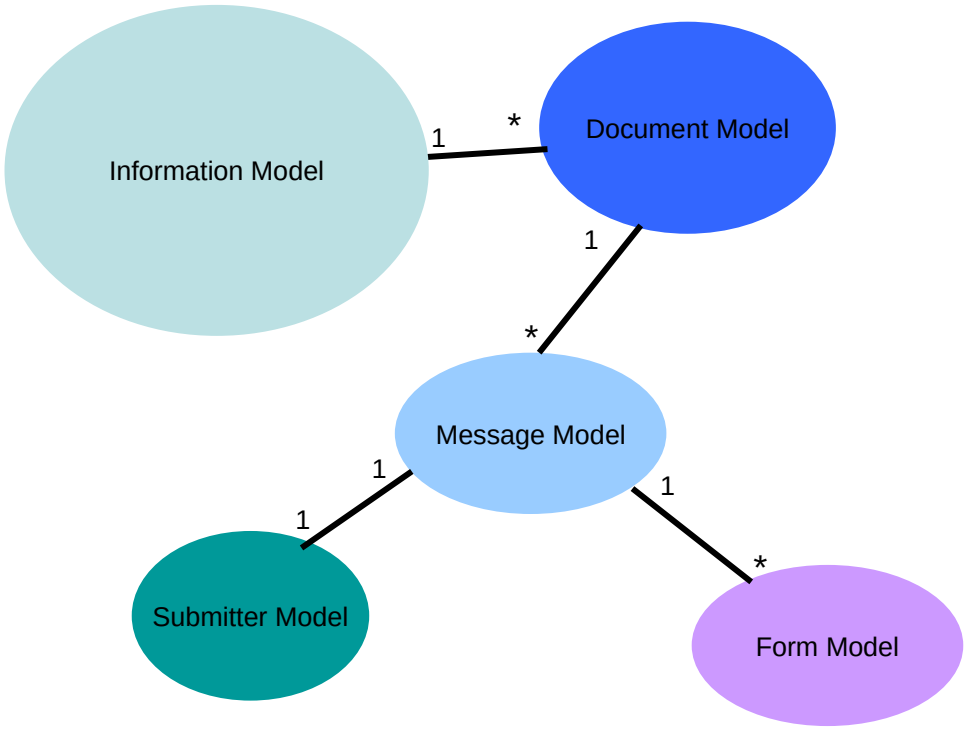

Figure 9 - The models in TOR

6.4.3 Message Model

A document model is the basis for a message model. A document model can be used as message model directly, but normally one wants to refine the messages further. Restrictions on values and the multiplicity of items in the messages can be defined here.

A message model can either be created manually from a document model, or it can be generated automatically for a form model.

A message model can be serialized into a XML Schema file.

6.4.4 Form Model

This model contains the graphical information needed to create an electronic form. When you design a form in TOR design this layout information is saved in a form model. The corresponding information needed to describe the information submitted through the form belongs to the message model.

A form model is serialized into X Forms.

6.4.5 Submitter Model

One of the main aims with the TOR project is to reduce the amount of multiple submission of the same information. If this is to work we need to know who is handing in which information, and when. This also makes it possible to manually generate personal forms in the longer run.

The submitter model is used for internal reasoning in the system, and can currently not be serialized.

7 Conclusion

When you have a semantic description of information, it can be expressed in any form and later compared with other forms. Hence it will be possible to compare the content of two different messages. If you also have mapped you internal information resources with the information model created in TOR you can easily exchange information with all other agencies having done the same.

We believe that TOR is a step towards semantic interoperability within e- Government in Norway.

8 References

[1] Altinn, http://www.altinn.no [last visited 11.02.2005]

[2] Cranefield, Steven & Purvis, Martin: UML as an Ontology Modelling Language, IJCAI-99

[3] KITH, Kompetansesenter for IT i helsevesenet, http://www.kith.no [last visited 14.01.2005]

[4] Park and Ram, Information Systems Interoperability: What lies beneath? ACM Transactions on Information Systems, Volume 22, Issue 4, October 2005.

[5] Topic maps, http://www.topicmaps.org/ [last visited 11.02.2005]

[6] UML, Unified Modelling Language, http://www.uml.org/ [last visited 14.01.2005]

[7] Wikipedia, http://www.wikipedia.org/ [last visited 14.01.2005]