Limber
Language
Independent
Metadata
Browsing of
European
Resources

| | |
|---|---|
| **Project ref. no.** | *IST-1999-11748* |
| **Project acronym** | **LIMBER** |
| **Project full title** | **Language Independent Metadata Browsing of European Resources** |

| | |
|---|---|
| **Security (distribution level)** | *Public* |
| **Contractual date of delivery** | |
| **Actual date of delivery** | |
| **Deliverable number** | *D-3* |
| **Deliverable name** | *LIMBER Metadata Environment Report* |
| **Type** | *Report* |
| **Status & version** | *Final Draft* |
| **Number of pages** | |
| **WP contributing to the deliverable** | *WP5* |
| **WP / Task responsible** | *NSD* |
| **Other contributors** | *UKDA, CLRC* |
| **Author(s)** | *Kyrre Moe, Terje Sande, NSD*<br>*Ken Miller, Titto Assini, UKDA*<br>*Brian Matthews, CLRC* |
| **EC Project Officer** | *Mr. Kimmo Rossi* |
| **Keywords** | *Metadata, DDI, Object-Oriented Modelling, Dublin Core, RDF, Semantic Web* |
| **Abstract (for dissemination)** | |

# Table of Contents

# Part 1: Summary of Metadata Work Package

## Objectives

To produce new metadata model for social sciences based on the existing ICPSR DDI DDT in XML. To investigate the expression of this model in RDF. To increase the interoperability of the datasets, and to extend beyond survey data.

## Description of work

The workpackage has worked towards the definition of a new metadata model which not only includes the items in the current NESSTAR prototype XML metadata descriptions based on the ICPSR Data Documentation Initiative (DDI), but also includes the semantic relationships and definitions of terms supported by RDF. This must capture the full range of requirements not only of the existing prototype, but also of those collected in WP3. This metadata model must accommodate those limitations noted in the development of the NESSTAR and DDI prototypes. Therefore this is not just the translation of existing metadata formats into RDF, or the integration of existing formats, but the complete definition of metadata formats for social science data, including attributes and possible values.

As the project progressed, it has become clear that the W3C activity in Metadata has not progressed as quickly as it might, and the concomitant slow take up of RDF in the wider community, it was decided to adopt a flexible approach and derive a generic OO model for DDI2 metadata. This is described in this document.

The RDF definitions of the concepts in the cut down version of the HASSET thesaurus (now about 5000 entries) can be incorporated into the metadata definition as thesaurus values to support the multilingual interface. The description of this work is included in WP6 deliverables and is omitted here.

# Part 2 Towards a semantic web

The existing World-Wide Web (WWW), based historically on HTML documents, and now increasing on XML based data formats, is perhaps the most successful open distributed system ever. However, the interpretation of documents and data on the web has been largely determined by the user reading documents or establishing domain specific formats for XML data descriptions. There is no notion of machine readable descriptions of the interpretation of text on the web. Historically, this has led to the "information overload" where search engines, using naïve textual matching, return too much information with little precision. Today as more sophisticated automated Web services are being demanded the need for *machine interpretation* of web based data is increasing. The move towards XML has done little to help; XML DTDs and XML Schemas allow more flexibility in the syntactic descriptions of data, but do little to agree the interpretation of that data; each application domain has to agree on the meanings of terms.

In response to this problem, the World Wide Web Consortium (W3C) has established a *Semantic Web* activity [Semantic Web]. The aim of this activity is to make the information presented on the web interpretable by machine. This activity is developing tools and techniques so that automatic agents can discover the relationship between machine resources, can interpret the meaning of those relationships and their attributes according to ontologies formally describing the properties of real world objects, and ultimately reason about the properties of these object in real-time to support services.

The basic mechanism underlying the Semantic Web activity is the *Resource Description Framework (RDF)* [Lassila & Swick 1999]*,* a language for describing relationships between objects. At its simplest level, RDF is a language for describing *triples* in the form of *(subject, predicate, object),* where the subject and the object of the triple are resources on the Web, and the predicate is some property of resources. More sophisticated statements can be made, including nested statements, bags and lists, and reified statements, allowing the assertion of statements about statements.

Particular classes of RDF models can be specified using *RDF Schemas.* This allows the user to specific *classes* and *subclasses* of resources, and *properties* and *sub-properties* between classes. However, RDF and RDF Schema are not sufficiently expressive to capture more than a simple graph model of objects and relationships. In order to realise the full potential of the semantic web, further components need to be constructed upon RDF.

- A means of defining more complex data models and ontologies with a more powerful constraint language.
- Logics to express rules and properties of RDF models and reasoning tools to derive true properties of RDF models.
- A language to express queries over set of RDF statements and have them evaluated in the logic.

The Semantic Web activity is planning to provide these technologies [Berners-Lee 1998] and tools that implement them are likely to appear over time. Currently, there is research activity in all these areas. The most advanced is the DAML+OIL proposal [DAML+OIL 2001] for capturing ontologies. The W3C is following this up within a Web Ontology activity.

Our vision is to harness the Semantic Web activity to provide support for deploying metadata within distributed systems, and using standard tools built to support the semantic web, to use those metadata descriptions for real time search and access within and across the social-science and related domains. This can be enabled by presenting Metadata using RDF, formally presented using RDF Schema within an ontology framework. The nature of RDF, which is designed for use with resources in a Web environment, is ideally suited when further information has to be retrieved from third parties. The work on Web Ontologies can also be used to support the description and use of thesauri, with more sophisticated constraints than currently allowed. This could then mediate the search process, with reasoning systems used via the semantic web could then be applied to the combination to determine whether data matches the required search in a more focused fashion.

## References

[Semantic Web] *Semantic Web Activity Statement*, World Wide Web Consortium, 2001 http://www.w3.org/2001/sw/Activity

[Lassila & Swick 1999] Ora Lassila , & Ralph R. Swick (eds), *Resource Description Framework (RDF) Model and Syntax Specification.*W3C Recommendation 22 February 1999.

[Berners-Lee 1998] Berners-Lee, T. *Semantic Web Road map,* W3C Note 1998 http://www.w3.org/DesignIssues/Semantic.html

[DAML+OIL 2001] DAML+OIL proposal, March 2001 http://www.daml.org/2001/03/daml+oil-index

# Part 3 The process

The metadata work of LIMBER has been based on the following assumptions:

- ❑ The necessity of a close co-ordination between FASTER and LIMBER (to avoid duplication of work as well as incompatibility of proposals).
- ❑ The necessity of positioning the development within the DDI process. Every step away from the DDI is removing us from the community that we are trying to support.
- ❑ A realisation of the fact that the current DDI 1.0 DTD is not solid and flexible enough to support the type of extensions needed to support FASTER and LIMBER functionality
- ❑ A very strong position within the DDI-committee that makes it possible to influence the direction of the standardisation efforts.
- ❑ An understanding of the "natural conservatism" that kicks in as soon as the first version of a standard is released. Organisations that have invested time and resources in a standard will always measure the potential gains from a change against the costs.

The first short presentation of the joint work of LIMBER and FASTER was done at the DDI full committee meeting in Chicago June 2000. Following this meeting, a workgroup was established to explore the possible move of the DDI DTD to either XML Schema or RDF. Unfortunately, the workgroup did never really establish itself so no concerted work was ever done within this group. However, the work was brought forward in LIMBER as well as FASTER and a first joint metadata meeting was held in connection with the LIMBER management meeting in Athens in September. Here it was decided to take a step back from the XML Schema versus RDF discussion and rather focus on the underlying model that the DDI-standard is built upon. The assumption of this decision was that a more object-oriented approach would prepare the ground for a more solid and extensible standard that eventually could be represented in either RDF or XML Schema. Without this step a migration to a new framework would only reproduce the problems and shortcomings of the existing version.

At the Athens meeting it was also decided to produce a document to be presented at the upcoming DDI workgroup meeting in Washington in November. The intention of this document was to argue for the approach as well as to describe how it could be achieved by means of concrete examples. The final document was discussed and improved at another joint FASTER/LIMBER meeting in Essex in mid November.

The Washington meeting was a joint meeting between the various DDI workgroups and not a full committee meeting. For this reason no major decisions could be taken at this stage. The meeting was attended by Ken Miller and Jostein Ryssevik (committee members), and Titto Assini and Jean Pierre Kent from FASTER. The joint FASTER/LIMBER document was presented by Titto. In addition a FASTER proposal regarding metadata descriptions of aggregate data (cubes) was presented by Jean Pierre.

The FASTER/LIMBER document was well received. Even though most of the participants had expected an exploration into the RDF versus XML Schema controversy, more or less everybody seemed to accept the need to prepare the foundation of the standard before the move to a new framework would make any sense. The most interesting discussion focused on the strains between the short- and longer-term perspectives of the DDI-work and on how this new approach could fit into the DDI-standards development cycle.

To understand this discussion the following facts are important:

- ❑ DDI 1.0 was released in February 2000 after about 5 years of work.

- The standard has been designed to meet the needs of the most important actors in the process; data archives and libraries archiving and disseminating social science micro-data.

- Several organisations have already invested a considerable amount of time and resources in the existing version (migrating data resources, changing internal procedures, developing software etc.). In many ways DDI 1.0 has already become a de facto standard within the archive world and people will be interested in safeguarding their investments.

- It has been proposed to develop the standard further under the DDI 2 label. The direction of this development and its relationship to the existing DDI 1.0 has however never been clarified. Two sets of objectives have been discussed:

  - To add support for other types of data than simple rectangular micro-data (this is a big issue than includes aggregate tables (cubes), time-series, data in relational databases etc. and also data coming out of modern CATI/CAPI type of environments)

  - To move the standard to a more appropriate framework (this is where XML Schema and RDF comes in).

- When it comes to the relationship between DDI 1 and 2, two positions was identified:

  - DDI 2 to be built incrementally adding more elements and content to the existing structure

  - DDI 2 seen as the next generation DDI, defining a clear break with the existing standard.

From one point of view LIMBER has achieved as much commitment from the DDI committee as could be hoped for. On the other hand there is an obvious contradiction between the long-term perspective of the DDI 2 process and the short-term demands of both the FASTER and LIMBER projects. Given this contradiction a two-track strategy was designed:

- **The short term track**: To propose and get acceptance for the minimum number of additions and changes to DDI 1.0 in order to meet LIMBERs requirements. This included a more explicit interface to external controlled vocabularies or thesauri, a more systematic use of keywords for all relevant metadata objects and a better solution to the multilinguality challenge.

- **The long term track**: To develop a more generic, flexible and extensible metadata object model that could serve as input to the DDI 2 process. This long term strategy was shared with FASTER and it was decided that the part of this model that was related to the structure of the data should be taken care of by FASTER. LIMBER should on the other hand focus on the part of the model that was relevant for human readable descriptions of data elements as well as resource location and classification.

Both of these tracks have been pursued with success. All short term proposals have been accepted for DDI 1.1. The longer term modelling efforts has ended with a model that will be delivered to the next full DDI 1.0 committee meeting.

# PART 4 Metadata for social science data resources: The DDI

Within the academic sector social science data archives and data libraries have been established to provide researcher and students with data for secondary analysis. Some of these institutions have been in existence for 2-3 decades and house the largest collections of accessible computer-readable data in the social sciences in their respective countries. The primary goals of the archives and libraries have been to safeguard the data and to make them as easily accessible as possible for teaching and research independent of whether the users are able to pay for the services or not.

The social science data archives are rarely engaged in the collection of primary data, but serve as brokers between various data providers and the academic community. Their holdings contain data from the public sector (statistical agencies, central government etc), the commercial sector (opinion and market research companies) and academic research. The archives do not only preserve data for future use but also add their own value to the collections:

- Data received by the archives goes through a variety of checks and cleaning procedures to ensure their integrity.
- Any system or software dependency are stripped away to make sure that data can be read at any time in the future.
- Comprehensive computer-readable metadata are developed.
- Data from various sources are often integrated and harmonised in order to produce easy-to-use information products (on-line databases, CD-roms etc.).
- Data are catalogued and made accessible through electronic search and retrieval systems.
- In order to encourage the use of statistical data among students, teaching packages and interactive statistical laboratories, are developed.

Due to the extensive refinements of the data sources, as well as a long-standing reputation of responsiveness to users' needs, data and related services from the archives are frequently requested by non-academic users. This includes users from the public sector, as well as from the mass media and private companies. To the extent that services to non-academic users do not run counter to the agreements with the data depositors, access is usually granted.

The characteristics of the user communities go a long way to explain the high priority that the archives have given to the development of metadata:

- Users of archived data have rarely been engaged in the creation of a dataset.
- Archived data will frequently be used for other research purposes than intended by the creators (secondary analysis).
- Archived data will frequently be used many years after they were created.
- Academic users are often comparing and combining data from a broad range of sources (across time and space).

The common denominator of the four characteristics is an emphasis on the relative distance between the end users of a statistical material and the production process. Whereas the creators and primary users of statistics might possess "undocumented" and informal knowledge, which will guide them in the analysis process, secondary users must rely on the amount of formal meta-data that travels along with the data in order to exploit their full potential. For this reason it might be said that social science data are only made accessible through their metadata. Without human language description of their various elements, data resources will manifest themselves as more or less meaningless collections of numbers to the end users. The metadata provides the bridges between the producers of data and their users and convey information that is essential for secondary analysts.

The metadata is also the pivotal point for any resource discovery system (paper-based as well as digital). Academic users will frequently shop for the most relevant data that might be used to shed light on a topic, substantiate a theory, test a hypothesis etc. Although basic catalogue information might provide information about the overall content of a data-source, detailed metadata on source- as well as variable-level are needed to increase the precision of the resource discovery process. This includes detailed information about concepts and definitions, methodologies and procedures, exact and complete question text (for survey based studies), relationships to other sources and studies, etc.

The conveying of the meaning of data through links to the production process and the facilitation of efficient and high-precision resource discovery can consequently be seen as the two most important reasons for the social science data archives to engage in metadata development.

A third reason, which also should be mentioned, is the potential role that metadata might play as a bridge between the data, the users and the intellectual production of the users. By embedding references and hyperlinks to the reports and scientific studies written on the basis of a dataset, as well as references and links to the researchers and institutes that have been responsible for the research, the metadata might become an important communication node and a vehicle in the process of knowledge accumulation. Too often the dataset or the table report is seen as the terminal station of the statistical production line. By including the (secondary) research activities in this process, an extension of the metadata concept along the lines described above, seems quite reasonable.

Over the years many initiatives have been taken within the data archive movement to create metadata standards. None of these have, however, reached the level of acceptance that is needed for a standard to be successful. The majority of social science data archives have documented their holdings according to a standard study description agreed in the mid 1970's by an international committee of data archivists. Unfortunately many local "dialects" of this standard have evolved and the archives have adapted their metadata holdings to fit the requirements of different storage and retrieval systems. As a consequence the level of standardisation across archives is rather low.

In order to improve this situation, a new international committee, the Data Documentation Initiative (DDI) was established in 1995 to create a universally supported metadata standard for the social science community. The committee was initiated and organised by the Inter-University Consortium for Political and Social Research (ICPSR). The members were coming from social science data archives and libraries in USA, Canada and Europe and from major producers of statistical data (like the US Bureau of the Census, the US Bureau of Labour statistics, Statistics Canada and Health Canada). Information about the work of the DDI-committee can be found at: http://www.icpsr.umich.edu/DDI/.

The original aim of the DDI was to replace the old-fashioned and obsolete standard study descripton with a more modern and Web-aware format. The first version of the new standard was consequently expressed as an SGML DTD. In 1997 it was translated to XML (Nielsen, Jan (1997) "From OSIRIS to XML. Markup and Internet Presentation of Structured Data Documentation". Unpublished thesis.) where it have stayed since. This was just a few months after the World Wide Web Consortium (W3C) released the very first working draft for this new language which according to the visions of the creators would add a new dimension to Web-publishing, especially related to resource discovery and metadata.

An XML DTD (Document Type Definition) provides the rules for applying XML to a document of a specific type. The DTD defines the elements that the document is composed of, the attributes of these elements, and their logical relationships to other elements. The DDI-tree contains five main branches, or sections:

- **The document description,** which describes the metadata document and the sources that have been used to create it (this section can thus be looked upon as a kind of metadata for the metadata , or mete-metadata if you like).
- **The study description,** which contains information about the entire study or data collection (content, collection methods, processing, versioning, sources, access conditions etc).
- **The file description,** which describes each single file of a data collection (formats, dimensions, processing information, missing data information etc.)
- **The variable description,** which describe each single variable in a datafile (format, variable and value labels, definitions, question texts, imputations etc.)
- **Other Study-Related Materials,** which can include references to reports and publications, other machine readable documentation that is relevant to the users of the study (referenced by URI’s) etc.

Each of these main branches is divided into a finer hierarchy of sub-branches. A graphical description of parts of the methodology branch, which is a sub-branch of the study description branch is shown below:

```
2.3 method* (ATT == ID, xml:lang, source)
     ---- 2.3.1 dataColl? (ATT == ID, xml:lang, source)
     ---- 2.3.1.1 timeMeth* (ATT == ID, xml:lang, source, method)
      ---- 2.3.1.2 frequenc* (ATT == ID, xml:lang, source, freq)
      ---- 2.3.1.3 sampProc* (ATT == ID, xml:lang, source)
      ---- 2.3.1.4 deviat* (ATT == ID, xml:lang, source)
      ---- 2.3.1.5 collMode* (ATT == ID, xml:lang, source)
      ---- 2.3.1.6 resInstru* (ATT == ID, xml:lang, source, type)
      ---- 2.3.1.7 sources? (ATT == ID, xml:lang, source)
```

As an example of the use of the DDI-standard, the timeMeth element (2.3.1.1 in graph above) is supposed to contain a description of the time method or time dimension of the data collection. The element text can be used to give a human language description of the method, whereas the method-attribute can include a controlled vocabulary, which more easily can be understood by a software system. An example of how this structure can be used in concrete mark-up is shown below:

```
 <method><dataColl><timeMeth method=’panel’>The study is a panel
survey where 50% of the sample are replaced at each
subsequent.....etc. </timeMeth></dataColl></method>
```

The example demonstrates one of the basic strengths of XML – the ability to bridge the gap between *human* and *machine* readability. The language itself does, however, not guarantee that a marked-up document can be read and understood as easily by a digital process as by a human being. In order to achieve this, the structure of the document (the DTD) must be designed with this dual aim in mind.

The choice of XML as a platform for the DDI standard seems to have been the right one. Since the freezing of the XML specification in February 1998, we have witnessed an explosive growth of web-oriented XML technology. Even more important, an increasing number of organisations, communities, domains and sciences are looking towards XML as a platform for domain-specific mark-up as well as vendor-neutral data exchange

The take-up of the proposed DDI-standard among the community of data archives and libraries has so far exceeded the expectations. At an increasing number of sites from all parts of the World efforts are being made to convert existing holdings of metadata to the new standard. The work of the DDI have also served as an instrument to revitalise the co-operation and sharing of know-how among the archives, as well as strengthening the ties to the data producers. The process of developing software that supports the new standard is also well under way.

Several shortcomings of the first release of the standard have been identified. The most important of these shortcomings are:

- **The survey data bias**. The DDI is well suited as a metadata standard for micro- or survey data, but lack the necessary elements and structure to describe more complex data like multidimensional cubes and time-series etc.
- **The lack of modularity**: The DDI is a comprehensive and detailed structure that allows data providers to describe any thinkable aspect of their data. The approach is however monolithic and the DDI cannot easily be broken into smaller modules to be combined more freely.
- **The lack of extensibility**: The DDI standard has no extensibility mechanism. There is no provision for development of local extensions like Dublin Core qualifiers.
- **The inflexibility of the DTD approach**: When the DDI made its move from SGML to XML there were no alternatives to the DTD. Later on XML Schema has entered the scene in addition to more dedicated metadata languages like RDF. Both of these are providing a more flexible and extensible framework than the document centric DTD.

Some of these shortcomings have already been addressed or in the process of being solved. Most progress has been made regarding the extension of the DDI to support more complex data types. A cube model has been added to the latest release partly developed in cooperation between the LIMBER and the FASTER project. The FASTER and LIMBER group has also delivered input to the DDI committee regarding a potential move from an XML DTD to a RDF or XML Schema. Additionally, a set of minor changes and additions regarding the interface between DDI and thesauri/controlled vocabularies has been delivered by representatives from the LIMBER project.

More about the division of labour between FASTER and LIMBER when it comes to metadata development can be found below.

# Part 4: Other Metadata Standards

## 1. Introduction

During the last years we have seen the emergence of a vast number of standards for describing both statistical- and metadata. Some of these standards are in scope regarded as too limited for our purpose, or simply outside our exploration area, typically standards like GILES, (US) MARC, CIMI, INDECS, etc. The primary objective of this part of the deliverable is to pull out and introduce standards which describes metadata in an appropriate manner. This claim could be assessed as at least a prerequisite whilst digging in the material.

In the following section we will introduce statistical (metadata) standards which can be appraised as alternatives, or even competitors, to the DDI standard framework at different levels.

We will subdivide this part of the deliverable into three parts. In the first section we will introduce two standards which have the ambition of being generic and, more or less, domain-independent. First, the Common Warehouse Metamodel (CWM), developed by the Object Management Group. Secondly, the ISO/IEC 11179, which the International Organization for Standardization and the International Electrotechnical Commission have developed as the main contributors. These two standards are of main interest for the development of metadadata models. Inside this partition wall, CWM would be treated being at a higher generic level than the ISO standard. Thus, ISO is more operationalised than the former.

In the second section we take a closer look at two real-world examples which highlights the very important concept of interoperability with relevant and central tools according to the LIMBER efforts.

In the third part of the paper, we will introduce standards used for different domains and purposes. These are Dublin Core, DDI DTD, idaresa - integrated documentation and retrieval environment for statistical aggregates, and GESMES – Eurostat (EDI), abbr. for Generic Statistical Message.

## 2. OMG – Object Management Group

In this first section we will introduce two standards treated as frameworks, i.e. principles of how one should build up a metadata standard. They are domain-independent, because they are describing ways of structuring and building up different standards, not the actual application; how it should be used for a special domain of data. The first of these is the Object Management Group (OMG). They are described as an effort to create a metamodel for easy interchange of warehouse metadata between warehouse tools, warehouse platforms and warehouse metadata repositories. Data Warehousing provides an excellent approach for transforming data into useful and reliable information to support the business decision process. One of the most important aspects of datawarehousing is metadata. The term “metadata” refers not only to the set of definitions of the data in the warehouse products, parts, prices, and so on but also to its formats, processing, transformations, and routing from origin to warehouse.

| Meta-level | MOF terms | Examples |
|---|---|---|

| | | |
|---|---|---|
| M3 | meta-metamodel | The “MOF Model” |
| M2 | Metamodel, meta-metadata | UML Metamodel,<br>CWM Metamodel |
| M1 | model, metadata | UML models,<br>CWM metadata |
| M0 | object, data | Modeled systems,<br>Warehouse data |

*Table1 OMG Metadata Architecture*

Thus, metadata is used for building, maintaining, managing, and using the data warehouse, e.g. everything, that is, except the data elements themselves in the data warehouse. Metadata management, and reconciliation of inconsistent metadata when data from different sources are merged, are the biggest problems facing enterprises working with data warehousing today.

Since there are many kinds of metadata in a typical system, the *Meta Object Facility* (MOF) framework needs to support many different MOF metamodels. The MOF integrates these metamodels by defining a common abstract syntax for describing metamodels. In the table above we are able to identify the MOF metadata framework through its four layers architecture. The MOF has been adopted as OMG`s standard for representing metamodels.

The CWM, abbr. for *Common Warehouse Metamodel* , has been designed conforming to this standard. This allows CWM to use other OMG specifications that are dependent on the MOF. In particular, it allows use of XMI to interchange warehouse metadata that is represented using the CWM metamodel. OMG’s Common Warehouse Metamodel or CWM provides a standard solution to this problem. The main purpose of CWM is to enable easy interchange of warehouse metadata between warehouse tools, warehouse platforms and warehouse metadata repositories in distributed heterogeneous environments. Building on three existing industry standards the OMG’s Unified Modeling Language (UML), the eXtensible Markup Language (XML), and OMG’s XML Metadata Interchange (XMI) the CWM starts by establishing a common metamodel for warehousing but then goes beyond this to also standardize the syntax and semantics needed for import, export, and other dynamic data warehousing operations.

The – MOF – is an OMG interface standard that can be used to define and manipulate a set of interoperable metamodels and their instances (models). The MOF also defines a simple meta-metamodel (based on the OMG UML) with sufficient semantics to describe metamodels in various domains. The CWM specification uses MOF as the meta-metamodel[1].

UML, the Unified Modeling Language, is an OMG standard modeling language for specification, construction, visualisation, and documentation of the artifacts of a software system. XMI, XML Metadata Interchange, is an OMG standard mechanism for the stream-based interchange of MOF-compliant metamodels[2]. Designed to work naturally with object,

[1] The CWM specification uses the UML notation as the graphical representation for its metamodel and reuses the UML where appropriate. All CWM object types are direct or indirect subtypes of appropriate UML object types, and so inherit their attributes and associations. This approach allows the CWM specification to capitalise on the substantial investment in developing and refining the UML metamodel. Further, it enables easy inclusion of UML models as part of the data warehouse metadata. The CWM Metamodel is designed, it is in fact one of its major design goals, to maximise the reuse of UML and the sharing of common modeling constructs where possible. The most prominent example is that CWM reuse/depends on UML for representing object-oriented data resources.

[2] XMI is based on XML and has two major components: First, the XML DTD Production Rules for producing XML DTDs for XMI encoded metadata, and secondly, the XML Document Production Rules for encoding metadata into an XML compatible form. The purpose of XMI is to allow the interchange of models in a serialised form. Since the MOF is the OMG`s adopted technology for representing metadata, it is natural that XMI focuses on the interchange of MOF metadata; i.e.

relational, record-based, multidimensional, and XML-based datastores, the CWM supports data mining, transformation, OLAP, information visualisation, and other end user processes. Metamodel support encompasses data warehouse management, process, and operation. The CWM specification extends to application programming interfaces (APIs), interchange formats, and services that support the entire lifecycle of metadata management including extraction, transformation, transportation, loading, integration, and analysis. And, users can resolve specific integration issues by taking advantage of the CWM metamodel's built-in extensibility.

CWM is a domain-specific extension of the OMG`s Metamodeling Architecture, and as such, implicitly supports the MOF, UML and XMI standards. Although CWM has certain "compatibilities" with various other standards, these should be regarded as touch points of mapping or integration; they do not represent dependencies of any kind. Thus, CWM is not dependent upon any standards outside those of the OMG Metamodeling Architecture.

| Management | Warehouse Process | | | Warehouse Operation | | |
|---|---|---|---|---|---|---|
| Analysis | Transformation | | OLAP | Data Mining | Information Visualization | Business Nomenclature |
| RESOURCE | Object-oriented UML | RELATIONAL | Record | Multidimensional | | XML |
| Foundation | Business Information | Data Types | Expression | Keys and Indexes | Type Mapping | Software Deployment |
| UML 1.3 (Core, Common_Behavior, Model_Management) | | | | | | |

*Table2 The CWM Metamodel*

The CWM metamodel is organized into 18 packages arranged in four layers on a UML base (see table above). The metamodel breaks new architectural ground by defining its sub-metamodel as individual packages. Because CWM uses modeling techniques that minimise the number of dependencies between its packages, tool integrators can select only those metamodel services they need while avoiding problems common to large, monolithic metamodels[3].

The four layers of the metamodel collect different sort of packages:

- The Foundation layer contains several general services that are shared by other packages.
- The Resource layer contains data models use for operational data sources and target data warehouses.
- The Analysis layer provides metamodels supporting logical services that may be mapped onto data stores defined by Resource layer packages. For example, the Transformation metamodel supports the definition of transformation between data warehouse sources and targets, and the OLAP metamodel allows data warehouses stored in either relational or multidimensional data engines to be viewed as dimensions and cubes.

metadata conforming to a MOF metamodel. In fact, XMI is really a pair of parallel mapping: one between MOF metamodels and XML DTDs, and another between MOF metadata and XML documents. XMI can be viewed as a common metadata interchange format that is independent of middleware technology. Any metadata repository or tool that can encode and decode XMI streams can exchange metadata with other repositories or tools with the same capability.

[3] One main objective among the developers is that similar modeling architectures can be leveraged to reduce the complexity of other monolithic models, for instance UML itself.

- The Management layer metamodels support the operation of data warehouses by allowing the definition and scheduling of operational tasks (Warehouse Process package) and by recording the activity of warehouse processes and related statistics (Warehouse Operation package).

As mentioned above, UML is the modeling foundation on which the complete CWM is built. Whenever possible, the CWM developers have directly reused existing UML classes and associations rather than creating CWM-specific versions of them. This choice both reduces the number of new CWM classes and associations and exploits the existing skills of UML-knowledgeable modelers[4].

CWM combines the power of enterprise data management and object modeling, making them available to data modelers, database designers, data warehouse users and administrators, and corporate portal developers and managers.
The OMG Modeling documents describe the OMG standards for modeling distributed software architectures and systems along with their CORBA Interfaces. There are two complementary specifications:

- UMLanguage Specification
- MOFacility Specification

The former defines a graphical language for visualising, specifying, constructing, and documenting the artefacts of distributed object systems. The specification includes the formal definition of a common Object Analysis and Design (OA&D) metamodel, a graphic notation, and a CORBA IDL facility that supports model interchange between OA&D tools and metadata repositories. The UML provides the foundation for specifying and sharing CORBA-based distributed object models. Thus, the CWM metamodel is described in UML and can be thought of as an extension specialising UML for data warehousing applications.

The latter defines a set of CORBA IDL interfaces that can be used to define and manipulate a set of interoperable metamodels and their corresponding models. These interoperable metamodels include the UML metamodel, the MOF metamodel, as well as future OMG adopted technologies that will be specified using metamodels. The MOF provides the infrastructure for implementing CORBA-based design and reuse repositories. The MOF specifies precise mapping rules that enables the CORBA interfaces for manipulating metadata in all phases of the distributed application development style.

Thus, the MOF goal is to allow interoperability across the application development cycle by supporting the definition of multiple metamodels, whereas the OA&DF focuses on supporting the definition of a single OA&D metamodel.

## 3) ISO/IEC11179: Specification and Standardization of Data Elements

The International Organization for Standardization (ISO) is a worldwide federation of national standards bodies from some 130 countries, one from each country. ISO is a non-governmental organisation established in 1947. The mission of ISO is to promote the development of standardisation and related activities in the world with a view to facilitating

[4] For example, the Object-Oriented package in the Resource layer is really just a reuse of existing UML classes that are already sufficient for describing object-oriented data sources, such as object-oriented DBMSs or application systems built in object-oriented languages like Java.

the international exchange of goods and services, and to developing co-operation in the spheres of intellectual, scientific, technological and economic activity.

ISO and IEC (the International Electrotechnical Commission) form the specialised system for worldwide systems. In the field of information technology, ISO and IEC have established Joint Technical Committee 1, ISO/IEC JTC1 which have been developing the standard.[5]

ISO/IEC 11179 describes the standardising and registering of data elements to make data understandable and shareable. The purpose of ISO/IEC 11179 is to give concrete guidance on the formulation and maintenance of discrete data element description and semantic content/metadata that shall be used to formulate data elements in a consistent, standard manner. Additionally, it provides guidance for establishing a data element registry (see below).

ISO/IEC 11179 seeks to describe the generic aspect of data. Data elements are the fundamental units of data. A data element is by definition a unit of data for which the identification, meaning, representation and permissible values are specified by means of a set of attributes. A precise, well-formed definition is one of the most critical requirements for shared understanding of a data element; well-formed definitions are imperative for the exchange of information.

For the purpose of ISO/IEC 11179, a data element is composed of the following three parts: an object class, properties, and representation. A definition is a natural language statement of the data element`s meaning; its predicate. The definition of a data element is an extremely critical aspect of data element development. To be shareable, a data element must have a well-formed, unambiguous, precise and commonly understood definition. A property has a definition and belongs to an object class. Conversely, a representation has no definition but has a format category, permissible data values, maximum character count, and, if a measurement, a unit of measure.

The combination of an object class and a property is a data element concept (DEC). A DEC can be represented in a form of a data element, described independently of any particular representation. In practice, the difference between a property and an object class is often not absolute. Thus, a data element can be seen to be composed of two parts: a data element concept and a representation. One fundamental aim in this standard would be that an element must be suitable for communication, interpretation, or processing by human or automated means. In an object model, data elements are expressed as object attributes. Further, these models provide one kind of classification scheme for the pertinent data elements.

The procedures and techniques specified in the International Standard will enable Registration Authorities to apply classification schemes that better enable one to

- analyse object classes, data element concepts, and data elements[6]

[5] Draft International Standards adopted by Joint Technical Committee 1 are circulated to national bodies for voting. Publication as an International Standard requires approval by at least 75% of the national bodies casting votes. ISO standards are developed according to the following principles:

- **Consensus**: The views of all interests are taken into account: manufacturers, vendors and users, consumer groups, testing laboratories, governments, engineering professions and research organisations.
- **Industry-wide**: Global solutions to satisfy industries and customers all over the world.
- **Voluntary**: International standardisation is market-driven and therefore based on voluntary involvement of all interests in the market-place.

[6] One particular part of OSI/IEC 11179 develops a set of principles, methods, and procedures for specifying what is needed, at a minimum, in a taxonomy/ontology for description of object classes,

- make comparisons within the following categories: object classes, properties, representations, data element concepts, and data elements
- reduce the variety of data element concepts and data elements
- identify, describe, and define data element concepts and data elements unambiguously
- assist in the analysis of data elements for the purposes of assigning registration status
- address synonym and homonym problems
- retrieve data element concepts and data elements from a data register
- recognise relationships among data element concepts and data elements
- support the unique and unambiguous identification and referencing of object classes, data element concepts, and data elements in a manner that is linguistically neutral and information technology enabled[7].

A group appointed the responsibility for registering data elements is called a Registration Authority (RA). It assigns a unique identifier, assures that all required metadata attributes are documented, and assign a status level depending upon the metadata quality and degree of integration. The RA is appointed to be responsible for a universe of discourse. This universe may be as narrow as an organisational sub-unit, or as an entire nation. The RA is responsible for the registration of any data element considered shareable within the universe of discourse for which it is appointed. The RA documents the description of data elements for which it is responsible in a Registry. The RA is responsible for the integrity of the data elements in the Registry. Every organisation wishing to become a RA shall possess an internationally recognised organisation code, assigned in accordance with the procedure prescribed in ISO/IEC 6523[8].

Every submitting organisations (SO) wishing to register data elements shall be able to do so in accordance within determined procedures. Applications to a RA for the registration of data elements shall be made by its SOs, after having consulted, as appropriate, with their corresponding Responsible Organizations (RO). RA shall decide whether an application is acceptable. After verification by the RA, the RO, and the SO, the Registration Status of the acceptable proposal shall be marked as “certified” or “standardised”, as appropriate. A data identifier is a generic tool exploring a data element. The identifier is assigned by the RA and would be concatenated by a combination of registration authority identifier, data identifier, and version identifier. Data element names can be formed from the names of components, each assigned meaning (semantics) and relative or absolute position (syntax) within a name.

Basic Attributes[9] of Data Elements specifies attributes of data elements. Some rules building up the structure of a data definitions: a data definition shall be unique, be stated in the singular, state what the concept is, not only what it is not, be stated as a descriptive phrase or sentence(s), contain only commonly understood abbreviations, and be expressed without embedding definitions of other data elements or underlying concepts. A data definition is guided by some general principles. It should state the essential meaning of the concept, be

property, representation, data element concepts, and data elements (herein called “information elements”).

[7] According to the (general) evaluation criterion-proposal, sent by Brian Matthews in mid-Dec, these items are highly relevant.

[8] For a thorough debate regarding this particular standard as a vehicle for RAI assignment, see the Annex C, i.e. FAQ, in part 6 of the Draft International Standard.

[9] The following attributes of data elements are mandatory: name, definition, representation category, form of representation, data type of data element values, maximum size of data element values, minimum size of data element values, permissible data element values.

precise, concise, unambiguous, and able to stand alone, be expressed without embedding rationale, functional usage, domain information, or procedural information, avoid circular reasoning, and use the same terminology and consistent logical structure for related definitions. It is limited to a set of basic attributes for data elements, independent of their usage in application systems, databases, data interchange messages, etc. The basic attributes are established to guarantee a shared view of the data elements, e.g. to increase the machine-understandability[10]. The concept means that they are essential in specifying a data element completely enough to ensure that it will be applicable for a variety of functions, such as:

- design of information processing systems;
- retrieval of data from databases;
- design of EDI-messages for data interchange;
- maintenance of data element dictionaries;
- data management;
- dictionary design;
- dictionary control;
- use of information processing systems

Metadata about data elements is stored in a data element registry. A data element registry supports data sharing with description of data. Registration is the process of documenting metadata to support data shareability. Registration should be carried out at the data element level to promote and maximise semantic value. ISO/IEC 11179 enables the end user to interpret the intended meaning confidently, correctly, and unambiguously. Use of one or more classification schemes is intended to provide a sound conceptual basis for the development of metadata having enhanced semantic purity and design integrity.

For users and managers of data, ISO/IEC 11179 specifies a basic set of data element characteristics necessary to share data. It places special emphasis on important element characteristics such as identifiers, definitions, and classification categories. ISO/IEC 11179 describes a data element registry to assist users of shared data to have a common understanding of a data elements meaning, representation, and identification. If data values are received, the user can discover the exact meaning of the data received. If users wish to retrieve data values from a database, they can identify the type of data desired.

## 4) Two real-world examples highlighting the concept of interoperability:

There are three main scenarios in which interoperability among metadata descriptions are required:

- To enable a single search across different and to some extent heterogeneous metadata descriptions;
- To enable the integration or merging of descriptions which are based on complementary but possibly overlapping metadata schemas or standards;
- To detect the common underlying model used by different metadata descriptions.

[10] The basic attributes specified are applicable for the following main activities: Definition and specification of the contents of data element dictionaries, design and specification of application-oriented data models, databases, and messages for data interchange, actual use of data in communications and information processing systems, and interchanging or referencing among various collections of data elements. A set of five related attributes serves to name and identify each data element for the purpose of differentiating data elements. These attributes are as follows: name, context, registration authority-, data-, and version identifier.

In this section we will the results of two twin-papers dealing with the concept of interoperability. The first is relevant according to the DDI discussion because it seeks solving the problem of integrating different standards into one representation. The latter directs its efforts to extract the best from two LIMBER-relevant tools: XML- and RDF Schema trying to find a solution to the problem of diversity and disparity.

## MetaNet[11]

Showing the importance of interoperability, the MetaNet paper focus the problem of hybrid mapping which seeks to combine the structural and syntactic mapping capabilities with tools familiar to the LIMBER project; XML and RDF[12]. More explicitly, the objective is to demonstrate how the semantic knowledge can be represented in a machine readable format (RDF Schema) and extracted with the syntactic and structural mapping capabilities of XSLT to enable the implementation of flexible dynamic mappings between different metadata descriptions within more or less the same domain[13].

The anticipation is that XSLT's ability to transform the data from one XML representation to another makes it appear to be ideal for metadata interchange applications. Nevertheless, the mapping implementation revealed that although XSLT works very well for performing the structural mapping from an event model to a resource-centric model[14], it is inadequate for implementing flexible dynamic semantic mapping between metadata vocabularies. The mapping revealed that if the input XML descriptions are relatively fixed and tightly constrained, then the semantic mappings can be hardwired and XSLT is adequate. But if the input XML descriptions are variable or unpredictable then XSL cannot cope.

The second approach seeks simply linking a matrix to the XSLT processor. As anticipated, this is not an ultimate solution as long as it does not scale, as the number of domains grow and the mappings become asymmetrical, then the matrix becomes excessively complex and unwieldy.

The third approach involves extracting the mapping dynamically from a thesaurus of metadata terms, generated by formally defining relationships between metadata terms from a number of different domains' standardized vocabularies[15]. The program flowchart, using Java, is outlined in figure below.

[11] MetaNet is a metadata term thesaurus, which provides the additional semantic knowledge required to enable semantic mapping between metadata terms from different domains or standards.

[12] Other different proposals have been developed for improving interoperability between domain-specific vocabularies, thesauri and ontologies. These are ranging from database schema integration, to the use of ontologies in organising and integrating networked information systems, to the merging of monolingual and multilingual thesauri. More recently the approach to merging thesauri has been to represent them formally using RDF Schemas and to use inference engines to automate the merging.

[13] The concrete example concerns music.

[14] The authors tried to transform the ABC model to DC, ID3 and MPEG-7 descriptions respectively. The ABC model defines a set of fundamental classes which provide the building blocks for expression, through sub-classing, of application-specific or domain-specific metadata vocabularies.

[15] The MetaNet contains preferred terms, equivalent terms, narrower terms and broader terms and attempts to encompass terms from the most significant and widely used metadata vocabularies (DC, IFLA, IEEE LOM, INDECS).

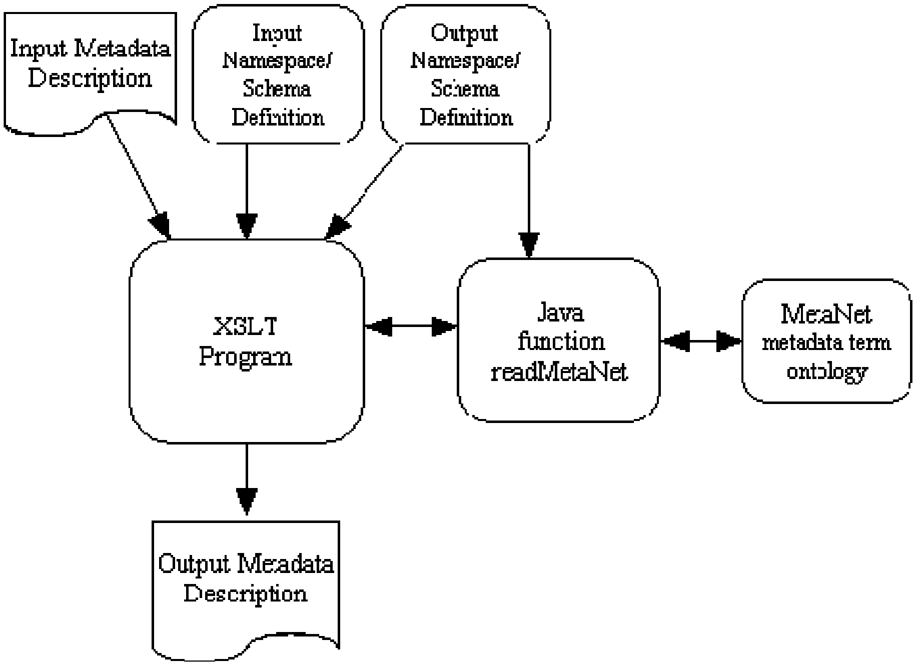

Input Metadata Description
Input Namespace/ Schema Definition
Output Namespace/ Schema Definition
XSLT Program
Java function readMetaNet
MetaNet metadata term ontology
Output Metadata Description

**Figure 1 Program Flow for Metadata Description Mappings**

In conclusion, the paper accentuate the shortcomings of XSLT (see above), further the limitations in the proposed MetaNet thesaurus solution, for instance that it is currently solely supporting one language (English). A situation which will become increasingly common in the future is the need to map from a schema which imports elements from multiple namespaces to another schema which imports a different set of elements from multiple namespaces. The author highlights her solution as being able to bridge this gap. Nevertheless, the major argument according too our approach is the kind of problem that rises when the standardisation level is low and a uniform description standard is lacking.

## Metadata Application Profiles[16]

The paper describes a hybrid collaborative approach combining the semantic knowledge of RDF Schemas with explicit structural, cardinality and datatyping constraints provided by XML Schemas in a complementary manner. More concrete, the paper, with use of examples and implementations, attempts to demonstrate how the two schema languages could be made work together, enabling flexible, dynamic mapping between complex, metadata descriptions which mix elements from multiple domains, i.e. application profiles. The authors propose that both metadata diversity and interoperability can more easily be accommodated across the web if each metadata domain defines both a RDF Schema and a XML Schema for their domain in their registered namespace[17].

[16] Full title*: Combining RDF and XML Schemas to Enhance Interoperability Between Metadata Application Profiles.*

[17] By expressing the semantic knowledge of each domain in a machine-understandable RDF Schema, it then becomes possible to merge these separate domain ontologies or vocabularies into a single encompassing ontology or vocabulary, also expressed as a RDF Schema, known as the MetaNet ontology.

There are two alternative schema languages for defining application profiles: RDF Schema and XML Schema[18]. The advantage of the former is that it provides rich semantic descriptions but limited support for the specification of local usage constraints, i.e. structural, cardinality and datatyping constraints. The advantage /disadvantage of the latter is the opposite of the former. Thus, the authors accentuate that the most logical approach is to use both Schemas for exploitation of their complementary features. An apparent shortcoming of this hybrid approach would be that there are no currently clearly defined mechanisms for smoothly and cleanly meshing RDF Schema and XML Schema definitions (Hunter and Lagoze, 2001, p.5).

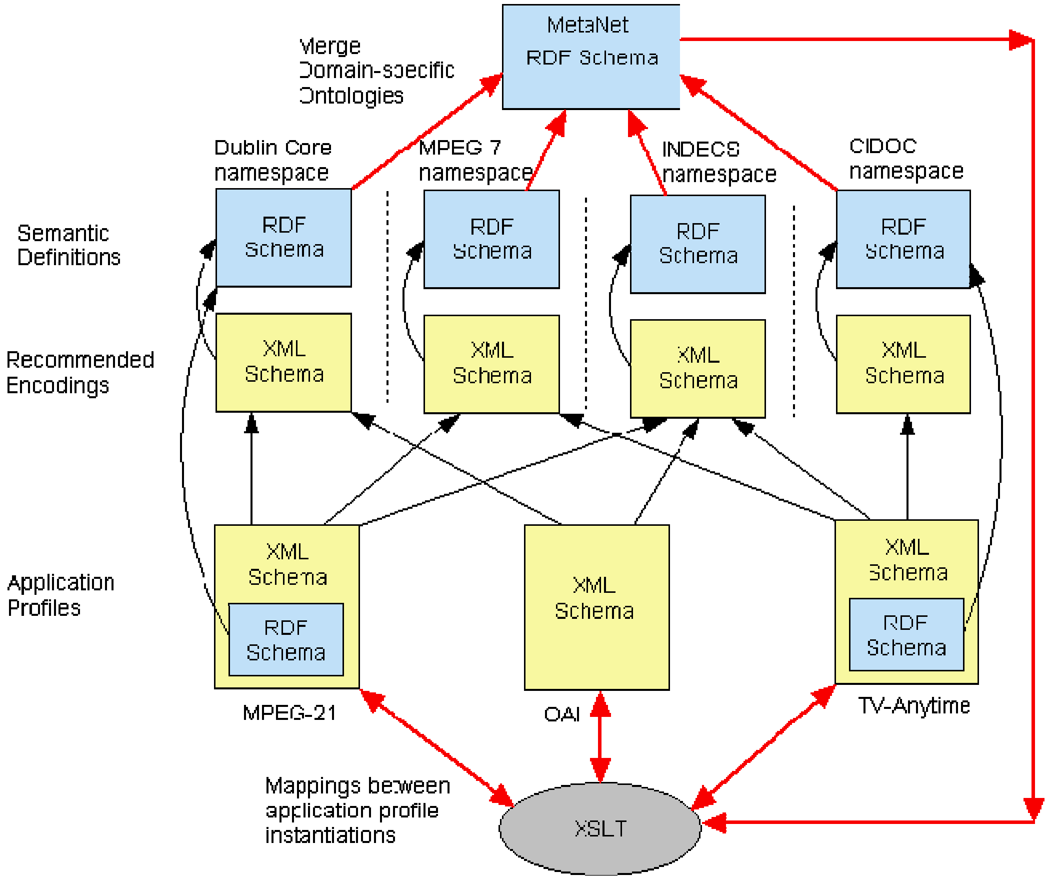


**Figure 2 Example of the Proposed Web Metadata Architecture**

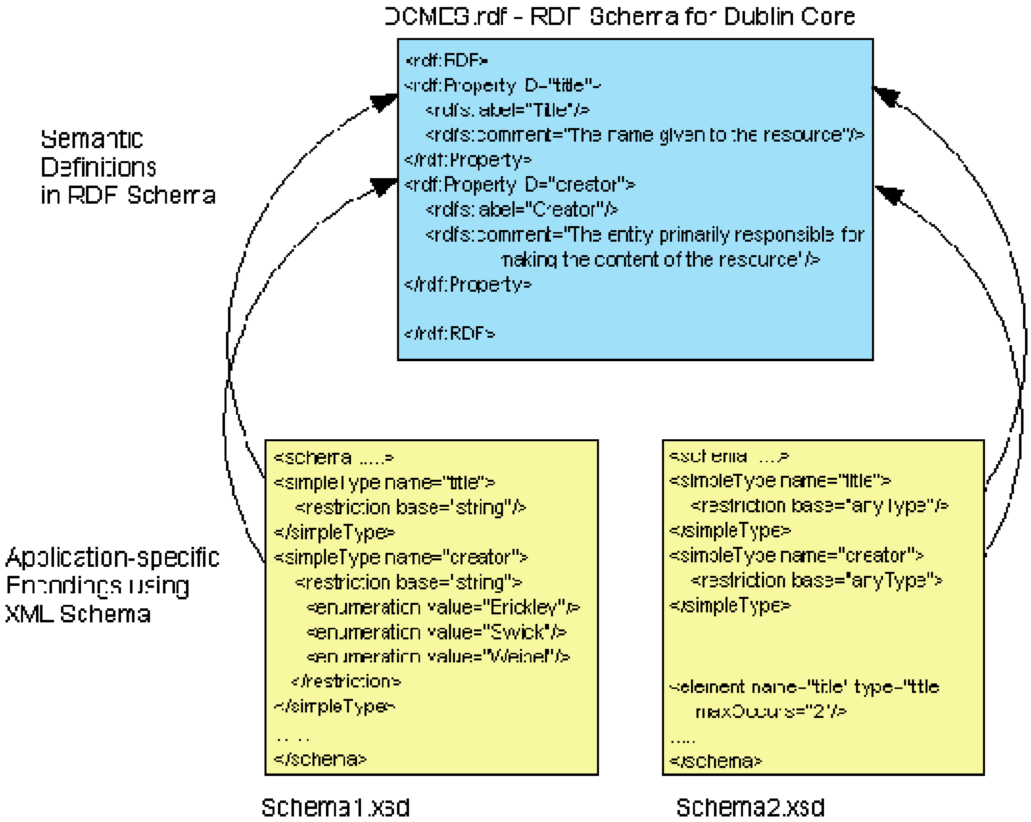
There are two methods for combining the two RDF Schema semantics with XML Schema local constraints:

1. Embedding the RDF Schema *Class/subClassOf*, *Property/subPropertyOf* definitions inside type annotations in the XML Schema file, or
2. Adding links from the XML Schema to an external RDF Schema file, see figure below.

The first approach has the advantage of combining both the semantic definitions and structural and syntactic constraints in a single file, whilst maintaining XML Schema conformance. However, the major limitation is more prevalent: Those RDF classes and properties defined explicitly within XML Schema annotations are solely local definitions and

[18] The authors exclude XML DTD as an alternative because it does not explicitly support the facility of namespaces.

cannot be reused or pointed to by other schemas because they are not globally-accessible named elements.


DCMES.rdf - RDF Schema for Dublin Core
Semantic Definitions in RDF Schema
Application-specific Encodings using XML Schema
Schema1.xsd
Schema2.xsd

```
<rdf:RDF>
<rdf:Property ID="title">
   <rdfs:label="Title"/>
   <rdfs:comment="The name given to the resource"/>
</rdf:Property>
<rdf:Property ID="creator">
   <rdfs:label="Creator"/>
   <rdfs:comment="The entity primarily responsible for
             making the content of the resource"/>
</rdf:Property>

</rdf:RDF>
```

```
<schema .....>
<simpleType name="title">
   <restriction base='string"/>
</simpleType>
<simpleType name="creator">
   <restriction base='string">
     <enumeration value="Brickley"/>
     <enumeration value="Swick"/>
     <enumeration value="Weibel"/>
   </restriction>
</simpleType>
.. ..
</schema>
```

```
<schema ....>
<simpleType name="title">
   <restriction base="anyType"/>
</simpleType>
<simpleType name="creator">
   <restriction base="anyType">
</simpleType>


<element name="title' type="title
   maxOccurs="2"/>
.....
</schema>
```

**Figure 3 Linking from Multiple XML Schema Definitions to a Common Base RDF Schema**

The authors did not find any optimal solution for linking the complementary functions of RDF- and XML Schema according to the second proposal either.

In conclusion, the authors sum up that XML Schema language is ideal for defining application profiles. RDF Schemas are used to express the semantics of domain-specific metadata models in a machine-understandable syntax. However, the efforts towards extracting out the complementary function failed. The suggested approaches for combining the two standards was not an easy task because 1) there is an overlap in functionality, and 2) there is a lack of clearly defined mechanisms or tools for linking the two standards, i.e. the lack of a hybrid parser checking for consistency.

Ideally the XML Schema language would provide an explicit built-in attribute, which is a *uriReference* to the corresponding semantics for that type, i.e. existing classes or properties in an external RDF Schema:

```
    <simpleType name="originator"

semantics="http://purl.org/dc/elements/1.1/dcmes.rdf#creator"/>
       <restriction base="string"/>
    </simpleType>
```

Additionally, the paper discovers that the current extensibility mechanisms for both XML- and RDF Schema are unclear and require clarification, simplification and implementation examples. There is a need for a re-examination of the two schema languages and the formulation of mechanisms which cleanly and smoothly integrate their complementary functionality.

## 5) Dublin Core[19]

In this section we will present relevant standards used by real-world applications. Dublin Core is an attempt at defining a generic standard for resource location. Many communities are interested in adopting a common core of semantics for resource description, and the Dublin Core has received some international and interdisciplinary support for this purpose. To date, the effort has succeeded on the force of the energy and personalities of the cadre of individuals who have invested their intellect and enthusiasm. In addition, expanding acceptance of the Dublin Core will require organisational stability and a formal process for refinement and change. The DCMI-Structure & Operation outlines formal procedures to support the evolution of the Dublin Core. These procedures are, mostly, borrowed from other standards communities. The Core contains just 15 elements, this could be seen as both the strength and the weakness of the approach. The metadata elements are divided into three distinct groups:

- Content: Title, Subject, Description, Source, Language, Relation (to another resource), Coverage (spatial or temporal characteristics of intellectual content)
- Intellectual Property: Creator, Publisher, Contributor, Rights
- Instantiation: Date, Type (such as home page, novel, working paper), Format (of data, to identify software and hardware required for use), Identifier (such as URL or ISBN).

DC is not a metadata standard for social science[20]. It concentrates on resource discovery, thus not covering other requirements, such as resource management or access restrictions. The Core was originally conceived for use by content creators, but interest has become widespread among specialised resource description groups, such as museums and libraries.

Some general shortcomings of the Core:

- Lack of formal responsibility for maintenance: Development takes place in the informal Dublin Core Workshop Series[21].
- The technical state is unstable. Changes are made on an ongoing basis, and stability does not appear to be improving.
- There are no guidelines for use. Differing interpretations will reduce interoperability.

DC enables interoperability across domains. Although the elements in the Core may be a subset of those required for social science or other datasets, they are clearly not a complete set. The role of DC has always been envisaged only to satisfy initial search criteria before more refined searches of metadata using domain specific metadata fields would be undertaken (Matthews, Miller, Ramfos, Ryssevik, Wilson, 2001, p.3).

[19] See part 6 for more

[20] There exists already a recommended mapping between DDI and Dublin Core

[21] The same critique as is addressed to the progress of the DDI standard.

## 6) DDI

The Data Documentation Initiative (DDI) is an effort to establish an international criterion and methodology for the content, presentation, transport, and preservation of metadata about datasets in the social and behavioural sciences. The aim of the DDI is to define a framework of codebooks with an uniform, highly structured format that lends itself to simultaneous use of multiple datasets, and will significantly improve the content and usability of metadata. Further, this specification may have far-reaching implications for improvement of the entire process of data collection, data dissemination, and data analysis[22]. In sum, this will create a new framework for secondary analysis.

The DDI is already in use by major international projects such as the European Networked Social Science Tools and Resources (NESSTAR). NESSTAR strives to provide a "seamless interface" between the user and the data and its documentation, through integrated data discovery, analysis, and dissemination tools. With NESSTAR, users are able to locate multiple data sources in every participating data archive in a single search operation, browse highly detailed structured metadata about these resources, conduct statistical analyses interactively, visualise data statistically and geographically, and download subsets of cases and/or variables in a number of formats. A standard like this is central to the ability of NESSTAR to deliver on these promises across countries and archives.

The ambitions of DDI have lately grown substantially. Codebooks are seen as structuring and supporting the entire data collection, distribution, and analysis process throughout the social and behavioural sciences, and the DDI provides the "glue" that will bring that process together. The DDI see themselves as enabling a new mode of conducting comparative and other research that uses multiple datasets. Thus, we see the DDI as offering, not only data producers and data archivists, but also data users a new power and flexibility to do their work, and to do it effectively and efficiently.

The adoption of the proposed DDI standard within the community of data archives and libraries have been successful. At an increasing number of sites from all over the world, efforts are being made to convert existing holdings of metadata to the new standard. The work of the DDI has also served as an instrument to revitalise the co-operation and sharing of know-how among the archives, as well as strengthening the ties to the data producers (Ryssevik, 1999).

Besides NESSTAR, additionally the CESSDA[23] members employs the DDI XML (eXtensible Mark-up Language) codebook DTD (Document Type Definition) as their metadata standard.

[22] Initially with support from the Inter-university Consortium for Political and Social Research (ICPSR) and then with support from an NSF grant (SBR-9617813) (and considerable in-kind contributions of staff time from institutions all over the world), the DDI committee has produced what is known as a Document Type Definition (DTD) for "markup" of social science codebooks. The DTD employs the eXtensible Markup Language (XML), which is a dialect of a more general markup language, SGML.

[23] Abbr. for the Council of European Social Science Data Archives. The community promotes the acquisition, archiving and distribution of electronic data for social science teaching and research in Europe. It encourages the exchange of data and technology and fosters the development of new organisations in sympathy with its aims. It associates and co-operates with other international organisations sharing similar objectives. In particular the German, French, Spanish and Greek Data Archives are involved in the evaluation of the multi-lingual thesaurus.

The XML DTD provides the rule for applying XML to a document of a specific type[24]. The DDI standard is specifically defined to describe flat single survey data files. However, problems arise when different types of data need to be described, when series of data need to be linked, when different output formats are required and when references to different metadata standards need to be made. All this would require major changes to the DTD, thus making metadata already created invalid against the new format. The LIMBER project have investigated the possibility of changing the DDI codebook to a more modular and extensible format such as RDF or XMLschemas.

# 7) GESMES – Eurostat (EDI)

The compilation of statistics is a continuous chain starting with raw data collection, followed by a number of harmonisation and aggregation processes leading to aggregated statistical data. Raw data collection is realised by means of surveys, sampling, direct reporting from individuals, economic operators, etc. Harmonisation and aggregation processes are often handled by statistical bodies or authorised economic operators in case of direct reporting. A Generic Statistical Message (GESMES)[25] is used by an organisation involved in this elaboration process to transmit a statistical data set. It permits the transmission of the following, either in the same and/or different messages:

- the statistical concepts comprising the data set and their structure,
- if required, all related information[26] (code sets, labels, methodological notes, footnotes, etc.),
- the statistical figures of the data set.

## 7.1 Principles

The structure of statistical (metadata) information to be exchanged could be defined as multidimensional array or chronological series. The difficulty of using or exchanging statistical data comes from the varieties of interpretations of those structures. The interpretation varies from one statistical domain to another depending on the content. Another key factor is the volume: statistical indicators have many dimensions (e.g. time, classifications depending on statistical nomenclatures), dense and sparse series of matrices can be presented simultaneously for the same indicator. The GESMES message may be used to exchange statistical data for all statistical domains in a standardised format together with it descriptions.

The data in the data set (i.e. the array) is contained in a generic segment (ARR - Array). The interpretation of the ARR segment is provided in segment group 8. GESMES is designed to support all types of statistical exchange, including time-series, which can be described by means of statistical concepts (i.e. statistical terms or objects)[27]. Associated with statistical data

[24] The DTD defines the elements that the document is composed of, the attributes of this elements, and their logical relationships to other elements. The elements will usually be arranged in a hierarchical or tree-like structure. The DDI-tree contains five main branches.

[25] The statistical office of the European Union, EUROSTAT, who has lead the development of statistical UN/EDIFACT messages has implemented GESMES into the data flows between it and the Member States of the European Economic Area.

[26] Concerning LIMBER, this point is at the crux of the metadata discussion.

[27] In the GESMES/CB application comprises the array section of the identification of the data set structure (IDE), the identification of the method used to place data values in the ARR segment, an

are textual information defining comprehensively the series and matrices, the underlying statistical concept or methodology[28].

GESMES comprises a number of parts (or segments which it is called in the documentation):

a) Identification of administrative information concerning the interchange partners;
b) Footnotes associated with value lists, statistical concepts and data set structures;
c) Value lists of coded or non-coded items to be used in this message or in a future message;
d) Definition of statistical concepts to be used in this message or in a future message;
e) Definition of the structure of array data to be used in this message or in a future message, and the allocation of a unique identity to this definition;
f) Identification of one or more data sets. For each data set: the identity of the structure definition to be used to interpret the array data; the scope, value lists or fixed values which are local to the data set; the array data (ARR); footnotes associated with the data set.

In our approach, we are treating GESMES as a highly domain-dependent, generic standard for exchange of statistical data. The argument could be exemplified with the highlighted application of the system: GESMES/CB[29], which constitutes a message used by the European Central Bank to exchange statistical data and metadata with its partners in the European System of Central Banks (ESCB) and other organisations world-wide[30](see example below showing the general structure of the message). The user communities of GESMES/CB have agreed on using a minimum set of common code lists. Thus, the exchanged messages are looking similar, not only from a syntactical point of view, but also, to a great extent, semantically. Generally, institutions willing to use GESMES/CB in practice, have to use a set of structural metadata: either structural metadata which have been already devised by a centre institution or –if new data flows are concerned- new (or additional) structural metadata need to be devised[31]. The application was further a key element in the statistical preparations for Monetary Union and has proved both efficient and effective in meeting the ESCB's rapidly evolving statistical requirements.

indication about the character used for the missing values and the data values (GESMES/CB User Guide, p.102).

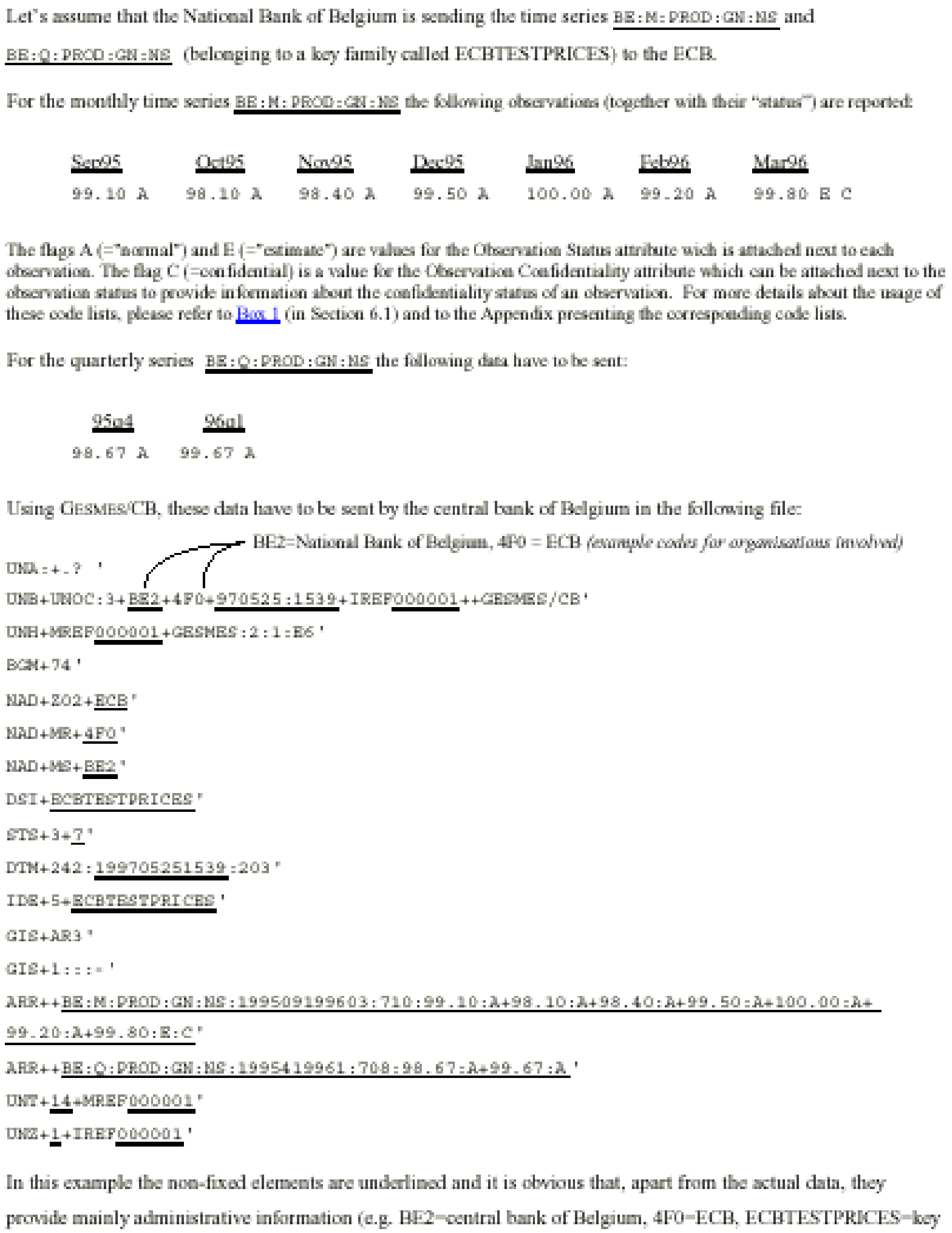
[28] A special feature in GESMES/CB is the possibility to distinguish between the concepts of a key family and of a data set. A key family comprises all time series that follow the same key structure, i.e. us the same dimensions in the same order for the time series keys, and which have the same statistical concepts as attributes on the same attachment level. A data set, i.e. technically speaking the information written in the Data Segment Identifier (DSI) of a given GESMES/CB message, may consist of all time series belonging to a key family (data and/or attributes at all levels) or a subset, for example, only data or only attributes for all series, or data and attributes for selected series only. Data exchange arrangements between institutions may stipulate the "grouping" of sibling groups and time series of a given key family into different data sets (for further reading, see GESMES/CB User Guide p.96).

[29] The last release of GESMES/CB was, July 30, 2000.

[30] The rules used in GESMES/CB are logically stronger and more restrictive than those used in the generic GESMES. GESMES/CB is accentuated being designed in a platform independent manner. The message implements a time series data exchange model (GESMES/CB Data Model) which provides for the exchange of multi-dimensional time series and a variety of associated metadata. It employs a GESMES profile and EDIFACT syntax. Although the system is able to accomplish an "update" and a "delete" message in the same interchange, one apparent shortcoming in GESMES/CB is that is not equipped with a mechanism to distinguish between "reporting new data" and "reporting corrections in previously reported data". Thus, it is up to the receiving application how to process the information contained in a GESMES/CB interchange.

[31] A message can only be one of the following three types: 1) "structural" message containing code lists, concepts and or/key family structure definitions; 2) an "update" message containing data/attributes or instructions for deletions; 3) a data request message.

Finally, we see GESMES in general, and especially the application, GESMES/CB, has a substantially narrower scope compared to the current state of the DDI. Already in the introduction in the User Guide it is emphasised that the application seeks “an easy adaptation to any economic domain and flexible coverage of all types of economic – statistical data” (GESMES/CB User Guide p.11). It is a pure technical format that seeks to establish routines and principles for the exchange of statistical data, i.e. import and export of data. GESMES is a standard for exchange of data, but not a complete integrated model. It discovers the most important sections, but as long as these are not dependent of each other in any manner, we can not call it an integrated data model.



Let's assume that the National Bank of Belgium is sending the time series BE:M:PROD:GN:NS and BE:Q:PROD:GN:NS (belonging to a key family called ECBTESTPRICES) to the ECB.

For the monthly time series BE:M:PROD:GN:NS the following observations (together with their "status") are reported:

| Sep95 | Oct95 | Nov95 | Dec95 | Jan96 | Feb96 | Mar96 |
|---|---|---|---|---|---|---|
| 99.10 A | 98.10 A | 98.40 A | 99.50 A | 100.00 A | 99.20 A | 99.80 E C |

The flags A (="normal") and E (="estimate") are values for the Observation Status attribute wich is attached next to each observation. The flag C (=confidential) is a value for the Observation Confidentiality attribute which can be attached next to the observation status to provide information about the confidentiality status of an observation. For more details about the usage of these code lists, please refer to Box 1 (in Section 6.1) and to the Appendix presenting the corresponding code lists.

For the quarterly series BE:Q:PROD:GN:NS the following data have to be sent:

| 95q4 | 96q1 |
|---|---|
| 98.67 A | 99.67 A |

Using GESMES/CB, these data have to be sent by the central bank of Belgium in the following file:

BE2=National Bank of Belgium, 4F0 = ECB *(example codes for organisations involved)*

```
UNA:+.? '
UNB+UNOC:3+BE2+4F0+970525:1539+IREF000001++GESMES/CB'
UNH+MREF000001+GESMES:2:1:E6'
BGM+74'
NAD+Z02+ECB'
NAD+MR+4F0'
NAD+MS+BE2'
DSI+ECBTESTPRICES'
STS+3+7'
DTM+242:199705251539:203'
IDE+5+ECBTESTPRICES'
GIS+AR3'
GIS+1:::-'
ARR++BE:M:PROD:GN:NS:199509199603:710:99.10:A+98.10:A+98.40:A+99.50:A+100.00:A+
99.20:A+99.80:E:C'
ARR++BE:Q:PROD:GN:NS:1995419961:708:98.67:A+99.67:A'
UNT+14+MREF000001'
UNZ+1+IREF000001'
```

In this example the non-fixed elements are underlined and it is obvious that, apart from the actual data, they provide mainly administrative information (e.g. BE2=central bank of Belgium, 4F0=ECB, ECBTESTPRICES=key

The National Bank of Belgium is sending a time series to the European Central Bank (ECB)

## 8) idaresa

idaresa - integrated documentation and retrieval environment for statistical aggregates - was one of the approved multinational European RTD co-operation set out to design and implement a metadata-based statistical information and data processing system targeted at the practical needs particularly of statistical agencies and offices in charge of supplying high-quality statistical information. In this RTD endeavour, special emphasis was laid on the harmonisation of statistical data originating from different sources and contexts. The limitation of idaresa is that it is a description of the standards for coodebooks, exclusively.

Main emphasis was laid on easy and powerful data dissemination capabilities as well as support in data retrieval, data aggregate formation and data export to third-party systems for subsequent data processing and analysis purposes. To this end, idaresa made heavy use of advanced meta information methodologies for processing data descriptive information - metadata - in parallel to primary data (observation data) and secondary data (statistics derived analytically from primary data). The envisioned approach encompasses both on-line data documentation and data consumption, including a range of tasks from simple data locating through sophisticated retrieval, aggregation, and tabulations. To this end, the most advanced information technologies and software engineering techniques were exploited to realise a major stride towards a truly unified, coherent, and interoperable European statistical system[32].

idaresa outcome is a running network software system prototype, inter-linking spatially separated data production sites - such as NSIs - in a heterogeneous distributed data management and processing environment based on a custom-tailored client-server architecture such that, ideally, clients (data consumers) are enabled to define information requests irrespective of actual storage structures and physical distribution of source data over the network. Practically, idaresa's end-user interface extends an existing and currently used thesaurus-based natural language front-end interface, supporting data access and navigation without explicit recourse to formal database query specifications and, thus, enabling non-technical subject domain-oriented interaction modes[33].

[32] idaresa is one of the approved research and development projects of EUROSTAT's DOSIS (Development of Statistical Information Systems) initiative, a special task of Esprit's Emerging Software Technologies track within the 4th EU Framework Programme for Research and Technological Development. The project commenced officially on January 1, 1996 and ended at February 28, 1999.

[33] Having in mind the vast majority of would-be data consumers seeking high-quality statistical information without necessarily being familiar with internal data holding organisations, idaresa provides a tailored tool for powerful and convenient wide-area public access to statistical information pools of European interest. Additionally, focusing on the pressing demands of statistical offices - like NSIs, EUROSTAT, or OECD - and data providers with similar task profiles, idaresa is prepared for implementation in real statistical data processing environments, especially data production sites being in charge of integrating inevitably heterogeneous statistical data originating from various non-centralised, autonomous data holdings.

## 9) IMIM[34] -

Meta-Information Systems Management - project presents an approach to the integration of statistical meta information and statistical production management control in the conduct of co-ordinated statistical projects based on surveys and the dissemination of their results. The project is based on exploiting the results in statistical meta information research over the past ten years, achieved in such projects as those undertaken under the DOSIS Programme of Eurostat, and utilising the possibilities for co-operative, distributed statistical data production and management afforded by the current generation of client/server architecture. While idaresa stresses the consumption of statistical data within an integrated distributed system of heterogeneous data sources, IMIM contributes especially to the production of data from a typical statistical data supplier's point of view. A particular synergy of the idaresa - IMIM co-ordination is expected to result from the jointly developed framework for metadata exchange.

One of the more important outputs of the software developments in the project is *Bridge*. This is an integrated metadata system for statistical purposes. It provides several tools to process statistical metadata as well as facilities for data exchange. The main ambition behind this part of the project is to show a way to the details of the information system for the external user as well as for the statistician[35].

## 10) Applications and infrastructures

An information system, where most processes are statistically processed is called a statistical information system (SIS). Typically, an information system is associated with a specific purpose, e.g. to produce a specific information product, like a certain data set and/or a certain report or publication. Such an information system is called an information system application, or an application, for short. A typical example of an application in a national statistical office would be the information system associated with a statistical survey conducted regularly, like the Labour Force Survey (LFS). Internationally, an analogous example would be the OECD information system for producing the Main Economic Indicators (MEI). In addition to information system applications with very specific end-products, there are information

[34] IMIM is one of the approved research and development projects of EUROSTAT's DOSIS (Development of Statistical Information Systems) initiative, a special task of Esprit's Emerging Software Technologies track within the 4th EU Framework Programme for Research and Technological Development. The project commenced officially on January 1, 1996, and lasted for three years. The project joined researchers from several national statistical offices and additionally from the private sector.

[35] Bridge is designed as homogeneous Integrated Metadata System (IMS) instead of a heterogeneous system like many traditional metadata systems in statistical offices. This seems to be the only way to ensure the consistency of metadata. The following basic requirements for statistical metadata systems had been established when designing and implementing Bridge: Statistical metadata systems have to be able to exchange data with traditional metadata databases; Statistical metadata systems need to be supported with consistent version control for metadata entries, relationships and indexes; As international information systems statistical metadata systems should be designed as multilingual system; and Statistical metadata systems should be knowledge-based systems. The idealistic approach of the Bridge system is based on these requirements and practical experiences have shown just in the modelling phase that such an IMS becomes a very complex system. To handle this complexity several metadata levels have been defined that are related to different types of "users".

systems with more general purposes, e.g. the information system associated with a database service to the external and/or internal users of the statistical data produced by a statistical organisation.

In conclusion, it is of major importance to emphasise that idaresa is a standard for documentation, thus comparatively different to the other application standard discussed in this section; the data exchange standard, GESMES. The latter is a domain-specific standard, nevertheless it is still able to handle datasets which is based on time series. Both GESMES and idaresa were originally carried out in an era that didn't have support of modeling tools like XML DTDs, and apparently not sophisticated tools like SOAP, RDF and XML Schemas[36]. The two standards have a much more limited scope than what is attempted in DDI.

## 11) Summary

We have presented six (statistical) metadata standards, in addition to two empirical examples, in this note. At first, we discovered that OMG and ISO/EDI 11179 are both, complete domain-independent standards, which need an additional application for usage. They emphasise defining objects and relations between objects representing generic concepts.

Our real-world examples introduced the concept of interoperability and underlined difficulties handling a range of standards simultaneously. Further, we notice that the papers were not able to present a solution combining the complementary functionality of two of the relevant tools/standards according to LIMBER: RDF Schema and XML Schema.

Dublin Core is a bibliographic standard, to a great extent domain-independent, while its fifteen elements could be used in a wide range of applications. Another advantage is that DC is extensible. Nevertheless, although the Core enables interoperability within domains, the standard introduce the problem of incompatibility between disparate and heterogeneous metadata descriptions or schemas across domains. Additionally, the role of the DC has always been envisaged only to satisfy initial search criteria before more refined searches would be accomplished.

GESMES and idaresa fulfil two different concerns, while they have in common being what we have called, relatively domain-dependent. The former, by its application its application GESMES/CB, is a standard for exchange of (economic) statistical data while the latter, with its application, IMIM, a documentation standard. They were originally both developed in an era which could not rely on a modeling tool like the XML. Therefore, they are both more old-fashioned than DDI and would, if seeking to implement them, soon be comprehended as strait jackets in such an environment.

It would be an exaggeration asserting that the current version of DDI is complete. Nevertheless, the standard fits most of the relevant data in the current state, in spite of the well-known shortcomings. With the proceeding extensions and improvements, such as the implementation of the cube-proposals, further progress according to MOM, and improvements/extensions of more controlled vocabularies, the DDI should be an appropriate standard able to in a sufficiently manner fulfil description of the rest of the surveys which it is not able to manage in the current version.

[36] We are quite aware that the UN/EDIFACT family standards later put efforts to move to an XML based platform. Nevertheless, the original standards were developed in an environment without these useful and often necessary tools.

# References

CWM *Common Warehouse Model* http://www.omg.org/technology/cwm/

DDI *Data Documentation Initiative: A Project of the Social Science Community.* http://www.icpsr.umich.edu/DDI/codebook.html

DC *The Dublin Core Metadata Initiative.* http://dublincore.org/

FASTER *Flexible Access to Statistical Tables and Electronic Resources* http://www.faster-data.org/

GESMES *Generic Statistical Message* http://www.unece.org/trade/untdid/d99b/trmd/gesmes_c.htm

GESMES/CB the time series data exchange message User Guide http://www.ecb.int/stats/gesmes/gesmes.htm

HASSET (1999). *Humanities and Social Science Electronic Thesaurus*. http://biron.essex.ac.uk/searching/zhasset.html

Hunter, Jane: *MetaNet – A Metadata Term Thesaurus to Enable Semantic Interoperability Between Metadata Domains,* Journal of Digital Information , 2001 http://archive.dstc.edu.au/RDU/staff/jane-hunter/harmony/jodi_article.html

Hunter, Jane and Lagoze, Carl: *Combining RDF and XML Schemas to Enhance Interoperability Between Metadata Application Profiles*, paper, 2001 http://www.cs.cornell.edu/lagoze/papers/Jane%20WWW10/Combining%20RDF%20and%20XML%20Schemas%20to%20Enhance%20Interoperability%20Between%20Metadata%20Application%20Profiles.html

idaresa - *integrated documentation and retrieval environment for statistical aggregates*, http://idaresa.univie.ac.at/outline.htm

IMIM *Integrated Meta-Information Management*, http://imim.scb.se/

ISO/IEC 11179 *The International Organization for Standardization and the International Electrotechnical Commission* http://www.sdct.itl.nist.gov/~ftp/l8/11179/ and http://www.iso.ch/

Matthews, B.M, Miller, K, Ramfos A, Ryssevik J, Wilson, M.D *Internationalising data access through LIMBER*. Forthcoming, http://brains.open.ac.uk/cfdocs/iwips/html/about.htm

Miller, K & Matthews B.M. (2001) *Having the right connections: the LIMBER project.* Forthcoming, Journal of Digital Information, http://jodi.ecs.soton.ac.uk/.

NESSTAR. *Networked European Social Science Tools and Resources*. http://www.nesstar.org/

OMG *Object Management Group* http://www.omg.org/

Ryssevik, J: *Global access to data: NESSTAR*, a special issue of the NSD Newsletter, 1999

Ryssevik, J & Musgrave, S (1999). *The Social Science Dream Machine: Resource discovery, analysis and delivery on the Web.* IASSIST Conference, Toronto. http://www.nesstar.org/papers/iassist_0599.html

# PART 6: A Metadata Framework: Requirements

## 1. Introduction

This part of the deliverable considers the current state of the Data Documentation Initiative (DDI) and also its future development. This is considered in the light of emerging recommendations from the World-Wide Web Consortium (W3C) in the domains of defining syntax schemas for validating XML data structures (XML Schema), and in web-compatible metadata (RDF).

## 2. DDI

The Data Documentation Initiative ([DDI]) is "*an effort to establish an international criterion and methodology for the content, presentation, transport, and preservation of 'metadata' about datasets in the social and behavioral sciences.*" After some 5 years of largely voluntary effort in the definition its metadata standards, this initiative is starting to take off with wider acceptance in the social science community, especially in the data archives. Software has appeared which supports the DDI, including the Nesstar suite of tools ([Nesstar]).

### 2.1 The DDI DTD

The DDI has produced Document Type Definition (DTD) to express the various aspects of social science metadata. This is divided into 5 sections.

- The documentation description: describing the XML DDI codebook used in the metadata.
- The study description: describing the study that generated the data. This is further subdivided into citation (title, responsibility, production etc), information (subject, abstract, scope etc), methodology (data collection methods used etc), access (availability and usage conditions) and other (related materials etc).
- The file description: detailing the physical format of the data file.
- The data description: describing the variables used within the data, the questions used to capture them and their possible ranges.
- The other description: providing a generic pointer to any other related material.

Each section is further subdivided into further sections to provide a hierarchical structure to the metadata.

### 2.2 Problems of the current DDI DTD

Some of these are discussed below in the context of RDF and DDI.

### 2.3 Proposals for extending DDI

The LIMBER project has focused on this XML metadata model to demonstrate how its objectives can be fulfilled. Eventually it is hoped to demonstrate that adoption of an object-oriented model, expressed in RDF or xmlSchema, can add consistency, flexibility and extensibility to the DDI standard. Also such a model should aid the adoption of the standard as an inter-operability tool making tighter links to the Dublin Core and the consistent use of controlled vocabularies.

However, in the short term, certain minor amendments to the 1.0 version of the DDI XML codebook standard can aid the adoption and assignment of controlled vocabularies to certain elements. This would greatly enhance the inter-operability between sites that are using the standard.

Within the DDI at present controlled vocabularies are assigned to certain attributes of a few elements, with recommendations to add more in the future. In addition to this there are three elements, <keyword>, <topcClas> and <concept>, to which keywords from a controlled vocabulary can be assigned, with attributes which name (vocab) and locate (vocabURI) the particular thesaurus or ontology chosen, and the ability to show language via the xml:lang attribute. These elements only allow assignment at either the study or variable level.

Present attribute controlled vocabularies:-
<fileStrc> type (rectangular|hierarchical|relational)
<varGrp> type (section|multipleResp|grid|display|repetition|subject|
                    version|iteration|analysis|pragmatic|record|file|
                   randomized|other)
<sumStat> type (mean|medn|mode|vald|invd|min|max|stdev)
<catStat> type (freq|percent|crosstab)
<varFormat> schema (SAS|SPSS|IBM|ANSI|ISO|XML-Data|other)
                 category (date | time | currency | other)

Proposed attribute controlled vocabularies (N.B elements with attributes already assign for these listings have them displayed)

<anlyUnit> unit
<timeMeth> method
<frequenc> freq
<resInstru> type
<dataChck>
<catgryGrp> missType

Present elements where controlled vocabularies can be assigned.

```
|    |    |---- 2.2.1 subject? (ATT == ID, xml:lang, source)
|    |    |    |
|    |    |    |---- 2.2.1.1 keyword*  (ATT == ID, xml:lang, source, vocab, vocabURI)
|    |    |    +---- 2.2.1.2 topcClas* (ATT == ID, xml:lang, source, vocab, vocabURI)


|    |---- 4.2 var* (ATT == ID, xml:lang, source, name, wgt, wgt-var, qstn, files,
|    |    |              vendor, dcml, intrvl, rectype, sdatrefs, methrefs, pubrefs, access)
|    |    |
|    |    |---- 4.2.21 concept*    (ATT == ID, xml:lang, source, vocab, vocabURI)
```

The LIMBER project proposes that the ability to assign keyword at the variable group level (<varGrp>) should also be included. Hence allowing:- stricter definition of groups of questions, the ability to use the structure of the thesaurus to refine searches, the ability to translate the keywords assigned into various languages and the choice for smaller archives or data centres, with scarce resources, to assign at this intermediate level. The proposed amendment follows the example of the <var> element and has no effect on existing mark-up.

```
|    |---- 4.1 varGrp* (ATT == ID, xml:lang, source, type, var, varGrp, name, sdatrefs,
|    |    |                    methrefs, pubrefs, access)
|    |    |
|    |    |---- 4.1.1 labl*     (ATT == ID, xml:lang, source, level, vendor)
|    |    |---- 4.1.2 txt*      (ATT == ID, xml:lang, source, level)
|    |    |---- 4.1.3 defntn?   (ATT == ID, xml:lang, source)
|    |    |---- 4.1.4 universe? (ATT == ID, xml:lang, source, level, clusion)
|    |    |---- 4.1.5 concept*    (ATT == ID, xml:lang, source, vocab, vocabURI)
|    |   +---- 4.1.6 notes*    (ATT == ID, xml:lang, source, type, subject, level, resp,
                                sdatrefs)
```

LIMBER also requests the addition of controlled vocabularies for elements from the <sumDscr> and <method><dataColl> sections, many of which already have listings or are earmarked to contain one in the future. However, the problems here are, which is the best method to employ, how to be compliant with previous versions and how to maintain such listings.

The problem with controlled vocabularies assigned to attributes and published within the DDI itself is that when the list requires updating a new version of standard has to be published after consensus of the committee. There are several other methods that could be employed:- inclusion of vocab and vocabURI within a.global, keeping all keywords within the <keyword> element and using the link mechanism to point from the appropriate element, assigning concept sub-element to specific elements or assigning the vocab and vocabURI just to specific elements.

LIMBER suggests the adoption of the vocab and vocabURI attributes to the following elements from the two sections below highlighted in red. However this does make the following attributes redundant:- unit, method and type (shown in blue). It suggests that the DDI recommends thesauri or even maintain their own listings at the DDI web site. It tries to be consistent in that all <concept> elements are used within the dataDscr section, all vocab/vocabURI elements are used within the stdyDscr section and all existing published controlled vocabularies remain. Also recommended is the inclusion of a format attribute on the date elements to contain "mm-dd-yyyy" etc definitions of the information found in the date attribute.

```
|    |---- 2.2.3 sumDscr* (ATT == ID, xml:lang, source)
|    |    |    |
|    |    |    |---- 2.2.3.1 timePrd*   (ATT == ID, xml:lang, source, event, format, date, cycle)
|    |    |    |---- 2.2.3.2 collDate*  (ATT == ID, xml:lang, source, event, format, date, cycle)
|    |    |    |---- 2.2.3.3 nation*    (ATT == ID, xml:lang, source, vocab, vocabURI, abbr)
|    |    |    |---- 2.2.3.4 geogCover* (ATT == ID, xml:lang, source, vocab, vocabURI)
|    |    |    |---- 2.2.3.5 geogUnit*  (ATT == ID, xml:lang, source, vocab, vocabURI)
|    |    |    |---- 2.2.3.6 anlyUnit*  (ATT == ID, xml:lang, source, unit, vocab, vocabURI)
|    |    |    |---- 2.2.3.7 universe*  (ATT == ID, xml:lang, source, vocab, vocabURI,
                                          level, clusion)
|    |    |    +---- 2.2.3.8 dataKind*  (ATT == ID, xml:lang, source, vocab, vocabURI)
```

```
|    |---- 2.3 method* (ATT == ID, xml:lang, source)
|    |    |
|    |    |---- 2.3.1 dataColl* (ATT == ID, xml:lang, source)
|    |    |    |
|    |    |    |---- 2.3.1.1 timeMeth*      (ATT == ID, xml:lang, source, method, vocab, vocabURI)
|    |    |    |---- 2.3.1.2 dataCollector* (ATT == ID, xml:lang, source, abbr, affiliation)
|    |    |    |---- 2.3.1.3 frequenc*      (ATT == ID, xml:lang, source, freq)
|    |    |    |---- 2.3.1.4 sampProc*      (ATT == ID, xml:lang, source, vocab, vocabURI)
|    |    |    |---- 2.3.1.5 deviat*        (ATT == ID, xml:lang, source)
|    |    |    |---- 2.3.1.6 collMode*     (ATT == ID, xml:lang, source, vocab, vocabURI)
|    |    |    |---- 2.3.1.7 resInstru*    (ATT == ID, xml:lang, source, type, vocab, vocabURI)
|    |    |    |---- 2.3.1.8 sources?      (ATT == ID, xml:lang, source)
|    |    |    |    |                                    |
|    |    |    |
|    |    |    |---- 2.3.1.9 collSitu*     (ATT == ID, xml:lang, source)
|    |    |    |---- 2.3.1.10 actMin*      (ATT == ID, xml:lang, source)
|    |    |    |---- 2.3.1.11 ConOps*      (ATT == ID, xml:lang, source, agency)
|    |    |    |---- 2.3.1.12 weight*      (ATT == ID, xml:lang, source)
|    |    |    +---- 2.3.1.13 cleanOps*    (ATT == ID, xml:lang, source, agency)
```

Further proposals for extending DDI are given at
http://www.icpsr.umich.edu/DDI/future.html;

# 3. XML Schemas

The current plan within the W3C is to phase out the use of DTDs as a means of describing the syntactic form of classes XML document for validation purposes. DTDs are seen as an old and inflexible method for specifying syntax, which requires users to learn a new syntax. XML Schemas should be used instead. These describe the syntactic form of XML in an XML syntax, and also provide a more flexible and powerful mechanism. For example, XML Schemas provide:

- An extended datatyping mechanism allowing true datatypes for attributes, such as integers, booleans and dates.
- Inheritance hierarchies.
- More flexible occurrence indicators for elements (more flexible than the ?, +, * indicators in DTDs).
- Schemas across multiple documents.

To maintain the DDI metadata standard inline with current directions from the W3C, an XML schema should be provided to express at least the same syntax as the current DTD allows. As XML Schema have a greater expressive power than DTDs this should be a straightforward task, and should be automatable. Standard converters from XML DTDs into XML Schemas are currently under development (for example, the IBM Alphaworks XML tool DdbE has such a capability [Alphaworks] – see appendix A for an example of delivering DDI as an XML Schema).

XML Schema became a recommendation of the W3C in May 2001 [XMLSchema1, XMLSchema2].

# 4. RDF

The Resource Description Framework (RDF) [RDFActivity] is the development of the W3C metadata activity. It represents an attempt to develop a framework for describing resources on the web (that is metadata) in a "machine-understandable" manner. RDF metadata is designed to be used in a variety of applications, for example (partially from [IntroRDF]):

- in resource discovery to provide better search engine capabilities;

- in cataloguing for describing the content and content relationships available at a particular Web site, page, or digital library;
- by intelligent software agents to facilitate knowledge sharing and exchange;
- in content rating; in describing collections of pages that represent a single logical "document";
- for describing intellectual property rights of Web pages.
- For describing profiles of devices (browsers, mobile devices) and preferences of the user (presentation styles and accessibility, privacy).

RDF is defined in two parts. The Model and Syntax recommendation [RDFModelSyntax] defines the model that underlies RDF and the syntax which is used to give it a machine representable form. The latter is a XML based syntax, allowing it to be read and processed by standard XML tools.

The underlying model of RDF is one of directed-labelled graphs. Nodes in the graph represent web resources, and arcs in the graph properties of resources. Thus the basic unit of RDF is the triple, consisting of a source resource, a property and a target resource, or valuation. Thus properties can be assign values to resources. This can be extended within the RDF syntax to allow aggregations of properties, alternative properties, and properties of properties (a process known as reification).

The basic assumption underlying RDF is that the Web is an open world; anybody can say anything about anything. Thus users can add new properties to resources at will, allowing multiple valuations of the same property to the same resources, even allowing contradiction. It is up to the user application to decide which statements can be trusted. This *property-centred* approach is in contrast with the object-oriented approach, which centres on the resources (*classes*), defines a fixed set of properties (*attributes*) of those resources, and has restricted access (*encapsulation*) to those properties.

The second part of RDF is the RDF Schema definition [RDFSchema]. This allows user to define a particular ontology for the classes of resources and the properties of those resources within a particular domain. Thus, particular RDF statements can be validated against the schema and then be said to "mean" what the schema defines. RDF schemas also allow subclasses and sub-properties, and constraints on the types of resources.

RDF thus defines a simple, abstract and powerful mechanism for defining Web based metadata. The RDF Schema effort within W3C after a long gestation period is now quite advanced, with the main threads of syntax and schema recommendations or proposed recommendations. There is now general acceptance of the purpose and scope of RDF, and its relationship with other activities (such as XML Schema) has now been clarified. The activity, however, still has much to complete. There is as yet limited tool availability (largely experimental free ware) and additional recommendations for querying, capturing ontologies, and reasoning are subjects of research. Nevertheless, this activity within W3C is now gaining momentum and should be explored seriously by the Limber project, and wider in the DDI.

## 4.1 General advantages of using RDF.

**Property oriented – extensible.**
The property-oriented nature of RDF means that it is easy to extend and adapt in a manner that does not break the existing processors, tailored to earlier versions of the domain ontology.

**Statements about statements ("reification").**

It is straightforward to extend the metadata hierarchy within RDF. Meta-metadata can be defined ("statements about metadata") through the reification process, where statements about the properties of resources are themselves turned into resources themselves.

**An XML format.**
RDF is an XML format, and so it can be processed via existing XML parsers and processors, transformed via XSLT and can adopt XML Schema datatypes for a base typing mechanism.

### Using work and tools of others.

There is now a wide community of users and developers surrounding RDF, represented for example through the RDF Interest Mailing list – www-rdf-interest@w3.org. By using RDF, the expertise of this community in methods of using RDF to represent metadata (specifically metadata, rather than other forms of data represented in XML) and also the tools developed within this community, can be brought to bear on the use of RDF within the social science domain.

### Using existing work on knowledge representation and ontologies.

There is a large body of existing expertise going back over 30 years on knowledge representation and ontologies in the computer science community, which is now heavily involved in the RDF activity. By adopting RDF, the social science community can take advantage of this work more easily.

### Ability to use reasoning across web objects using RDF.

A very active line of research within the RDF community is into how to use RDF to reason about web resource across the web. This is especially important to applications concerning trust, authentication and negotiation, whereby the parties have to determine from the statements place on the web about resources, and statements about those statements concerning their authenticity (from third parties say) whether that the resources satisfy the properties desired by the parties. An application of reasoning for social science data archives may whether a party has satisfied the access conditions to access a certain dataset. Conversely, the user may be able to reason about whether the content of the dataset may satisfy the criterion on the study he or she desires.

### Capturing semantics using RDF ("the Semantic Web").

XML is a syntactic format; it lacks the machinery to capture a representation of the semantics of the entities it is representing. XML thus relies on human intelligence and consistency to process it correctly; the semantics would then be buried in applications. The approach of RDF is to bring the semantics (i.e. the directed graph model) closer to a machine-readable form, so that "semantics-aware" negotiation processing can be performed on a machine to machine basis.

**Adding constraints to metadata**.
Constraints can be added to the metadata in the schema.

# 5. RDF and DDI

We consider some issues under consideration for extending the DDI and discuss how the use of RDF could assist.

**Extensibility without changing.**

There is a requirement that DDI should evolve in such as way that it can be extensible without modifying existing DDI descriptions. XML DTDs are rigid structures. If elements are introduced in one version of DDI at some point in the hierarchy, then all the data written against the new version are likely to be rejected by other interpreters using other versions of the DTD, even those components which conform to the original DTD.

RDF is designed with extensibility in mind: if a triple is encountered which does not conform to the known RDF schema, the processor should ignore it and continue to process the rest of the RDF description in the context of the old schema. RDF also has a concept of an object hierarchy built in (as does XML Schema). Thus reusability and extensibility can be achieved by adding new subclasses or subproperties.

**Relationships between surveys**
The current DTD is tailored to describe one dataset in isolation. However, surveys are often in families or series, and it would be useful to provide metadata that connects the sets together. Extending the DTD in a fairly straightforward manner could do this. However, in this approach, the data of the various surveys will be gathered in one place and connected together as one unit; this makes accessing the data about one survey alone more difficult, and makes it harder to extend the set of datasets as the resource evolves.

In RDF, the metadata for each dataset can be provided independently of its family, and through reification the metadata itself can be regarded as a resource which the family metadata can refer to. This can be extended easily and in a distributed manner (i.e. the descriptions of the datasets can be located anywhere).

**Hierarchical data sets**
The current DDI is tailored towards rectangular survey data. However, there is a requirement to accommodate other forms of data structuring such as hierarchical data structures (derived from CAI for example) or time series data.

This could again be achieved through developing the DTD, but with the problems of extensibility outlined above. XML Schema with its use of inheritance hierarchies may also allow a more flexible extension of the number of types of datasets. RDF has the advantages here of extensibility as outlined above.

## Multilinguality

This is a particularly important aspect to the Limber project. The capacity for multilinguality is not built into the current DDI. All elements have an xml:lang attribute allowing the language of the element to be specified. However, some elements are not duplicable, making it difficult to provide language alternatives. This can be resolved in an extension of the DTD. There is some debate on how language attributes can best be represented in RDF (see [TimBL98]) whether via the xml:lang attribute, which has no model in RDF, or through some other mechanism. Nevertheless, it should be straightforward to represent multiple languages alternatives (using the rdf:alt element) of the same metadata.

## Integration with other domains

An important requirement of the LIMBER metadata activity is the facility to interchange metadata with other domains. This would allow broader searches across subject disciplines, cross correlating data and deepen the insights offered by analysis of the social science data, opening the data for a much wider range of questions. Some example domains which Social Science metadata may wish to integrate data with include:

- Libraries for references to published literature (especially using the Dublin Core).

- Grey literature for research reports and other informal documentation (for example using the CERIF format).
- Geographical information systems for references to locations.
- Health data for cross-referencing between social science factors and public health.
- Environmental information (meteorological, geological, land use, pollution etc) for cross-referencing between environmental and social science factors.

By offering an integrated metadata capability, the social science system retrieval system becomes much richer, for example allowing questions which correlate environmental, health and social factors (e.g. "incidence of disease with poverty in wind shadow of nuclear power plant").
When using data from two or more domains, it would be desirable to be able to:
Use the data from the different domains in the same format, so that the same tools can be to present and manipulate each, in an integrated manner.

- The fields from the different domains are clearly distinguished so that the data from each can be discovered without ambiguity.
- Metadata descriptions that integrate fields from two or more domains can be processed by applications tailored from one or other of the domains without causing that application to fail. That is, the processors can access the fields relevant to that domain and ignore any others.
- The common components of metadata from two or more domain can be mapped together to link the information from domains, using an "ontology mapping".

RDF is ideally suited for use in such a scenario.

- It provides a common format for defining (using RDF Schemas) and representing (RDF Syntax) metadata from different domains.
- The XML Namespace mechanism, central to RDF, allows fields from different domains to be added to a single RDF description, whilst clearly distinguishing them as coming from different namespaces.
- Again the XML namespace mechanism allows elements from different domains to be passed onto different processors, without breaking (note that conventional XML DTD validation can be fail otherwise).
- RDF itself can capture the mapping between different domains.

Further, there is existing work in using RDF to capture metadata in some of the domains. See for example the expression in RDF of the Dublin Core metadata standard for published literature [DC-RDF], and the CERES (California Environmental Resources Evaluation System) and USGS Biological Resource Division joint programme to build digital environmental thesauri using RDF [CERES]. By expressing the DDI within RDF, this existing work can be used.

## 5.1 Issues on using RDF for DDI

Some issues remain which need to be resolved before using RDF to express the DDI codebooks

### Converter Software

There is a clear need to have appropriate converter software to allow for backward compatibility with existing DTD based DDI definitions. Given an appropriate formal RDF schema, this should not be a complicated piece of software. The given that both DDI 1.0 and RDF use a XML based syntax, the W3C standard XML transformation language XSLT

([XSLT]) would be an appropriate vehicle for defining such a transformation. Public domain software already exists for interpreting XSLT scripts.

## Immaturity of RDF

RDF is as yet immature. Whilst the model and syntax recommendations have been in existence for sometime, the vital schema recommendation has (as of Nov 2001) reached the candidate recommendation phase; indeed this part of the activity has not matured significantly during the course of the Limber project. The W3C has formally started a Semantic Web activity, and within this activity, the whole of the basis of RDF (syntax and semantics) has been subject to review, within the Core RDF activity. However, further developments are also being made, especially within the areas on supporting ontologies on the Web, such as the DAML+OIL project. When such initiatives reach maturity, then they will have a significant impact on the development of Metadata within Social Sciences and beyond.

The level of understanding of RDF within the computing community remains relatively small, and many issues remain to be sorted out, before being widely used outside a specialist community. Further, tool support remains currently experimental and limited. Nevertheless, there is great momentum behind RDF: a noted enthusiast is Tim Berners-Lee, Director of the W3C and inventor of the WWW, who sees this technology as a key (with digital signatures) to the "web-of-trust". In this model, intelligent decisions can be made across the Web, using RDF as the medium for passing semantics based information and also passing trust information to test the reliability of statements.

## Learning curve for users

RDF represents a higher-level of abstraction than XML alone. XML is a mechanism for representing the structure of information in syntax. It does not however, attempt to communicate any meaning associated with the names and structuring thus defined, which are defined outside the language, in user guides etc. RDF does attempt to communicate more about the semantics of objects. This represents a steep learning curve for users unfamiliar with computer science, let alone knowledge representation techniques. XML already represents a raising of the level of abstraction into a formal computer representation, which is hard for some users to understand; RDF is harder again for users to understand. This may represent a significant barrier to introducing the benefits of RDF to the wider community.

# 6. A metadata framework

As a framework, I follow the categorisation of metadata from Keith Jeffery, dividing metadata into three main divisions and several subdivisions beneath that.

– **Schematic**

The data model . A (logical) description of the structure of the resource, the relationship between the elements of the resource, and any constraints.

- **Concrete**

  Provides a machine view close to physical representation - data formats, fields, strings, code interfaces.

- **Abstract**:

  Provides a user view, near the real-world - abstract entities and relationships between them

- **Mappings**

  Capturing the relationships between levels and domains; one abstract to many concrete.

– **Navigational**

Provides the information on where resources and their sub-components are located. Often tends to be mixed up in other metadata. Ideally, kept separate from other information.

– **Associative**

All other information about a resource.

- **Descriptive**:

  what the resource is about and where it comes from.

- **Restrictive**:

  how the resource can be used.

- **Supportive**:

  the context in which the resource sits.

The metadata is structured into categories which correspond to this framework.

This provides a set of top-level categories. To specialise to a particular domain of interest, we use an *object inheritance* mechanism, providing an initial class hierarchy of metadata objects in this document.
We shall use UML class diagrams to illustrate the relationships between the components of the metadata model.

A complementary classification is according to Swetland (2000). We could divide different types of metadata and their functions into formal but nevertheless generic boxes:

- **Administrative**, which covers matadata used in managing and administrating information resources, examples here would be acquisition information, rights and reproducing tracking, location information, selection criteria for digitization, version control and differentiation between similar information objects, audit trails created by record-keeping system, and documentation of legal access requirement

- **Descriptive**, i.e. metadata used to describe or identify information resources, examples would be cataloging records, specialised indexes, finding aids, annotations by users, metadata for record-keeping systems generated by records creators, and hyperlinked relationships between resources

- **Preservation**, i.e. metadata related to the preservation management of information resources, for example documentation of physical condition of resources, and documentation of actions taken to preserve physical and digital versions of resources, i.e. data refreshing and migration

- **Technical**, i.e. metadata related to how a system functions or metadata behave, for example documentation of hard- and software, tracking of system response times, digitization information, e.g. formats, compression ratios, scaling routines, and authentication and security data, e.g. encryption keys, access control and passwords

- **Use**, i.e. metadata related to the level and type of use of information resources, for example exhibit records, use and user tracking, and content re-use and multi-versioning information

# References

[Alphaworks] IBM alphaWorks technology development site http://www.alphaworks.ibm.com

[TimBL98] Tim Berners-Lee http://www.w3.org/DesignIssues/InterpretationProperties.html

[CERES] The CERES/NBII Thesaurus Partnership Project http://ceres.ca.gov/thesaurus/

[DC-RDF] Guidance on expressing the Dublin Core within the Resource Description Framework (RDF) http://www.ukoln.ac.uk/metadata/resources/dc/datamodel/WD-dc-rdf/

[DDI] The Data Documentation Initiative homepage http://www.icpsr.umich.edu/DDI/codebook.html

[IntroRDF] Introduction to RDF Metadata, W3C NOTE 1997-11-13 http://www.w3.org/TR/NOTE-rdf-simple-intro-971113.html

[NESSTAR] The European Networked Social Science Tools and Resources project http://www.nesstar.org

[RDFActivity] W3C Resource Description Framework activity page http://www.w3.org/RDF/

[RDFModelSyntax] *Resource Description Framework (RDF) Model and Syntax Specification,* W3C Recommendation, 22 February 1999, http://www.w3.org/TR/REC-rdf-syntax/

[RDFSchema] *Resource Description Framework (RDF) Schema Specification 1.0*, W3C Candidate Recommendation 27 March 2000 http://www.w3.org/TR/2000/CR-rdf-schema-20000327/

[XML] *Extensible Markup Language (XML) 1.0,* W3C Recommendation 10-February-1998, http://www.w3.org/TR/1998/REC-xml-19980210

[XMLSchema1] XML Schema Part 1: Structures , W3C Recommendation 2 May 2001, http://www.w3.org/TR/xmlschema-1

[XMLSchema2] XML Schema Part 2: Datatypes , W3C Recommendation 2 May 2001, http://www.w3.org/TR/xmlschema-2

[XSLT] *XSL Transformations (XSLT) Version 1.0,* W3C Recommendation 16 November 1999, http://www.w3.org/TR/1999/REC-xslt-19991116

# Part 7: Towards an Object Oriented Model for the DDI Metadata

## 1 Introduction

A thorough debate regarding the lacks and shortcomings of the current DTD a new approach was introduced for the partners, at the Limber metadata-meeting in Athens in September 2000. Instead of a prolonged arguing regarding the superiority of RDF and/or XML Schema it was emphasised a proposal to focus our endeavour converting/transforming the current DTD into a object-oriented (OO), more or less representation-independent platform.
This approach was forwarded as proposal to the DDI-TC (technical committee) meeting in Washington DC, 26 November 2001, and adopted as an approach towards the development of DDI 2.0.

This part of the deliverable will highlight some of the motivations for change, and present the important arguments for why an object oriented approach should be adopted for the development of the DDI standard, compared to other possible solutions. Examples will be given of the object-oriented model, structured as diagrams with textual explanations. However, the object-classes in our example models are solely proposals, examples of how the model could be structured and how the metadata-stream could flow.

# Part 7a: Arguments for a conversion of the current DDI DTD towards a representation-independent object oriented model approach

## 1. Representation-independence

The superiority of an object-oriented model approach, in comparison with not only the current DDI DTD, but additionally XML Schema, SOAP (Simple Object Access Protocol) and RDF, is the concept of independence of representation. Adoption of an object-oriented model would allow selection of the data representation best suited for specific purposes, independent of time and location. For instance, for one purpose representation in RDF might be the most suitable, for another XML Schema, for a third a database the most appropriate. With such a solution we will be able to select more or less whatever representation we want not exclusively today but also in the future where other specifications and applications will predominate. For instance, if RDF in the future becomes the market leader then it would be an easy task to use this framework as the fundamental representation. The representation-independence is not a prevalent feature or a possible solution with any of the competitive alternatives.

There are at least two different tasks to take into consideration. First the task of the actual conversion of the current DTD, an established standard and secondly the conversion of material already marked-up to this standard. A conversion from the current DTD to, for instance XML Schema leads us into in a sub-optimal solution under the prevailing conditions. Unless one uses exactly the same elements and attributes as in the current DTD, additional work would be required for every single already marked-up survey after conversion to the new format. As of today, there are no complete tools available to carry through the conversion in an appropriate manner. However, we are aware that some tools are under development, like the IBM XML tool Dbde for a conversion from XML DTDs to XML Schemas. If, however the standard is strictly translated element for element, attribute for attribute into the new format then the focal arguments for changing the DTD are missing.

With an object-oriented solution there is an extra level of abstraction that is independent of storing format. Thus, we have the opportunity to generate different kinds of file-formats from the same model. In our opinion, this is distinctly a great benefit from the current situation, and as far we can see, also from the alternatives brought into the discussion.

## 2. Machine-understandable, not solely machine-readable

The current DTD was not developed with the object of being machine-understandable, only of being machine-readable. The current DDI DTD describes how the information should be contained, and not how it should be organised when an application needs to understand and use it. This feature has become a prevalent shortcoming during the practical use of the current DDI DTD. As long as we don’t have reliable and complete knowledge of the future, we cannot be sure if the content of the attributes with controlled vocabularies would remain sufficient and adequate.

With an object oriented approach it is possible to explicitly take into consideration which attributes we want to give objects and additionally which restriction we want to put on the values of those attributes.

# 3. Consistency/ Backward Compatibility

The conversion to the new object-oriented approach should be incremental. Thus, we don't have as our prominent aim building the entire model immediately. The process starts implementing a few objects with progressive additions over time seeing the model in practical use. Thus, focus is concentrated in limited but nevertheless important aspects and not a total standard. What we need to highlight is that the proposed temporary model is based on constructional engineering which support extensibility. That means that our starting-bricks are fundamental to the model but additionally should permit prospective object-classes being implemented at a later stage. Therefore, documentation marked-up in earlier versions remains usable, whilst the documentation marked-up in later versions will be of even better value, provided of course that the model is constructed satisfactory. However, information in the earlier versions could easily be converted into the latter, since the model in the latest version would be better specified with more proper object-classes and the information will be of better value and greater validity. As mentioned above the most important aspect besides setting up the model in an adequate and appropriate manner should be avoiding putting unnecessary bindings in our model today which could lead to unfortunate and damaging constraints tomorrow.[37]

Additionally, backward compatibility should be accomplished in an easier manner than the current state. As long as the newest version of the model is solely a more specified and thoroughly version of the older, and the model is from the commencement carefully designed avoiding strong bindings concerning prospective versions, backward compatibility could be carried out without any decisive problems during an object-oriented approach. As mentioned above, the first version shouldn't put restrictions on later versions. Thus, in the first version we are aware that some information needs to be put in some kind of a collection-container (some *Other* object-class), logically because we know that the model is not complete yet. However, we don't feel obliged to characterise this aspect as a serious weakness of the approach. If the specification claims changes regarding this collection group, aspects from this "other bag" could be pulled out and developed as unique object-classes and getting fully integrated in the model if necessary or desirable. We want to identify this present model being part of an incremental process, only during practical use could the model be in continuing progress.

Binding representation to a finite standard could result in obstacles when it comes to backward compatibility. This standard will appear fixed and newer version could contain features that are not present in the former and/or the reverse, the older version could have features that are not captured by the latter. Thus, there are potential compatibility problems. In the worst case, the tagged surveys would need to be, at least, partly marked-up once again in the latter, new version.

# 4. Extensibility

In moving towards an object-oriented solution the ability to extend the model concerning changed specification, prospective shortcomings in the model, etc, becomes available. This is not an available option in the current DTD, neither is it a built-in option in the XML DTD-technology itself. For every possible change one needs to write a new DTD, which is inconvenient and time consuming. Let us take a look at our diagrams. For instance, in figure 1

[37] While we are convinced that our proposal will be an improvement – not only compared to the current DDI DTD, but additionally to the other proposed alternatives, we don' t exactly know what units that should be implemented in the model in the first place. In our opinion, the proposed specification will be valuable, nevertheless, we need a more thorough discussion before picking out the fundamental bricks which will constitute the very basis for the model. Therefore, the model is by no means fixed or deadlocked, but will be an incremental process starting with a few, fundamental and hopefully agreeable, units, subsequently being extended.

we are not allowed to integrate a new element on any level without a re-write of the current DTD. In any meaning of re-structuring or re-specification of the model, this striking shortcoming results in at least a heavy workload. Contrary, if we want to re-structure our suggested model, specifically our example dealing with *File*, we are able to do so regarding changes we will establish on every level in the model. For instance, we have established an *Other* object-class with additional information we are not sure are relevant. After a while we identify that some information in this object-class is repeated for a large amount of the marked up surveys. Therefore, we find it necessary to introduce a new object-class dealing with this common information in our model. Our open-ended approach allows us to do so. Too many serious changes could unavoidably result in trouble when it comes to converting. Nevertheless, while we are starting with a few, but fundamental bricks building up our model, new object-classes are allowed to be implemented in the model in latter versions.

The lack of options for extensibility in the DTDs are well known and often mentioned as a salient reason to move away from this tool. However, the extensibility-potential is not exclusively an object-oriented domain. Additionally, it is a prevalent feature of for instance RDF and XML Schema.

# 5. Local (re)use of international standards

Another important feature of an object-oriented approach is the property of its *open-ended* structure. Basically, it means that every definition of an object belonging to a model is located as a resource somewhere at the web. Implementation of such a feature conveys a lot of advantages. For instance, it occurs an object called *Backup*, with x attributes in the international standard. Neither of these attributes tells us something about the actual placement of the backup, placement in a fireproof cabinet/cupboard, safe-deposit box, etc. A particular organisation, say NSD, are then able to establish an URI with its own namespace. By now, one creates a new object, lets for convenience call it *BackupStore*, which is defined as a subclass of *Backup* identified with various attributes identifying where the backup is located. Thus, the local object inherits all features from the international *Backup*, but is still able to establish its distinguishing features.

Another apparent gain with this open-ended approach is the rule that anyone who wants to participate will be accessed to do so. Therefore, if you want definitions to your object-classes in your model, but are not properly sure what would be a sufficient definition, searching through already established definitions, identified as resources stored somewhere on the web, could save time during your efforts towards establishing a model. Thus, you are in a fortunate situation where you are able to take advantage of other's experience simultaneously developing your own. As soon as this technology are spread on a wider scope and applications starts using the definitions of the object-classes will be extended on a wider range and probably tailor the particular definition you need in your specific model.

# 6. Summary

In conclusion, an object-oriented model for the DDI standard would not only allow the desired objectives of the mode flexibility and extensibility, but also give the standard a consistency and formal procedure for upgrades, take it beyond any short-lived present technology whilst still allowing compatibility between versions.

# PART 7b: Towards a object-oriented, representation-independent data model

## 1 Approach

We now start to define an object-oriented approach to defining the data model underlying the DDI. We use a subset of the UML notation to define the model. In keeping with the incremental approach discussed above, we start the model with the *file description* component of the DDI. The model fragments found in this section are primarily just examples of how OO models can be used to describe the characteristics of statistical metadata. At this stage we have not made any attempt to list all relevant attributes of the objects. Mainly for demonstration purposes we have for some objects indicated what operations may be used.

In Figure 1, we represent the *File Description* level from the current DDI DTD and its elements below. The total number of elements directly under the header *file Dscr* is 30. One half is shown in the diagram; for reasons of space, we have limited the diagram to two decimal places in the DTD structure. Later we will introduce an experiment trying to reverse engineer the *File* level into an OO model.

Relationships provide a pathway for communication between objects. If two objects need to "talk" there must be a link between them. Four common types of relationship:

a) A*ssociation*: a bi-directional connection between classes,
b) *Aggregation*: a stronger form of relationships where the relationship is a whole and its parts,
c) *Dependency*: a weaker form of relationship showing a relationship between a client and a supplier where the client does not have semantic knowledge of the supplier.
d). *Generalisation*: a sub-class relation – the sub-class inherits the properties and relationships of its generalisation; this is the major modularization mechanism in the model.

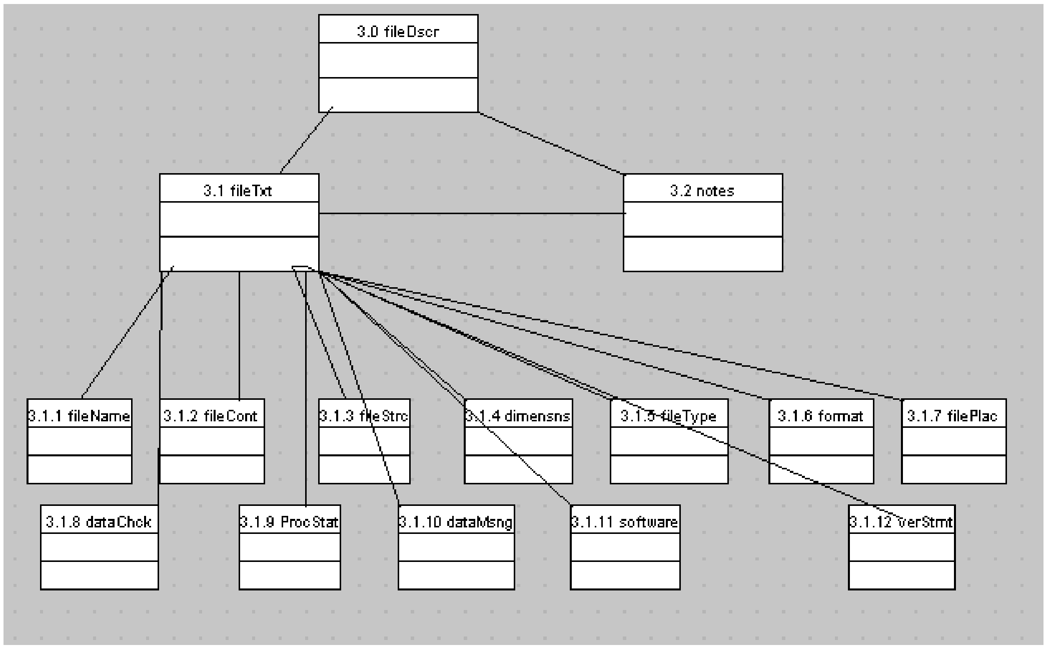


***Figure 1** DDI DTD fileDscr*

Inheritance is the crux of the matter in our approach, as in every OO-model. Thus, lower level object-classes (*subclass* or *descendants*) inherit attributes and methods from units above (*superclass* or *ancestors*) in the hierarchy. As customary in OO model approaches the most urgent challenge is to find the adequate level of abstraction when picking out and naming the object-classes. One convenient method would be to take the current DDI as a point of departure and use the prevailing Elements as objects (see the reverse engineering below). In our opinion, this approach would lead into a logical fallacy.[38]

On a higher level, classes and associations can be organised into *packages* representing those elements which are conceptually connected. In this document, packages are used to group concepts from different domains, representing the level of support which may be assumed from those domains. Thus, there are packages of classes for the Dublin Core elements, and for the DDI codebook for example. Further, we could expect that a specific data provider would provide a package of their own specific classes, inheriting as appropriate from the DDI package.

## 2 Top-level model

[38] Our motivation transforming the DTD is based on the assumption that we find it provably inappropriate as standard for different purposes (hierarchical files, time series, multidimensional tables, e.g. files that are not of the straightforward rectangular, survey-type vs. for instance the state of the Cristal Model which is capable of handling other file types as well, but certainly have other lacks and shortcomings). Therefore, we need another level of abstraction establishing the object-classes which should enter into the model. Finding the object-classes is not a very easy task, but should nevertheless be the first step in our efforts towards a solution we find more appropriate than the current DTD.

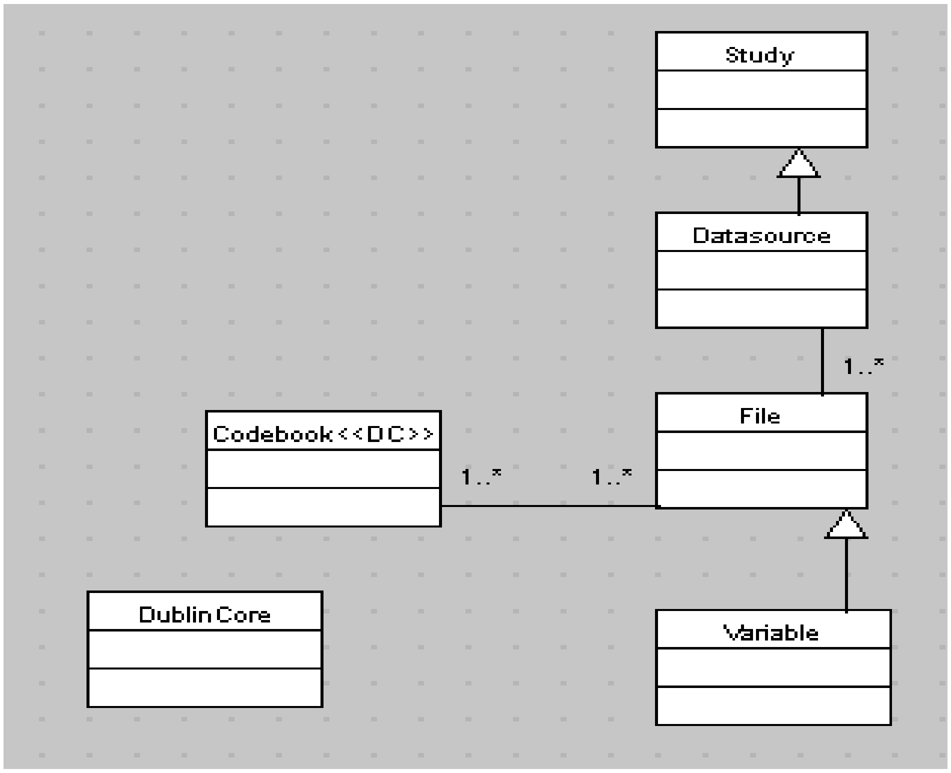


***Figure 2*** *Top level Codebook*

Figure 2 demonstrates the relationship between the main elements in the model. The *Codebook* is described at least at the *Dublin Core* level. The *Codebook* itself documents one or more files and may be the result of none or many Studies. A *Study* may use one or more *Datasource* and the *File* may contain information from one or more sources of data. Basically, this is how this information is organised in the current DTD, and in our opinion, there are few if any reason to change it.

# 3 The Filegroup Class

Within the DDI package, we have the filegroup class.

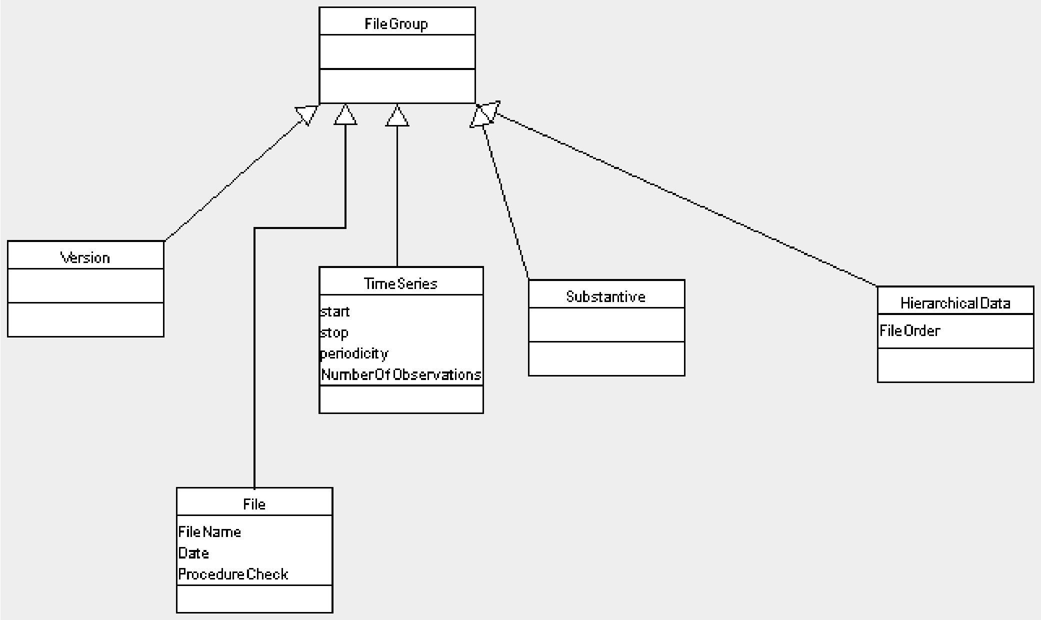


***Figure 3*** *Top level File-FileGroup*

Figure 3 introduces some new objects that are only partially found in the current DDI-DTD. The top object is *FileGroup*, and below we have suggested four possible examples that may prove useful when reporting searches as well as automated operations based on marked up metadata.

- *Version*. Contains pointers to all version of a given file.
- *TimeSeries*. Description of files that form a time-series, for example Level of living-studies.
- *Substantive*. Documents a group of files that have some meaningful similarities, for example ISP-studies, Eurobarometer etc.
- *Hierarchical Data*. Description on how a set of files form a hierarchy.

# 4 The Datasource classes

Also within the DDI package are the Datasource classes, and the related File and Variable classes.

The model fragment in figure 4 shows the major relationships between *Datasource*, *File* and *Variable*. The file contains a set of variables that comes from one or more sources of data. The three main objects *DataSource*, *File* and *Variables* will be documented differently depending on their type. A statistical table will have different attributes than questionnaires and similarly a standard rectangular file will have other attributes than a cube. If each of these objects are modelled separately it is easier for the metadata writer to know that information must be supplied and it is easier for an application to locate the information.

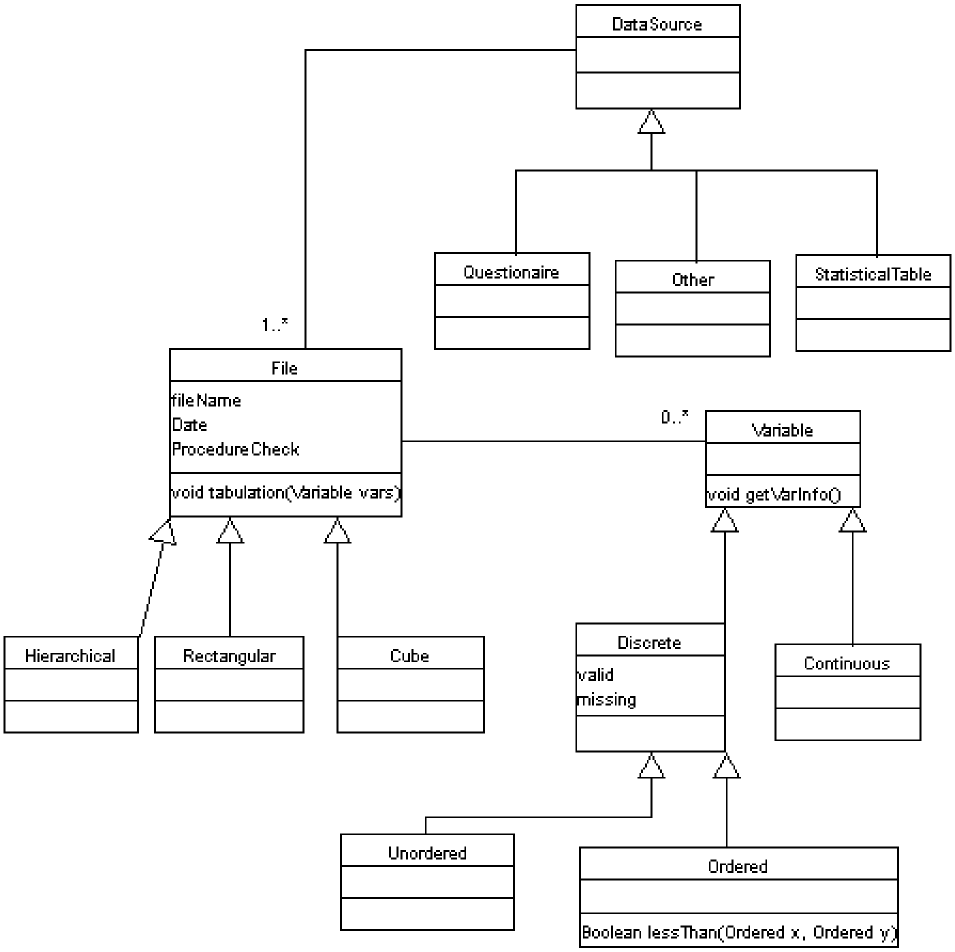


***Figure 4*** *Top level Variable*

# 5 Access Control and Inheritance

The model provides generic classes for providing access control features within the DDI package. Thus any object can inherit from the class *Restricted Object* which provides the association to an *Access Condition* class, defining the access control policy being used. The DDI package in the model also provides some common access control policies, such as embargo until a fixed date, or testing the IP number of the user, and also some constructors (*and*, *or*, *not*) for combining access conditions. Further access conditions can be provided in specific user packages, such as for instance those provided by the UKDA.

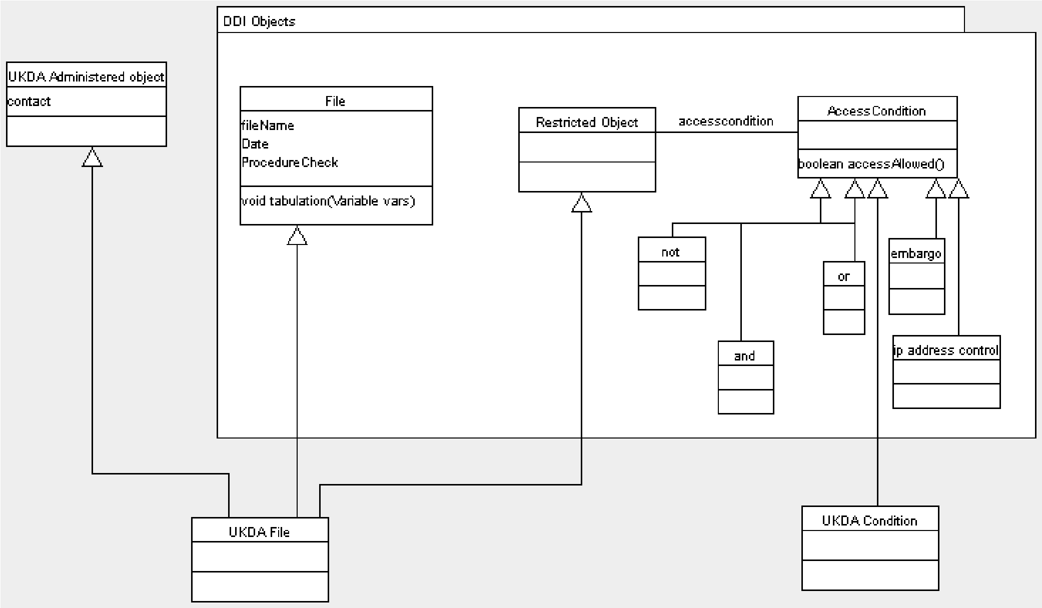


***Figure 5** access control model*

Thus in Figure 5 we can identify the inheritance structure between the object-classes. Starting from the *File* object-class within the DDI package, the UKDA provides a new class for their own files, *UKDA File*, inheriting from the standard DDI file description and thus still accessible from outside the UKDA in those terms. This new class has additional access control features by multiply inheriting from the general *Restricted Object* class, thus allowing an access condition to be added. This can be one of the ones within the DDI package, or one outside, such as the UKDA's own policy, or some combination of these. Further, the UKDA File class can inherit from the UKDA's own *Administrated Object* class, to provide administrative capabilities, such as the contact of the person within UKDA responsible for the file; this is a class within the UKDA's own package, and thus not accessible from outside.

# 6 File Format Class

From the (*File*) *Format* object-class, which is associated with the *File*, we can recognise the subclasses: *System Files* and *Common ASCII*, as in Figure 6. There may also be other subclasses of files. Taking a closer look at the difference in inheritance concerning the two classes of files, in Figure 7, while SPSS-, NSD, and other files where metadata is documentation in a systemfile on its own, we have common variables which we are able to identify on the upper level (*SystemFile*). Regarding *CommonASCII*, the variables for the two kinds of files are different. Thus, we have to identify them separately on the level below. One important question we need to take into consideration is how far we should go in finding common features between formats.

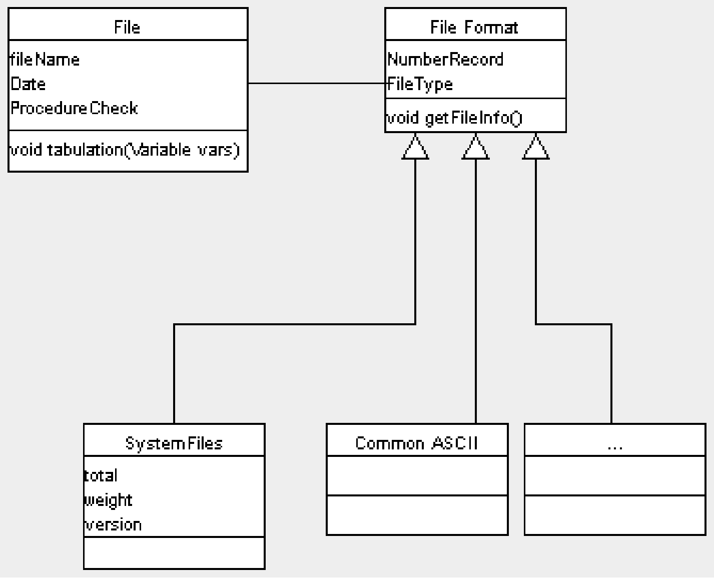


***Figure 6*** *Intermediate level for File*

For instance, we are able to find common attributes and variables in the first place between SPSS-, NSD-, and SAS-files. Below we are able to find similarities between SPSS and NSD-files but not SAS simultaneous on this particular level. Accordingly, we need to establish a new object-class for SPSS- and NSD-files, showing how they are mutually contradictory from SAS. Further, if we have to handle lots of different file-formats, which probably should be a characteristic feature of our model, our hierarchy will at some point be difficult to follow. Thus, the aim of parsimony could be overruled.

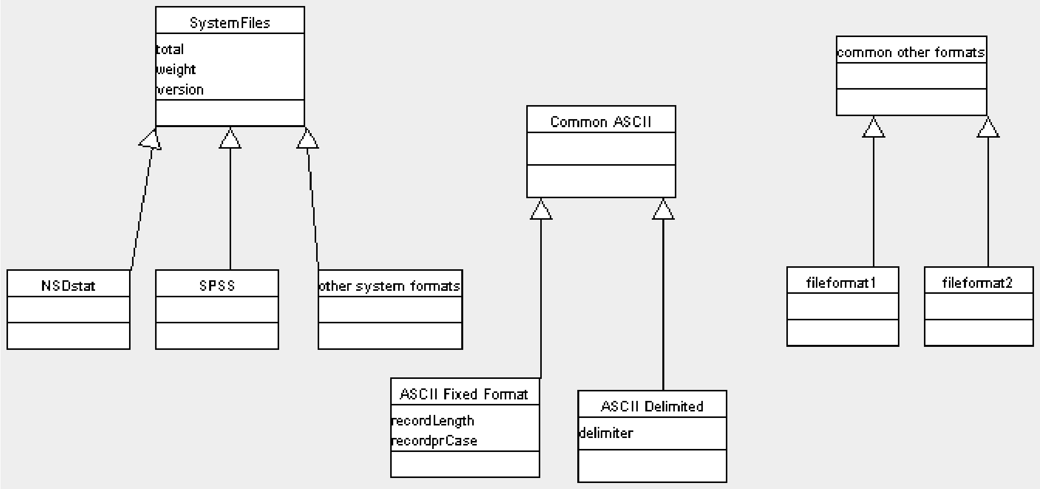


***Figure 7*** *Lower level for File*

# 7 The Study Description: Extending the Dublin Core[39]

In Figure 8 we sketch the possible relationship between the Dublin Core classes, collected together in one package for *Dublin Core Objects*, and the *DDI Objects*.

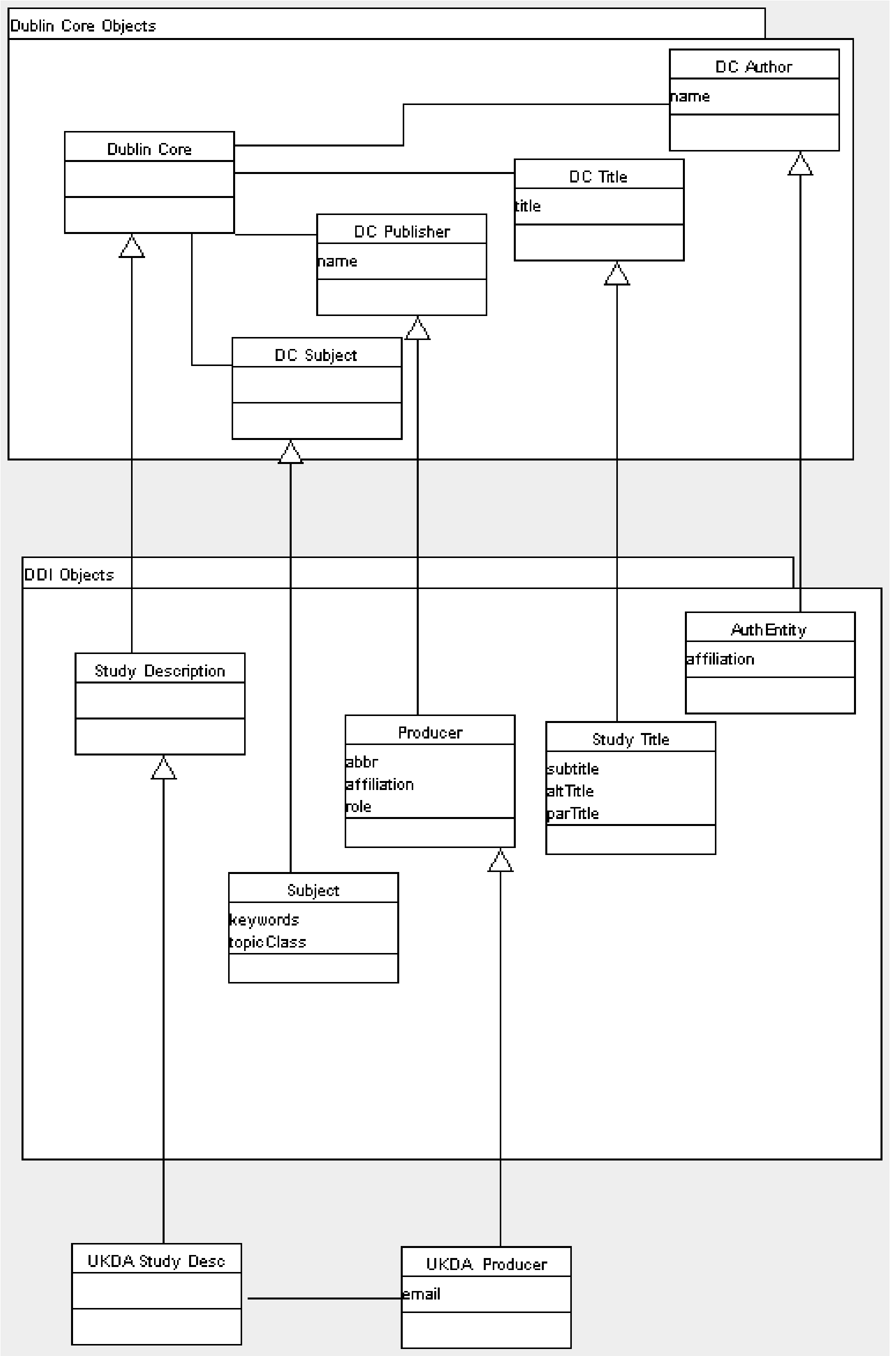


***Figure 8** The Study description and the Dublin Core*

[39] See part 5 for a more formal mapping of the DDI and the DC.

Thus the Dublin Core record object has an association with an object for each of the (in total – not shown) 15 fields within the original Dublin Core definition. The DDI *Study Description* is then a subclass in the DDI package, which inherits from the *Dublin Core* class, thus inheriting those associations. Each class for each field in the Dublin Core record then has its own specialisation within the DDI Objects package, providing the additional information specified by the DDI. Thus the same record will be accessible via a DDI or a Dublin Core view on the system. This can be carried out further. A study description as defined by the UKDA might provide additional information, using its own description of a producer for example.

We shall return to the problem of representing the Dublin Core as an abstraction or specialisation of the DDI, in more detail later in this deliverable.

# 8 Definition of concepts distributed on the net: An open-ended approach

This model follows the logic of (for instance) RDF and XML Schema. It means that every definition of a concept should appear public on the net visible for all possible, interested users. The basic assumption is that the Web is an open world; anybody can say anything about anything. The open-ended approach allows every conceivable participant filling in a definition of a concept appropriate for her model. Thus, the web works out as reference book of the definition of concepts in different OO-models. In such a manner, users will be able to reuse concepts already established, and not being coerced to necessarily start from scratch developing their concepts with adequate definitions, if it already exists somewhere on the web.

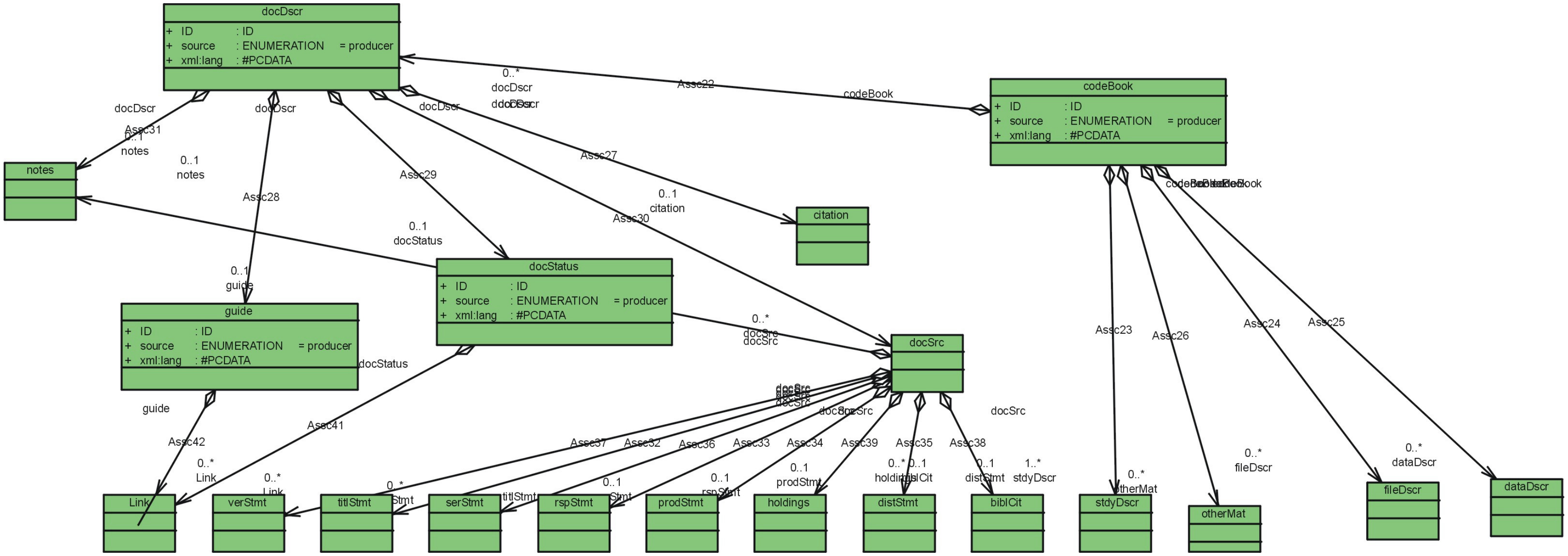


**Figure 9** Reverse engineering of the *File* level in the DDI DTD.

Generated in *Power Designer 7.5*

We have tested out different solutions converting the current DTD with an option called reverse engineering which we hoped should give us some ideas building up a new OO-model. The effort turned out being rather difficult to accomplish in an appropriate manner. By now, we are by no way convinced that Power Designer (used in the diagram) is the appropriate tool for this purpose. Therefore, we additionally tested market-leader Rational Rose for the same functionality. One apparent benefit of RR is the ability to generate different representations, for instance by the so-called “round-engineer”-option, e.g. both reverse- and forward-engineering, object-based and -oriented (program) languages (C++, Java, Visual Basic, etc.), XML, and other representation as well. After some bother we were able to read the model into RR. But unfortunately, it was tricky getting an actual diagram out of the model, and the model was totally unreadable. One fundamental shortcoming and logical defect of the nature of reverse engineering related to our approach: The different programs take naturally point of departure in the current DTD. We have elsewhere argued that this is not an optimal solution as starting point for our model.

# Part 8: Interoperating between DDI and the Dublin Core.

## 1. Introduction

Through the compilation of the WP3 deliverable plan there were suggested a draft outline for the metadata environment in the project. It was proposed that there should be a division of labour between FASTER and LIMBER, LIMBER input to the DDI-process, and how DDI should relate to other standards. Concerning the latter moment our motivation through this paper will be to prepare a standard gateway between the DDI and the Dublin Core standards. Our main concern would be to find the relevant objets in the area of statistical data, and further, how these objects should be described adequately. The basic assumption is that as long as the forthcoming Metadata Object Model (MOM) in the FASTER project encompass the administrative and technical parts of the DDI DTD, LIMBER should put their efforts against the documentation section, or the descriptive metadata type. We introduce a new classification which contains partially already existing sections of the DDI, and partially some new sections. We call our new proposed model the *Resource Discovery Object Model*.

As far as we are able to see there are three main areas in the current DDI DTD according to Swetland's scheme. The first area concerns the administrative-technical part, the second the documentation-technical, and the latter the statistical-technical, access control and its like. Further, our approach is supported by (one of) the deliverable concerning the Metadata Object Model (MOM) in the FASTER project which emphasise the Documentation Model as one of totally six models in the forthcoming MOM. According to the different types of metadata above, the proposed MOM in the FASTER project takes care of both the statistical-technical part and the administrative parts, simultaneously. According to this scenario, the LIMBER concern should be addressed against the descriptive parts, i.e. the documentation for resource-location.

## 2. Dublin Core

The Dublin Core Metadata Initative is a cross-disciplinary international effort to develop mechanism for the discovery-oriented description of diverse resources in an electronic environment. The DC Element Set comprises fifteen elements which together capture a representation of essential aspects related to the description of resources. Each element is optional and repeatable, and may further appear in any order.

Another way to look at Dublin Core is a "small language for making a particular class of statements about resources" (Baker, 2000). In this language, there are two classes of terms – elements and qualifiers, i.e. noun and adjectives – which can be arranged into a simple pattern of statements. The resources themselves are the implied subjects in the langauge.

# 3. DDI and DC

Thus, according to the LIMBER project the documentation technical part needs to be full-fledged. Our approach suggest be that a more or less restricted modification of the fifteen elements in the Dublin Core attends to this task. The proceeding arguments take point of departure according to this assumption. The table below shows the existing overlap between the current standards:

| DC | Docum.desc | Study desc | File desc | Data desc | Other mat |
|---|---|---|---|---|---|
| Title | **Titl 1.1.1.1**<br>subTitl* 1.1.1.2<br>altTitl* 1.1.1.3<br>parTitl* 1.1.1.4 | **Titl 2.1.1.1**<br>subTitl* 2.1.1.2<br>altTitl* 2.1.1.3<br>parTitl* 2.1.1.4 | **fileName 3.1.1** | | |
| Creator | **AuthEnty*1.1.2.1** | **AuthEnty* 2.1.2.1** | | | |
| Contributor | **othId* 1.1.2.2** | **othId* 2.1.2.2** | | | |
| Publisher | **Producer*1.1.3.1** | **producer*2.1.3.1** | | | |
| Date | **prodDate*1.1.3.3** | **prodDate*2.1.3.3** | | | |
| Subject | | **keyword*2.2.1.1**<br>topcClass*2.2.1.2 | | **concept 4.2.21** | |
| Description | | **abstract*2.2.2** | **fileCont 3.1.2**<br>notes 3.2 | **notes 4.3** | |
| Coverage | | **timePrd*2.2.3.1**<br>collDate*2.2.3.2<br>nation*2.2.3.3<br>geogCover*2.2.3.4<br>universe*2.2.3.7 | | universe 4.2.12 | |
| Rights | **copyright?1.1.3.2** | **copyright?2.1.3.2** | | **imputation 4.2.3**<br>**security 4.2.4**<br>**embargo 4.2.5** | |
| Type | | | **fileType 3.1.5** | | |
| Identifier | | **anlyUnit 2.2.3.6** | | **respUnit 4.2.6**<br>**analysUnit 4.2.7** | |
| Relation | | | **fileStrc 3.1.3** | **varGrp 4.1** | |
| Format | | | **format 3.1.6** | **varFormat 4.2.23** | |

# 4. Correlating the Dublin Core elements and the corresponding elements in the DDI

The elements Language and Source are omitted of reasons given elsewhere in this document. In our opinion, solely the first eight elements are of explicit LIMBER concern. The Rights element is a administrative task while the four remaining are of technical matters. Thus, the forthcoming MOM will encompass these certain areas. Explanation is given elsewhere.

The main issue with this deferment is to show that with a slightly more gentle definition of the respective elements, major parts of the current DDI are already covered by the Dublin Core, at least when it comes to the documentation-technical parts LIMBER aims to deal with.

Additionally, each Dublin Core element is defined using a set of ten attributes from the ISO/IEC 11179 standard for description of data elements. These include:

- Name – The label assigned to the data element
- Identifier – The unique identifier assigned to the data element
- Version – The version of the data element
- Registration Authority – The entity authorised to register the data element
- Language – The language in which the data element is specified
- Definition – A statement that clearly represents the concept and essential nature of the data element
- Obligation – Indicates if the data element is required to always or sometimes be present (contain a value)
- Datatype – Indicates the type of data that can be represented in the value of the data element
- Maximum Occurrence – Indicates any limit to the repeatability of the data element
- Comment – A remark concerning the application of the data element

Six of the above ten attributes are common to all the Dublin Core elements. These are with their respective values:

| | |
|---|---|
| Version: | 1.1 |
| Registration Authority | Dublin Core Metadata Initative (DCMI) |
| Language | en |
| Obligation | Optional |
| Datatype | Character String |
| Maximum Occurrence | Unlimited |

The definitions here include both the conceptual and representational form of the DC elements. The Definition attribute captures the semantic concept and the Datatype and Comment attributes capture the data representation.
The current recommended mapping between the two standards

| **DC** | **DDI** | |
|---|---|---|
| Title | titl | 1.1.1.1 |
| Creator | AuthEnty* | 1.1.2.1 |
| Contributor | othId* | 1.1.2.2 |
| Publisher | producer* | 1.1.3.1 |
| Rights | copyright? | 1.1.3.2 |
| Date | prodDate* | 1.1.3.3 |
| Subject | keyword* | 2.2.1.1 |
| | topcClass* | 2.2.1.2 |
| Description | abstract* | 2.2.2 |
| Coverage | timePrd* | 2.2.3.1 |
| | collDate* | 2.2.3.2 |
| | nation* | 2.2.3.3 |
| | geogCover* | 2.2.3.4 |
| | universe* | 2.2.3.7 |

Thus, nine out of fifteen elements in the DC have their identical equivalent/corresponding element in the DDI and a mapping could therefore easily be achieved if desirable. Note that the Source and Language elements in the DC and the *Source* and *Lang* attributes in the DDI, does not necessarily express the same kind of information. The former refer to the language of the source document, while the latter refer to the documentation language in the actual markup. For instance, if we want to describe a Spanish survey in German, Spanish would the element setting in Dublin Core, while German would be the attribute setting, in DDI. However, in the majority of the cases, the actual outcome would be identical. Keywords appear at the Subject Element, which has its corresponding elements in keyword (2.2.1.1) and topic Class (2.2.1.2) in the DDI. The remaining elements in the DC are more bibliographical oriented but not necessarily of minor importance.

One shortcoming in the current DC is the prevalent lack of controlled vocabularies. Shortly, we can define controlled vocabularies as a limited set of consistently used and carefully defined terms. The advantages of introducing such vocabularies are apparent: The improvement of search results, diminishment of incorrect and/or inconsistent data, reducing the likelihood of spelling errors when recording the metadata. Namespace could be considered as a controlled vocabulary. Nevertheless, the introduction of qualifiers (see below) would to some extent resolve this shortcoming.

## Namespaces used by the DC community

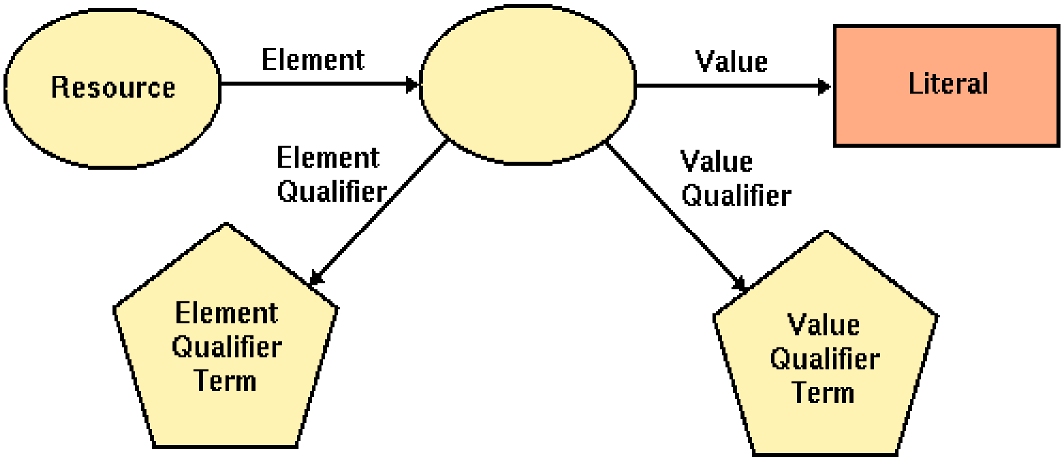
| Namespace<br>Definition | Namespace<br>Prefix | Namespace<br>URI | Namespace<br>Declaration |
|---|---|---|---|
| The Dublin Core<br>Namespace | dc | *http://purl.org/dc/elements/1.0/* | *xmlns:dc=" http://purl.org/ dc/elements/1.0/"* |

| | | | |
|---|---|---|---|
| The Dublin Core Qualifier Namespace | dcq | *http://purl.org/dc/qualifiers/1.0/* | *xmlns:dcq=" http://purl.org/ dc/qualifiers/1.0/"* |
| The Dublin Core Terms Namespace | dct | *http://purl.org/dc/terms/1.0/* | *xmlns:dct=" http://purl.org/ dc/terms/1.0/"* |

Users and implementors are free to adopt and support the basic Dublin Core, as represented by dc: on its own. Additionally, they may choose to embrace the extensibility mechanisms offered through dcq:. Finally, those using dcq: are further able to make use of commonly requested terms and values within the dct: namespace, and/or to utilise their own lists.

Another critique has been directed against the DC as long as the majority of work on the DC has addressed the definition of semantics rather than syntax or structure, allowing rapid development free of the constraints imposed by specific implementation environments. Whilst beneficial in many ways, this has lad to certain lack of clarity at times, especially in relation to development of 'qualification' mechanisms which enrich descriptions in the DC. It has also made interoperable implementation difficult, as individual implementors have typically developed their own internal mechanism for actually encoding DC; mechanisms which are not always compatible with those of their potential collaborators elsewhere.

Each Dublin core definition refers to the resource being described. A resource is defined as "anything that has identity". For the purpose of Dublin Core meatadata, a resource will typically be an information or service resource, but may be applied more broadly.



**An illustration of the Dublin Core Data Model**

Of major importance according to our approach are the *Elements Qualifier* and the *Value Qualifier*. The former is the solution proposed in order to meet the requirement of many implementors to specify aspects of a given resource with greater precision than is offered by the fifteen DC elements. The Element Qualifiers does not change the definition of the element, nor does it modify the value. Recommended practice is to use an Element Qualifier from a controlled list of Element Qualifier Terms defined for that element or from a domain specific list approved and maintained by a particular community.

Value Qualifiers are the solution proposed in order to meet the requirement of many implementors to specify the manner in which a value is encoded, normally with reference to some list of controlled terms or to a set of parsing rules. A Value Qualifier refers to the value, specifying either an encoding rule or a controlled vocabulary to aid in interpreting the value. Recommended practice is to use a Value Qualifier generally understood by the target community, or a standard encoding or parsing scheme in wide use across communities. Value Qualifier Terms represents permissible values for any given Value Qualifier. By separating the generic enabling mechanisms (the Value Qualifier) from the actual qualifications themselves (the Value Qualifier Terms) is remains possible to provide guidance on the ways in which qualification should be undertaken without the necessity to create an all-encompassing set of terms from which all communities should select.

Therefore, the DC Metadata Initative developed the “Dumb-Down Principle”. We will recommend co-opting this principle in our approach. First however, we need to establish what we mean by the notion of *qualifiers*. There are recognised two classes of terms in the DC; Elements and qualifiers, or nouns and adjectives. Each element in DC is optional and may be repeated. Additionally, each element has a limited set of qualifiers, attributes that may be used for further refinement of the elements meaning. A qualifier specifies an element with more restricted semantics.

### An example of qualifier lists (comprised of DC qualifier list 1 and 3)

| Element | Element Refinement | Encoding Scheme |
|---|---|---|
| Title | Alternative | |
| Subject | | LSCH,MeSH,DDC,LCC,UDC, (DDI) |
| Description | Table of Contents, Abstract | |
| Date | Created, Valid, Available, Issued, Modified | DCMI Period, W3C DTF, (DDI) |
| Relation | Is Version of, Has Version | URI |

| | | |
|---|---|---|
| | Is Replaced By, Replaces<br>Is Required By, Requires<br>Is Part Of, Has Part<br>Is Referenced By, References<br>Is Format Of, Has Format | |
| Coverage | Spatial | DCMI Point, ISO 3166,<br>DCMI Box, TGN, (DDI) |
| Coverage | Temporal | DCMI Period<br>W3C-DTF |

The DC Metadata Initative issued its lists of recommended DC Qualifiers in July of 2000. A set of recommended qualifiers is available. There are developed two main classes of qualifiers:

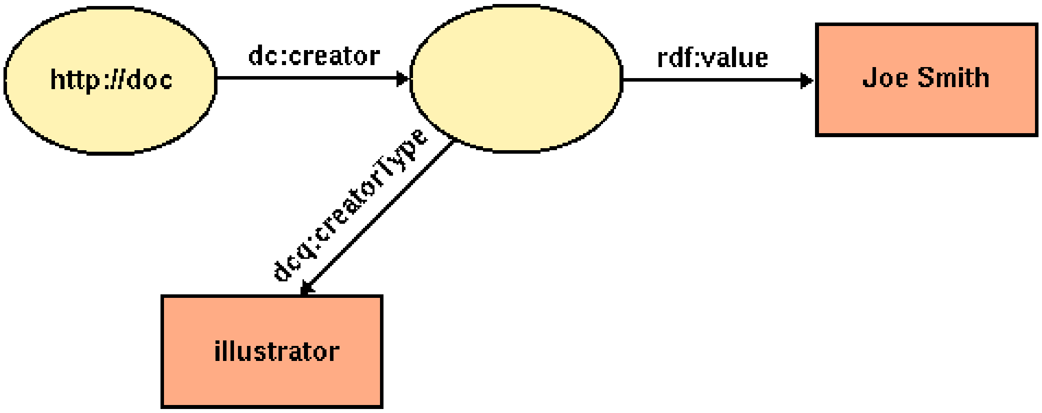
- **Element Refinement**, which makes the meaning of an element more specific or narrower. Thus, a refined element shares the meaning of the unqualified element, but with a more restricted scope, i.e. with established constraints. A client that does not understand a specific element refinement term should be able to ignore the qualifier and treat the metadata value as if it were an unqualified element. The definitions of element refinement terms for qualifiers need to be publicly available, i.e. somewhere on the web. Example: date – date created, dated published. An apparent parallel could be drawn to the suggestions concerning the organisations of the object-definitions in the MOM parlance.

- **Encoding Scheme**, which identifies schemes that aid in the interpretation of an element value. These schemes include controlled vocabularies and formal notations or parsing rules. A value expressed using an encoding scheme will thus be a token selected from a controlled vocabulary (e.g., a term from a classification system or set of subject headings) or a string formatted in accordance with a formal notation. If a encoding scheme is not (appropriately) understood by a client or an agent, the value may still be useful for human readability. The definitive description of an encoding scheme for qualifiers must be clearly identified and available for public use, eligible after the same principles as the Element Refinement. Example: LSCH-ISO 8601.

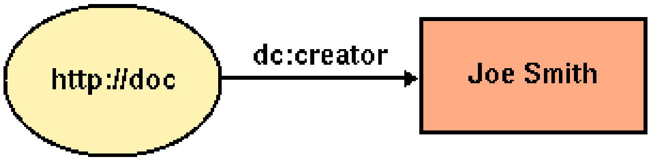
Lets move back taking a narrower interpretation of what the concept of the "Dumb-Down-Principle" contain. The approach states that the use of qualifiers as an additional level of detail introduces the situation where a client are able to encounter collections of resources that are described using the DC with qualifiers that are unknown to the client application. This can happen either because the client does not support qualifiers and the collection does, or the collection supports specialised qualifiers developed by implementors for specific local or domain needs. The useful interpretation of such descriptions will depend on the ability to ignore the unknown qualifiers and fall back on the broader meaning of the element in its unqualified form. The guiding principle of the qualification of DC elements, i.e. the “Dumb-Down Principle”, is that a client should be able to ignore any refinement and use the description as if it were unqualified. While this result in some loss of specific meaning, the remaining element value (minus the qualifier) must continue to be generally correct.

The former figure draws a picture of a qualified DC statement, whilst the latter shows the statement after dumbing down.

1)



2)



**The “Dumb-Down Principle”**

A syntactic RDF Representation of the above figure is as follows:

```
<?xml version='1.0'?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-
syntax-ns#"
         xmlns:dc = "http://purl.org/dc/elements/1.0/"
         xmlns:dcq =
"http://purl.org/dc/qualifiers/1.0/">
 <rdf:Description rdf:about = "http://doc">
  <dc:creator>
   <rdf:Description>
    <rdf:value> Joe Smith </rdf:value>
    <dcq:creatorType> Illustrator </dcq:creatorType>
   </rdf:Description>
  </dc:creator>
 </rdf:Description>
</rdf:RDF>
```

Note that the CreatorType is conveyed from the Element Qualifier (Element Refinement Qualifier), therefore the prefix dcq:, i.e. a namespace specifying qualifiers. CreatorType does not exist at the native DC, but the standard is enhanced and implementation would be allowed as long as the user points to the URI: xmlns:dcq = http://purl.org/dc/qualifiers/1.0/.

(An example of an Encoding Scheme qualifier could be:
<dc: Type ss:scheme="dcq:DCMIType"> Text </dc:Type>)

The Core elements are listed in the order they were developed originally, but there are other, more relevant ways to group or organise them. The proceeding classification divides the elements according to their content, intellectual property, and instantiation.

| **Content** | **Intellectual Property** | **Instantiation** |
|---|---|---|
| Coverage | Contributor | Date |
| Description | Creator | Format |
| Type | Publisher | Identifier |
| Relation | Rights | Language |
| Source | | |
| Subject | | |
| Title | | |

In the DDI OO model, the Dublin Core elements, become searchable objects, as in the following figure.

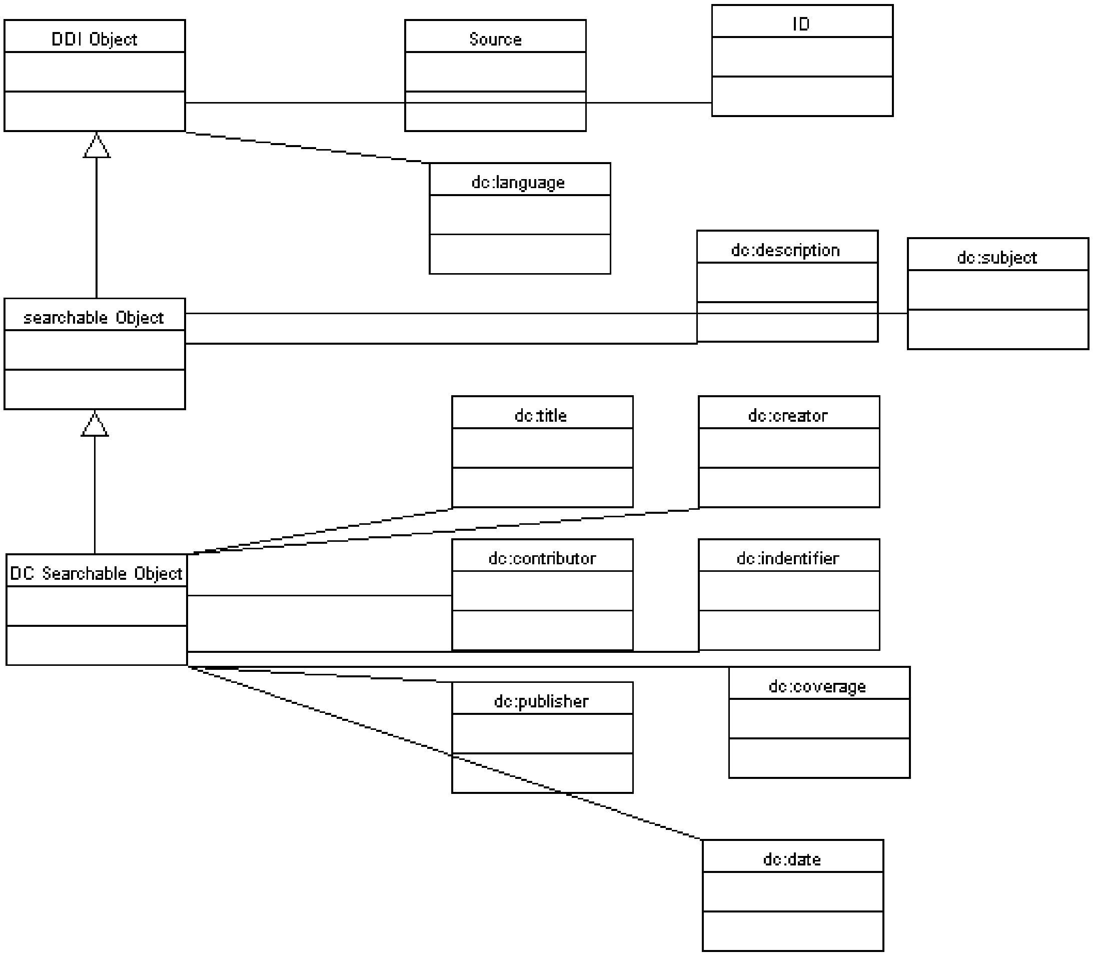


***Figure: The Dublin Core searchable objects in the DDI Metadata Model***

# 5. The Resource Discovery Object Model

Our proposal for a new design of the DDI, the *Resource Discovery Object Model:*

| | **SINGLE** | **GROUP** |
|---|---|---|
| **STUDY** | Election Study | Euro-barometer |
| **FILE** | Rectangular, survey-based type | Hierarchical, aggregation |
| **QUESTION** | | Background, battery |
| **ITEM/VARIABLE** | | |

| | | |
|---|---|---|
| **CUBE/TABLE** | | |

## 5.1 Discussion

Our suggestion will cover the Descriptive part in the table describing different types of metadata above. However, we recommend the label Resource Discovery Object recognising our model. Study, File, Question, Item, and Cube/Table appear as seeking-objects, i.e. those objects are apprehended as the relevant once for retrieving (meta)data. We propose the Dublin Core would be mapped to all cells in the table describing the Discovery part. Units stored below the header Group includes pointers to one or more other groups, automatically. Single does not by nature need the establishment of such a feature. The notions SINGLE and GROUP are not mutually exclusive as long as one survey could be contained in many Groups, i.e. fixed defined subgroups, and additionally one study could be extracted as the only relevant in one search but as part of a GROUP in another. For example, an ISSP survey could be interesting as both SINGLE and GROUP. If a researcher want to investigate the concept of *Trust in Authorities*, say, exclusively one out of this contains this specific term. Thus, none of the remaining ISSP surveys would be of interest in this current search.

However, this approach loses much of the structuring available in DDI 1.0 (and DDI 1.1). The solution to this is make the relationship to the searchable components in the Dublin Core at a lower level, so that the searchable components are at the leval of *distStmt, prodStmt etc,* as in the following diagram.

This approach has been trialed in RDF; a partial schema and a partial DDI codebook example are given in appendices C and D.

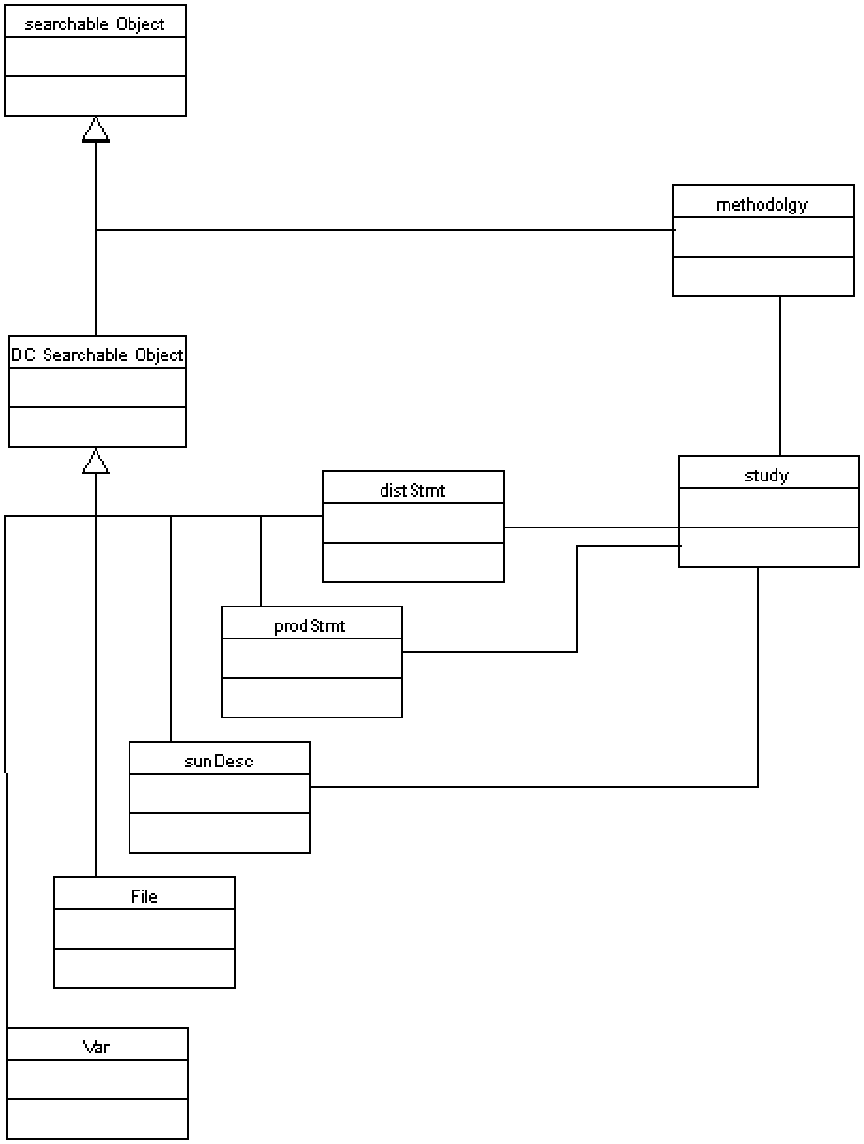


***Figure: DDI Components as searchable Dublin Core objects***

## 5.2 Motivations for change

A great advantage following our approach connecting to the established DC would be that one receives a set of formal, standardised definitions trough this widely used and accepted standard. Necessarily, we need to develop modifications of the DC elements, a more or less special tailored DDI modification of the DC. This modification would be stored on the web and invoked by use of the namespace tool in a RDF- or XML Schema framework. In prolongation of this argument there would be an apparent advantage that there exists a multitude of tools supporting the Dublin Core documentation standard.

Another benefit would be that we are able to distinguish between SINGLE and GROUP at every level. For instance, in the current DDI there are no way describing weather a survey is part of a bigger picture, say no. 4 of 9 surveys, totally. In our proposal, such a feature would be captured, distinguishing between the concept of SINGLE and the concept of GROUP. Additionally, the two dimensions are overlapping, not mutually exclusive, and could further be finer granulated into adequate and appropriate sub-populations.

We find it pretty bothersome and less flexible focusing the codebook as the crux of the construction of the DDI in an automatic, machine-comprehensible and –understandable world. The current DDI is concentrated around the replication of the codebook developed by data librarians in a human-created era and milieu. In an electronic world, the codeooks as entity would be of minor importance, almost obliterated, unless one needs to retrieve data from the printed archive. Therefore, we want to put our endeavours transforming the DDI into an automatic directed surrounding through our proposed Resource Discovery Object Model. Surely, we would not all propose neglecting the elements under section 1, but maintain that it is possible to generate the codebook anyway. The arguments emphasise the way elements are organised and how it possible in a more fruitful manner reconstructing a machine-understandable standard; in our view, a deft tool to raise the existing standard to another level of abstraction, i.e. a more generic solution[40].

[40] The current DDI is a blurred construction of technical, documentation, and administrative information. In our solution, where we separate these tasks, we are able to avoid such a mix of information, often comprehended as redundancy.

# APPENDIX A: a Fragment of the DDI in XML Schema

The following is a fragment of the result of passing the existing XML DTD for DDI1.0 through the IBM Alphaworks' tool DdbE. Here we have elements: `codeBook, docDscr, guide, docStatus, docSrc, stdyDscr, stdyInfo, subject,` and `keyword.` It can be seen that this automated method has produced a schema with a some redundancy: for example the attributes `xml:lang, source` and `ID` have been repeated for each element. This is because the parameterised entities in the DTD have been expanded out before the translation has taken place – introducing redundancy into the DTD. In a handcrafted schema, this could be factored out. Nevertheless, this does demonstrate that XML schema is a straightforward development path for DDI.

```
<schema xmlns='http://www.w3.org/1999/XMLSchema'>
<annotation>
  <documentation>
   XML Schema based on codebook.dtd
  </documentation>
</annotation>

<element  name="codeBook">
  <complexType  content="elementOnly">
    <sequence>
      <element name="docDscr"  maxOccurs='*' minOccurs="0"/>
      <element name="stdyDscr"  maxOccurs='*'/>
      <element name="fileDscr"  maxOccurs='*' minOccurs="0"/>
      <element name="dataDscr"  maxOccurs='*' minOccurs="0"/>
      <element name="otherMat"  maxOccurs='*' minOccurs="0"/>
    </sequence>
     <attribute name="xml:lang" type="NMTOKEN"  minOccurs="0"
/>
    <attribute name="source" type="NMTOKEN"  minOccurs="0"
       default="producer" > <simpleType base="NMTOKEN">
         <enumeration value='archive|producer' />
       </simpleType>
    </attribute>
    <attribute name="ID" type="ID"  minOccurs="0" />
  </complexType>
</element>

<element  name="docDscr">
  <complexType  content="elementOnly">
    <sequence>
      <element name="citation"   minOccurs="0" maxOccurs="1"/>
       <element name="guide"   minOccurs="0" maxOccurs="1"/>
      <element name="docStatus"   minOccurs="0" maxOccurs="1"/>
      <element name="docSrc"  maxOccurs='*' minOccurs="0"/>
      <element name="notes"   minOccurs="0" maxOccurs="1"/>
    </sequence>
    <attribute name="xml:lang" type="NMTOKEN" minOccurs="0"/>
    <attribute name="source" type="NMTOKEN"  minOccurs="0"
      default="producer" > <simpleType base="NMTOKEN">
        <enumeration value='archive|producer' />
      </simpleType>
    </attribute>
    <attribute name="ID" type="ID"  minOccurs="0" />
  </complexType>
```

```
</element>

<element  name="guide">
  <complexType  content="mixed">
    <group  maxOccurs='*' minOccurs="0">
      <choice>
        <element name="Link"/>
      </choice>
    </group>
    <attribute name="xml:lang" type="NMTOKEN"  minOccurs="0"
/>
    <attribute name="source" type="NMTOKEN"  minOccurs="0"
      default="producer" > <simpleType base="NMTOKEN">
        <enumeration value='archive|producer' />
      </simpleType>
    </attribute>
    <attribute name="ID" type="ID" minOccurs="0"/>
  </complexType>
</element>

<element  name="docStatus">
  <complexType  content="mixed">
    <group  maxOccurs='*' minOccurs="0">
      <choice>
        <element name="Link"/>
      </choice>
    </group>
    <attribute name="xml:lang" type="NMTOKEN"  minOccurs="0"
/>
    <attribute name="source" type="NMTOKEN"  minOccurs="0"
      default="producer" > <simpleType base="NMTOKEN">
        <enumeration value='archive|producer' />
      </simpleType>
    </attribute>
    <attribute name="ID" type="ID"  minOccurs="0" />
  </complexType>
</element>

<element  name="docSrc">
  <complexType  content="elementOnly">
    <sequence>
      <element name="titlStmt"/>
      <element name="rspStmt"   minOccurs="0" maxOccurs="1"/>
       <element name="prodStmt"  minOccurs="0" maxOccurs="1"/>
      <element name="distStmt"  minOccurs="0" maxOccurs="1"/>
      <element name="serStmt"   minOccurs="0" maxOccurs="1"/>
      <element name="verStmt"   maxOccurs='*' minOccurs="0"/>
      <element name="biblCit"   minOccurs="0" maxOccurs="1"/>
      <element name="holdings"  maxOccurs='*' minOccurs="0"/>
      <element name="notes"     minOccurs="0" maxOccurs="1"/>
    </sequence>
    <attribute name="xml:lang" type="NMTOKEN"  minOccurs="0" />
    <attribute name="source" type="NMTOKEN"  minOccurs="0"
      default="producer" > <simpleType base="NMTOKEN">
        <enumeration value='archive|producer' />
      </simpleType>
    </attribute>
    <attribute name="ID" type="ID"  minOccurs="0" />
    <attribute name="MARCURI" type="string"  minOccurs="0" />
  </complexType>
```

```
</element>

<element  name="stdyDscr">
  <complexType  content="elementOnly">
    <sequence>
      <element name="citation"  maxOccurs='*'/>
      <element name="stdyInfo"  maxOccurs='*' minOccurs="0"/>
       <element name="method"  maxOccurs='*' minOccurs="0"/>
      <element name="dataAccs"  maxOccurs='*' minOccurs="0"/>
      <element name="othrStdyMat"  maxOccurs='*'
      minOccurs="0"/>
    </sequence>
     <attribute name="xml:lang" type="NMTOKEN"  minOccurs="0"
/>
    <attribute name="source" type="NMTOKEN" minOccurs="0"
      default="producer" >
      <simpleType base="NMTOKEN">
        <enumeration value='archive|producer' />
      </simpleType>
    </attribute>
     <attribute name="access" type="IDREFS"  minOccurs="0" />
    <attribute name="ID" type="ID"  minOccurs="0" />
  </complexType>
</element>

<element  name="stdyInfo">
  <complexType  content="elementOnly">
    <sequence>
      <element name="subject"   minOccurs="0" maxOccurs="1"/>
      <element name="abstract"  maxOccurs='*' minOccurs="0"/>
      <element name="sumDscr"  maxOccurs='*' minOccurs="0"/>
      <element name="notes"   minOccurs="0" maxOccurs="1"/>
    </sequence>
    <attribute name="xml:lang" type="NMTOKEN"  minOccurs="0"
/>
    <attribute name="source" type="NMTOKEN"  minOccurs="0"
      default="producer" > <simpleType base="NMTOKEN">
        <enumeration value='archive|producer' />
      </simpleType>
    </attribute>
    <attribute name="ID" type="ID"  minOccurs="0" />
  </complexType>
</element>

<element  name="subject">
  <complexType  content="elementOnly">
    <sequence>
      <element name="keyword"  maxOccurs='*' minOccurs="0"/>
      <element name="topcClas"  maxOccurs='*' minOccurs="0"/>
    </sequence>
    <attribute name="xml:lang" type="NMTOKEN"  minOccurs="0"
/>
    <attribute name="source" type="NMTOKEN"  minOccurs="0"
      default="producer" > <simpleType base="NMTOKEN">
        <enumeration value='archive|producer' />
      </simpleType>
    </attribute>
    <attribute name="ID" type="ID"  minOccurs="0" />
  </complexType>
</element>
```

```
<element  name="keyword">
  <complexType  content="mixed">
    <group  maxOccurs='*' minOccurs="0">
      <choice>
        <element name="Link"/>
      </choice>
    </group>
    <attribute name="xml:lang" type="NMTOKEN"  minOccurs="0"
/>
    <attribute name="source" type="NMTOKEN"  minOccurs="0"
      default="producer" > <simpleType base="NMTOKEN">
        <enumeration value='archive|producer' />
      </simpleType>
    </attribute>
    <attribute name="ID" type="ID"  minOccurs="0" />
    <attribute name="vocab" type="string"  minOccurs="0" />
    <attribute name="vocabURI" type="string"  minOccurs="0" />
  </complexType>
</element>
```

# Appendix B: The Dublin Core Elements.

We give an overview of the DC elements and their respective attributes (with additional RDF-representation for the Dublin Core elements, from the *Guidance on expressing the Dublin Core within the Resource Description Framework (RDF)* <URL: http://www.ukoln.ac.uk/metadata/resources/dc/datamodel/WD-dc-rdf/WD-dc-rdf-19990701.html>, and RDF-representation for the qualifiers namespace for coverageType and coverageScheme):

## 1.1 Elements of the Dublin Core

### *1.1.1 Element: Title*

Name: Title
Identifier: Title
Definition: A name given to the resource.
Comment: Typically, a Title will be a name by which the resource is formally known.

```
<rdf:Description ID="title">
    <rdf:type rdf:resource="http://www.w3.org/TR/REC-rdf-
syntax#Property"/>
    <rdfs:label>Title</rdfs:label>
    <rdfs:comment>The name given to the resource, usually by the
Creator
    or Publisher.</rdfs:comment>
    <rdfs:isDefinedBy = ""/>
 </rdf:Description>
```

### *1.1.2 Element: Creator*

Name: Creator
Identifier: Creator
Definition: An entity primarily responsible for making the content of the resource.
Comment: Examples of a Creator include a person, an organisation, or a service. Typically, the name of a Creator should be used to indicate the entity.

```
 <rdf:Description ID="creator">
   <rdf:type rdf:resource="http://www.w3.org/TR/REC-rdf-syntax#Property"/>
   <rdfs:label>Author/Creator</rdfs:label>
   <rdfs:comment>The person or organization primarily responsible for creating the
intellectual content of the resource. For example,
authors in the case of written documents, artists, photographers, or illustrators in the case of
visual resources.</rdfs:comment>
   <rdfs:isDefinedBy = ""/>
 </rdf:Description>
```

### *1.1.3 Element: Subject*

Name: Subject and Keywords
Identifier: Subject

Definition: The topic of the content of the resource.
Comment: Typically, a Subject will be expressed as keywords, key phrases or classification codes that describe a topic of the resource. Recommended best practice is to select a value from a controlled vocabulary or formal classification scheme.

```
<rdf:Description ID="subject">
    <rdf:type rdf:resource="http://www.w3.org/TR/REC-rdf-
syntax#Property"/>
    <rdfs:label>Subject</rdfs:label>
    <rdfs:comment>The topic of the resource. Typically, subject will
be
    expressed as keywords or phrases that describe the subject or
    content of the resource. The use of controlled vocabularies and
    formal classification schemes is encouraged.</rdfs:comment>
    <rdfs:isDefinedBy = ""/>
  </rdf:Description>
```

### *1.1.4 Element: Description*

Name: Description
Identifier: Description
Definition: An account of the content of the resource.
Comment: Description may include but is not limited to: an abstract, table of contents, reference to a graphical representation of content or a free-text account of the content.

```
<rdf:Description ID="description">
    <rdf:type rdf:resource="http://www.w3.org/TR/REC-rdf-
syntax#Property"/>
    <rdfs:label>Description</rdfs:label>
    <rdfs:comment> A textual description of the content of the
resource,
    including abstracts in the case of document-like objects or
content
    descriptions in the case of visual resources.</rdfs:comment>
    <rdfs:isDefinedBy = ""/>
  </rdf:Description>
```

### *1.1.5 Element: Publisher*

Name: Publisher
Identifier: Publisher
Definition: An entity responsible for making the resource available
Comment: Examples of a Publisher include a person, an organisation, or a service. Typically, the name of a Publisher should be used to indicate the entity.

```
<rdf:Description ID="publisher">
    <rdf:type rdf:resource="http://www.w3.org/TR/REC-rdf-
syntax#Property"/>
    <rdfs:label>Publisher</rdfs:label>
    <rdfs:comment>The entity responsible for making the resource
    available in its present form, such as a publishing house, a
    university department, or a corporate entity.</rdfs:comment>
    <rdfs:isDefinedBy = ""/>
  </rdf:Description>
```

### 1.1.6 Element: Contributor

Name: Contributor
Identifier: Contributor
Definition: An entity responsible for making contributions to the content of the resource.
Comment: Examples of a Contributor include a person, an organisation, or a service. Typically, the name of a Contributor should be used to indicate the entity.

```
<rdf:Description ID="contributor">
    <rdf:type rdf:resource="http://www.w3.org/TR/REC-rdf-
syntax#Property"/>
    <rdfs:label>Other Contributors</rdfs:label>
    <rdfs:comment>A person or organization not specified in a Creator
    element who has made significant intellectual contributions to
the
    resource but whose contribution is secondary to any person or
    organization specified in a Creator element (for example, editor,
    transcriber, and illustrator).</rdfs:comment>
    <rdfs:isDefinedBy = ""/>
  </rdf:Description>
```

### 1.1.7 Element: Date

Name: Date
Identifier: Date
Definition: A date associated with an event in the life cycle of the resource.
Comment: Typically, Date will be associated with the creation or availability of the resource. Recommended best practice for encoding the date value is defined in a profile of ISO 8601 and follows the YYYY-MM-DD format.

```
<rdf:Description ID="date">
    <rdf:type rdf:resource="http://www.w3.org/TR/REC-rdf-
syntax#Property"/>
    <rdfs:label>Date</rdfs:label>
    <rdfs:comment>A date associated with the creation or availability
of
    the resource. Such a date is not to be confused with one
belonging
    in the Coverage element, which would be associated with the
resource
    only insofar as the intellectual content is somehow about that
    date. Recommended best practice is defined in a profile of ISO
8601
    [Date and Time Formats (based on ISO8601), W3C Technical Note,
    http://www.w3.org/TR/NOTE-datetime] that includes (among others)
    dates of the forms YYYY and YYYY-MM-DD. In this scheme, for
example,
    the date 1994-11-05 corresponds to November 5,
1994.</rdfs:comment>
    <rdfs:isDefinedBy = ""/>
  </rdf:Description>
```

### 1.1.8 Element: Type

Name: Resource Type
Identifier: Type
Definition: The nature or genre of the content of the resource.
Comment: Type includes terms describing general categories, functions, genres, or aggregation levels for content. Recommended best practice is to select a value from a controlled vocabulary (for example, the working draft list of Dublin Core Types. To describe the physical or digital manifestation of the resource, use the FORMAT element.

```
<rdf:Description ID="type">
    <rdf:type rdf:resource="http://www.w3.org/TR/REC-rdf-
syntax#Property"/>
    <rdfs:label>Type</rdfs:label>
    <rdfs:comment>The category of the resource, such as home page,
    novel, poem, working paper, technical report, essay, dictionary.
For
    the sake of interoperability, Type should be selected from an
    enumerated list that is currently under development in the
workshop
    series.</rdfs:comment>
    <rdfs:isDefinedBy = ""/>
  </rdf:Description>
```

### 1.1.9 Element: Format

Name: Format
Identifier: Format
Definition: The physical or digital manifestation of the resource.
Comment: Typically, Format may include the media-type or dimensions of the resource. Format may be used to determine the software, hardware or other equipment needed to display or operate the resource. Examples of dimensions include size and duration. Recommended best practice is to select a value from a controlled vocabulary (for example, the list of Internet Media Types defining computer media formats).

```
<rdf:Description ID="format">
    <rdf:type rdf:resource="http://www.w3.org/TR/REC-rdf-
syntax#Property"/>
    <rdfs:label>Format</rdfs:label>
    <rdfs:comment>The data format of the resource, used to identify
the
    software and possibly hardware that might be needed to display or
    operate the resource. For the sake of interoperability, Format
    should be selected from an enumerated list that is currently
under
    development in the workshop series.</rdfs:comment>
    <rdfs:isDefinedBy = ""/>
  </rdf:Description>
```

### 1.1.10 *Element: Identifier*

Name: Resource Identifier
Identifier: Identifier
Definition: An unambiguous reference to the resource within a given context.
Comment: Recommended best practice is to identify the resource by means of a string or number conforming to a formal identification system. Example formal identification systems include the Uniform Resource Identifier (URI) (including the Uniform Resource Locator (URL)), the Digital Object Identifier (DOI) and the International Standard Book Number (ISBN).

```
  <rdf:Description ID="identifier">
    <rdf:type rdf:resource="http://www.w3.org/TR/REC-rdf-
syntax#Property"/>
    <rdfs:label>Identifier</rdfs:label>
    <rdfs:comment>A string or number used to uniquely identify the
    resource. Examples for networked resources include URLs and URNs
    (when implemented). Other globally-unique identifiers, such as
    International Standard Book Numbers (ISBN) or other formal names
are
    also candidates for this element.</rdfs:comment>
    <rdfs:isDefinedBy = ""/>
  </rdf:Description>
```

### 1.1.11 *Element: Source*

Name: Source
Identifier: Source
Definition: A Reference to a resource from which the present resource is derived.
Comment: The present resource may be derived from the Source resource in whole or in part. Recommended best practice is to reference the resource by means of a string or number conforming to a formal identification system.

```
  <rdf:Description ID="source">
    <rdf:type rdf:resource="http://www.w3.org/TR/REC-rdf-
syntax#Property"/>
    <rdfs:label>Source</rdfs:label>
    <rdfs:comment>Information about a second resource from which the
    present resource is derived. While it is generally recommended
that
    elements contain information about the present resource only,
this
    element may contain a date, creator, format, identifier, or other
    metadata for the second resource when it is considered important
for
    discovery of the present resource; recommended best practice is
to
    use the Relation element instead.  For example, it is possible to
    use a Source date of 1603 in a description of a 1996 film
adaptation
```

```
    of a Shakespearean play, but it is preferred instead to use
Relation
    "IsBasedOn" with a reference to a separate resource whose
    description contains a Date of 1603. Source is not applicable if
the
    present resource is in its original form.</rdfs:comment>
    <rdfs:isDefinedBy = ""/>
 </rdf:Description>
```

### 1.1.12 *Element: Language*

Name: Language
Identifier: Language
Definition: A language of the intellectual content of the resource.
Comment: Recommended best practice for the values of the Language element is defined by RFC 1766 which includes a two-letter Language Code (taken from the ISO 639 standard), followed optionally, by a two-letter Country Code (taken from the ISO 3166 standard. For example, 'en' for English, 'fr' for French, or 'en-uk' for English used in the United Kingdom.

```
  <rdf:Description ID="language">
    <rdf:type rdf:resource="http://www.w3.org/TR/REC-rdf-
syntax#Property"/>
    <rdfs:label>Language</rdfs:label>
    <rdfs:comment>The language of the intellectual content of the
    resource. Where practical, the content of this field should
coincide
    with RFC 1766 [Tags for the Identification of Languages,
    http://ds.internic.net/rfc/rfc1766.txt ]; examples include en,
de,
    es, fi, fr, ja, th, and zh.</rdfs:comment>
    <rdfs:isDefinedBy = ""/>
 </rdf:Description>
```

### 1.1.13 *Element: Relation*

Name: Relation
Identifier: Relation
Definition: A reference to a related resource.
Comment: Recommended best practice is to reference the resource by means of a string or number conforming to a formal identification system.

```
<rdf:Description ID="relation">
    <rdf:type rdf:resource="http://www.w3.org/TR/REC-rdf-
syntax#Property"/>
    <rdfs:label>Relation</rdfs:label>
    <rdfs:comment>An identifier of a second resource and its
    relationship to the present resource. This element permits links
    between related resources and resource descriptions to be
    indicated. Examples include an edition of a work (IsVersionOf), a
    translation of a work (IsBasedOn), a chapter of a book
(IsPartOf),
    and a mechanical transformation of a dataset into an image
    (IsFormatOf). For the sake of interoperability, relationships
should
    be selected from an enumerated list that is currently under
    development in the workshop series.</rdfs:comment>
    <rdfs:isDefinedBy = ""/>
  </rdf:Description>
```

### *1.1.14 Element: Coverage*

Name:  Coverage
Identifier: Coverage
Definition: The extent or scope of the content of the resource.
Comment: Coverage will typically include spatial location (a place name or geographic coordinates), temporal period (a period label, date, or date range) or jurisdiction (such as a named administrative entity). Recommended best practice is to select a value from a controlled vocabulary (for example, the Thesaurus of Geographic Names [TGN]) and that, where appropriate, named places or time periods be used in preference to numeric identifiers such as sets of coordinates or date ranges.

```
 <rdf:Description ID="coverage">
   <rdf:type rdf:resource="http://www.w3.org/TR/REC-rdf-syntax#Property"/>
   <rdfs:label>Coverage</rdfs:label>
   <rdfs:comment>The spatial or temporal characteristics of the
intellectual content of the resource. Spatial coverage refers to a
physical region (e.g., celestial sector); use coordinates (e.g.,
longitude and latitude) or place names that are from a controlled
list or are fully spelled out. Temporal coverage refers to what the
resource is about rather than when it was created or made available
(the latter belonging in the Date element); use the same date/time
format (often a range) [Date and Time Formats (based on ISO8601), W3C Technical Note,
http://www.w3.org/TR/NOTE-datetime] as recommended for the Date element or time
periods that are from a controlled list or are fully spelled out.</rdfs:comment>
   <rdfs:isDefinedBy = ""/>
 </rdf:Description>
```

```
<rdf:Description ID="coverageType">
 <rdf:type rdf:resource="http://www.w3.org/TR/REC-rdf- syntax#Property"/>
 <rdfs:label>Coverage Element Qualifier</rdfs:label>
 <rdfs:comment>The type of coverage.</rdfs:comment>
 <rdfs:isDefinedBy = ""/>
</rdf:Description>
```

```
<rdf:Description ID="coverageScheme">
 <rdf:type rdf:resource="http://www.w3.org/TR/REC-rdf-syntax#Property"/>
 <rdfs:label>Coverage Value Qualifier</rdfs:label>
 <rdfs:comment>The encoding scheme or processing hint associated with the
coverage.</rdfs:comment>
 <rdfs:isDefinedBy = ""/>
</rdf:Description>
```

### *1.1.15 Element: Rights*

Name: Rights Management
Identifier: Rights
Definition: Information about rights held in and over the resource.
Comment: Typically, a Rights element will contain a rights management statement for the resource, or reference a service providing such information. Rights information often encompasses Intellectual Property Rights (IPR), Copyright, and various Property Rights. If the Rights element is absent, no assumptions can be made about the status of these and other rights with respect to the resource.

```
<rdf:Description ID="rights">
    <rdf:type rdf:resource="http://www.w3.org/TR/REC-rdf-
syntax#Property"/>
    <rdfs:label>Rights</rdfs:label>
    <rdfs:comment>A rights management statement, an identifier that
    links to a rights management statement, or an identifier that
links
    to a service providing information about rights management for
the
    resource.</rdfs:comment>
    <rdfs:isDefinedBy = ""/>
  </rdf:Description>
```

## 1.2 Qualification mechanisms

The following are the machine-readable, thus often effectively invisible Element Qualifiers. In opposition, the human readable labels are application specific, and will thus be more user friendly in terminology. The structure between the element in Dublin Core, the Element- and Value Qualifiers are as illustrated below

| **Element** | **Element Qualifie r** | **Value Qualifier** | **Value T ype** |
|---|---|---|---|
| *dc:*title | *dcq:*titleType | *dcq:*titleScheme | *rdf:*type |
| *dc:*creator | *dcq:*creatorType | *dcq:*creatorScheme | *rdf:*type |
| dc:*rights* | *dcq*:rightsType | dcq:rightsScheme | *rdf:type* |

The Dublin Core Terms Namespace (dct) at http://purl.org/dc/terms/1.0/ is reserved for these declarations, and the terms themselves are under discussion within the DC Working Group, continually. The exact manner in which terms will be drawn from the dct: namespace is still under discussion within the Working Group. The pointing-to- location from which the value will be drawn, does not scale well, and will probably be replaced by a more elegant solution.

# Appendix C: RDF Schema for DDI version 2.

```
<rdf:RDF xml:lang="en"
      xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
      xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">
<!--

This is the RDF Schema for the Data Documentation Initiative (DDI)
version 2.

http://www.icpsr.umich.edu/DDI

The Data Documentation Initiative (DDI) is an effort to establish an
international criterion and methodology for the content,
presentation, transport, and preservation of "metadata" about
datasets in the social and behavioural sciences. Metadata (data about
data) constitute theinformation that enables the effective,
efficient, and accurate use of those datasets.

Brian Matthews,

Limber Project,

Information Technology Department,
CLRC Rutherford Appleton Laboratory

-->


<!-- Generic class DDIObject: all classes in the DDI will be
subclasses of class DDIObject. -->
<!--
This allows us to:
   1. make any rdf class a resource, and
   2. define generic properties of any DDI class, such as language or
source.
-->


<rdfs:Class rdf:ID="DDIObject">
        <rdfs:comment>All class will be an element of the top-level
class.</rdfs:comment>
      <rdfs:subClassOf   rdf:resource="http://www.w3.org/2000/01/rdf-
schema#Resource"/>
</rdfs:Class>

<rdfs:Class rdf:ID="SearchableObject">
        <rdfs:comment>Those object which can be searched on within
the DDI.</rdfs:comment>
      <rdfs:subClassOf rdf:resource="#DDIObject"/>
</rdfs:Class>

<rdfs:Class rdf:ID="DCSearchableObject">
```

```
        <rdfs:comment>Those object which can be searched using Dublin
Core properties.</rdfs:comment>
      <rdfs:subClassOf rdf:resource="#SearchableObject"/>
</rdfs:Class>

<rdfs:Property
rdf:resource="http://purl.org/dc/elements/1.0/description">
        <rdfs:comment>We want to allow DC descriptions elements on
any searchable DDI Object.</rdfs:comment>
        <rdfs:domain rdf:resource="#SearchableObject"/>
</rdfs:Property>

<!-- DC Searchable objects. -->

<rdfs:Property
rdf:resource="http://purl.org/dc/elements/1.0/subject">
        <rdfs:comment>We want to allow DC subject elements on any
searchable DDI Object.</rdfs:comment>
        <rdfs:domain rdf:resource="#SearchableObject"/>
</rdfs:Property>

<rdfs:Property rdf:resource="http://purl.org/dc/elements/1.0/title">
        <rdfs:comment>We want to allow DC title elements on any DC
searchable DDI Object.</rdfs:comment>
        <rdfs:domain rdf:resource="#SearchableObject"/>
</rdfs:Property>

<rdfs:Property
rdf:resource="http://purl.org/dc/elements/1.0/creator">
        <rdfs:comment>We want to allow DC creator elements on any DC
searchable DDI Object.</rdfs:comment>
        <rdfs:domain rdf:resource="#SearchableObject"/>
</rdfs:Property>

<rdfs:Property
rdf:resource="http://purl.org/dc/elements/1.0/contributor">
        <rdfs:comment>We want to allow DC contributor elements on any
DC searchable DDI Object.</rdfs:comment>
        <rdfs:domain rdf:resource="#SearchableObject"/>
</rdfs:Property>

<rdfs:Property
rdf:resource="http://purl.org/dc/elements/1.0/identifier">
        <rdfs:comment>We want to allow DC identifier elements on any
DC searchable DDI Object.</rdfs:comment>
        <rdfs:comment>This identifier is an extenally assigned id -
not the same as the objects unique ID.</rdfs:comment>
        <rdfs:domain rdf:resource="#SearchableObject"/>
</rdfs:Property>

<rdfs:Property
rdf:resource="http://purl.org/dc/elements/1.0/publisher">
        <rdfs:comment>We want to allow DC publisher elements on any
DC searchable DDI Object.</rdfs:comment>
        <rdfs:domain rdf:resource="#SearchableObject"/>
</rdfs:Property>
```

```
<rdfs:Property
rdf:resource="http://purl.org/dc/elements/1.0/coverage">
        <rdfs:comment>We want to allow DC coverage elements on any DC
searchable DDI Object.</rdfs:comment>
        <rdfs:domain rdf:resource="#SearchableObject"/>
</rdfs:Property>

<rdfs:Property rdf:resource="http://purl.org/dc/elements/1.0/date">
        <rdfs:comment>We want to allow DC date elements on any DC
searchable DDI Object.</rdfs:comment>
        <rdfs:domain rdf:resource="#SearchableObject"/>
</rdfs:Property>


<!-- All DDI Objects can have a unique identifier, language and
source. -->

<rdfs:Property
rdf:resource="http://purl.org/dc/elements/1.0/language">
        <rdfs:comment>We want to allow DC language elements on any
DDI Object.</rdfs:comment>
        <rdfs:domain rdf:resource="#DDIObject"/>
        <rdfs:range rdf:resource="#LanguageCode"/>
</rdfs:Property>

<rdfs:Property rdf:ID="ID" >
      <rdfs:comment>The unique identifier of the DDI
Object.</rdfs:comment>
        <rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-
schema#Literal"/>
</rdf:Property>

<rdfs:Property rdf:ID="Source" >
      <rdfs:comment>The source which provided this item of
information about the class.</rdfs:comment>
      <rdfs:domain rdf:resource="#DDI"/>
        <rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-
schema#Literal"/>
</rdf:Property>


<!-- should be one of these for each language in ISO 639 Language
Codes -->

<LanguageCode rdf:ID="de"/>
<LanguageCode rdf:ID="en"/>
<LanguageCode rdf:ID="es"/>
<LanguageCode rdf:ID="fr"/>


<!-- three useful superclasses -->

<rdfs:Class rdf:ID="StudyData">
        <rdfs:comment>
All elements within the summary data description (2.2.3). These
elements include: time period covered, date of collection, nation or
country, geographic coverage, geographic unit, unit of analysis,
universe, and kind of data.
```

```
        </rdfs:comment>
      <rdfs:subClassOf rdf:resource="#DCSearchableObject"/>
</rdfs:Class>

<rdfs:Class rdf:ID="MethodologyData">
        <rdfs:comment>
All elements within the study methodology and
processing section (2.3) including information on data collection and
data appraisal (e.g., sampling,sources, weighting, data cleaning,
response rates, and sampling error estimates).
        </rdfs:comment>
      <rdfs:subClassOf rdf:resource="#SearchableObject"/>
</rdfs:Class>

<rdfs:Class rdf:ID="DistStmt">
      <rdfs:subClassOf rdf:resource="#DCSearchableObject"/>
</rdfs:Class>

<rdfs:Class rdf:ID="ProdStmt">
      <rdfs:subClassOf rdf:resource="#DCSearchableObject"/>
</rdfs:Class>

<rdfs:Class rdf:ID="SumDesc">
      <rdfs:subClassOf rdf:resource="#DCSearchableObject"/>
<</rdfs:Class>

<rdfs:Class rdf:ID="File">
      <rdfs:subClassOf rdf:resource="#DCSearchableObject"/>
</rdfs:Class>

<rdfs:Class rdf:ID="Var">
      <rdfs:subClassOf rdf:resource="#DCSearchableObject"/>
</rdfs:Class>

<rdfs:Class rdf:ID="CodeBook">
        <rdfs:comment>  Top  level  Object  in  a  metadata  description
</rdfs:comment>
      <rdfs:subClassOf rdf:resource="#DDIObject"/>
</rdfs:Class>

<rdfs:Property rdf:ID="Citation">
      <rdfs:domain rdf:resource="#CodeBook"/>
        <rdfs:range rdf:resource="#StudyData"/>
</rdfs:Property>


<rdfs:Property rdf:ID="MethStudyRef">
      <rdfs:domain rdf:resource="#MethodologyData"/>
        <rdfs:range rdf:resource="#StudyData"/>
</rdfs:Property>

<rdfs:Property rdf:ID="StudyInfoRef">
      <rdfs:domain rdf:resource="#CodeBook"/>
        <rdfs:range rdf:resource="#StudyData"/>
</rdfs:Property>

<rdfs:Property rdf:ID="StudyDistRef">
      <rdfs:domain rdf:resource="#StudyData"/>
```

```
        <rdfs:range rdf:resource="#DistStmt"/>
</rdfs:Property>

<rdfs:Property rdf:ID="StudyProdRef">
      <rdfs:domain rdf:resource="#StudyData"/>
        <rdfs:range rdf:resource="#ProdStmt"/>
</rdfs:Property>

<rdfs:Property rdf:ID="StudySumRef">
      <rdfs:domain rdf:resource="#StudyData"/>
        <rdfs:range rdf:resource="#SumDesc"/>
</rdfs:Property>


</rdf:RDF>
```

## Appendix D: A Codebook example in RDF

```
<!-- Brian Matthews (Rutherford Appleton Laboratory) - Limber Project
-->
<rdf:RDF  xml:lang="en"  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-
syntax-ns#"                xmlns:rdfs="http://www.w3.org/2000/01/rdf-
schema#"
         xmlns:dc="http://purl.org/dc/elements/1.0/"
         xmlns:dcq = "http://purl.org/dc/qualifiers/1.0/"
         xmlns:ddi2="http://www.clrc.ac.uk/Limber/RDF/ddi2#"
>

<!--

This example captures the fragment of DDI 1.0 Code book as follows.

<![CDATA[
<codeBook>
<docDscr>
<citation>
<titlStmt>
<titl>XML codeBook for SN:67017</titl>
</titlStmt>
<prodStmt>
<prodDate date="1999-02-02">2 February 1999</prodDate>
</prodStmt>
</citation>
</docDscr>
<stdyDscr>
<citation>
<titlStmt>
<titl>Images of the World in the Year 2000</titl>
<subTitl>Great Britain National</subTitl>
<IDNo agency="UKDA">67017</IDNo>
</titlStmt>
<rspStmt>
<AuthEnty   affiliation="University   of   Essex.   Department   of
Sociology">Matthews, D.</AuthEnty>
<AuthEnty>Jenkins, R.</AuthEnty>
<dataCollector>Research Services Limited</dataCollector>
</rspStmt>

<distStmt>
<dataDist>The Data Archive</dataDist>
<depositr affiliation="University of Essex. Department of
Sociology">Matthews, D.</depositr>
</distStmt>

</citation>
<stdyInfo>
<subject>
<keyword>AGE DIFFERENCES attitudinal</keyword>
<keyword>AGGRESSIVENESS attitudinal</keyword>
<keyword>ALLIANCES attitudinal</keyword>
<keyword>ARMED FORCES attitudinal</keyword>
</subject>
<abstract>This inquiry into the views of the year 2000 held by the
younger generation took place under the auspices of the European
Coordination Centre for Research and Documentation in the Social
Sciences, established at Vienna, which was founded by UNESCO and
which is a division of the International Social Science Council at
Paris. The technical coordination was in the hands of the
International Peace Research Institute, Oslo, under the direction of
Johan Galtung.</abstract>
<abstract>To examine attitudes of people in the age group 15 - 40
years towards various aspects of the future, with particular
reference to war, peace and disarmament. The great attractiveness of
such an inquiry lies in comparing the results of countries with very
different political and philosophical backgrounds. Eleven countries
are covered by this study.</abstract>

<sumDscr>
<timePrd date="1967-00-00" event="single">1967</timePrd>

<nation>Cross-national</nation>
<nation>Great Britain national</nation>
<geogCover>GREAT BRITAIN</geogCover>
<anlyUnit>Individuals</anlyUnit>
<universe level="study">Adults</universe>
<universe level="study">British population 15 - 40 years
old</universe>
</sumDscr>
</stdyInfo>

</codeBook>

]]>


Brian Matthews,

Limber Project,

Information Technology Department,
CLRC Rutherford Appleton Laboratory

-->
```

```
<rdf:Description rdf:ID="">
     <rdf:type
rdf:resource="http://www.clrc.ac.uk/Limber/RDF/ddi2#DDIObject"/>
     <ddi2:ID></ddi2:ID>
     <dc:language
rdf:resource="http://www.clrc.ac.uk/Limber/RDF/ddi2#en"/>
     <ddi2:Citation>
      <ddi2:StudyData>
            <dc:title>XML codeBook for SN:67017</dc:title>
            <ddi2:StudyProdRef>
              <ddi2:ProdStmt>
                <dc:date>1999-02-02</dc:date>
              </ddi2:ProdStmt>
            </ddi2:StudyProdRef>
      </ddi2:StudyData>
     </ddi2:Citation>
     <ddi2:Citation>
      <ddi2:StudyData>
            <dc:title>Images of the World in the Year 2000</dc:title>
            <dc:title>Great Britain National</dc:title>
            <dc:identifier>UKDA 67017</dc:identifier>
            <dc:author>Matthews, D.</dc:author>
            <dc:author>Jenkins, R.</dc:author>
            <dc:contributor>Research Services Limited.</dc:author>
          <ddi2:StudyDistRef>
              <ddi2:DistStmt>
               <dc:publisher>The Data Archive</dc:publisher>
               <dc:contributor>
                  <rdf:Description>
                    <rdf:value>Matthews, D.</rdf:value>
                    <dcq:contributorType> Depositor
                    </dcq:contributorType>
                  </rdf:Description>
               </dc:contributor>
               </ddi2:DistStmt>
          </ddi2:StudyDistRef>
      </rdf:Description>
     </ddi2:Citation>
  <ddi2:StudyInfoRef>
    <ddi2:StudyData>
     <dc:subject>AGE DIFFERENCES attitudinal</dc:subject>
     <dc:subject>AGGRESSIVENESS attitudinal</dc:subject>
     <dc:subject>ALLIANCES attitudinal</dc:subject>
     <dc:subject>ARMED FORCES attitudinal</dc:subject>
     <dc:description>
This inquiry into the views of the year 2000 held
by the younger generation took place under the auspices of the
European Coordination Centre for Research and Documentation in the
Social Sciences, established at Vienna, which was founded by UNESCO
and which is a division of the International Social Science Council
at Paris. The technical coordination was in the hands of the
International Peace Research Institute, Oslo, under the direction of
Johan Galtung.  To examine attitudes of people in the age group 15 -
40 years towards various aspects of the future, with particular
reference to war, peace and disarmament. The great attractiveness of
such an inquiry lies in comparing the results of countries with very
different political and philosophical backgrounds. Eleven countries
are covered by this study.
```

```
     </dc:description>
     <ddi2:SumDesc>
         <dc:date>1967</dc:date>
         <dc:coverage>Cross-national</dc:coverage>
         <dc:coverage>Great Britain national</dc:coverage>
         <dc:coverage>GREAT BRITAIN</dc:coverage>
         <dc:coverage>Individuals</dc:coverage>
         <dc:coverage>Adults</dc:coverage>
         <dc:coverage>British population 15 - 40 years
old</dc:coverage>
     </ddi2:SumDesc>
    </ddi2:StudyData>
  </ddi2:StudyInfoRef>

</rdf:Description>
</rdf:RDF>
```

# Appendix E: BAZAAR STYLE METADATA IN THE AGE OF THE WEB - AN 'OPEN SOURCE' APPROACH TO METADATA DEVELOPMENT

**STATISTICAL COMMISSION and** Working
paper No. 4
**ECONOMIC COMMISSION FOR EUROPE** English
only

**CONFERENCE OF EUROPEAN STATISTICIANS**

**UN/ECE Work Session on Statistical Metadata**
(Washington D.C., United States, 28-30 November 2000)

Topic (i): Statistical metadata for dissemination

**BAZAAR STYLE METADATA IN THE AGE OF THE WEB - AN 'OPEN SOURCE' APPROACH TO METADATA DEVELOPMENT**

Submitted by The Norwegian Social Science Data Services[41]

**Invited paper**

## I. INTRODUCTION

1. Metadata is all about communication. Metadata might be looked upon as a structured conversation between the different persons, offices, organisations and even software processes working with a dataset all the way from the design process to the final users. The main purpose of this structured conversation is to make sure that all relevant information are

[41] Prepared by Jostein Ryssevik.

passed on from one station to the next and that all participants have a chance to add their own relevant knowledge to this information exchange.

2. Most producers of statistical data will look upon the end users as legitimate *receivers* of relevant metadata. The idea that end users also might *contribute* to the metadata conversation is more unfamiliar. What we might envisage are feedback systems where users of statistical data are allowed to share their experiences with other users as well as with people engaged in the creation of the data. This will include the ability to create links from the metadata to reports and other products of the research process, as well as systems where users are allowed to append comments, advises or warnings to the core body of the metadata. Metadata should consequently be looked upon as open and dynamic over the entire life span of a data source and the metadata conversation as multi-directional.

3. The aim of this paper is to discuss metadata standards and metadata development in the light of this communication perspective. We will also explore the consequences of the move to Internet and the Web as the dominant communication and dissemination medium for statistical information. Our assumption is (following the ideas of Marshall McLuhan) that the introduction of a new communication medium like the Web has an impact far beyond the structuring and packaging of content. New technology changes the very models of communication and creates new methods and patterns of collaboration. In the final section of the paper we are using one of the most interesting models of co-operation rooted in the Web-revolution - the open-source software development movement - to challenge our traditional monolithic view on metadata development as well as metadata standards.

## 2. METADATA AS COMMUNICATION

4. Many discussions on the nature of metadata are set in the conceptual triangle: "*finding*", "*understanding*" and "*assessing*".

- *Finding*: Metadata is facilitating high precision resource discovery. A user is never searching for numbers, but for concepts represented by numbers. Trough catalogue information, study descriptions, question texts, definition of concepts or descriptions of sampling procedures etc, users are able to locate the collection of numbers that might fulfil their data needs.

- *Understanding*: Metadata is giving meaning to numbers. Without human language descriptions of their various elements, data resources will manifest themselves as more or less meaningless collections of numbers to the end users. The metadata provides the bridges between the producers of data and their users and convey information that is essential for secondary analysts.

- *Assessing*: Metadata is giving end-user a chance to assess the quality and relevance of a collection of numbers. By describing methodologies and procedures, as well as features related to the context of a particular study, end users are allowed to decide whether or not a data collection is meeting their professional or scientific standards.

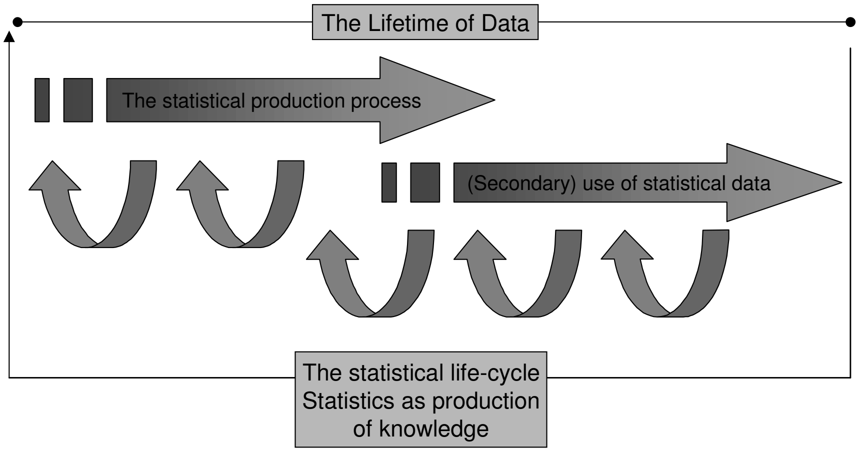
5. There is a fourth concept that is highly relevant to this discussion, and that is "*sharing*". The final product of the statistical production process is not the dataset (the collection of numbers) or even the table report. The dataset and the table report are both artifacts of a more general knowledge production process where the ultimate goal is to

improve the understanding of our physical and social environment. The evidence based knowledge production process is an activity with many groups of participants, each bringing different skills and resources to the table. It is also an activity that normally will be distributed in space as well as time. The majority of actors that are involved in this process have not been engaged in the creation of the data and have therefore no access to the "undocumented" and informal knowledge that follows from direct participation in data production. They might also be using the data for other research purposes than intended by the creators (secondary analysis) and will frequently do their analysis many ears after the data were collected.

6. The metadata might be seen as a facilitator for the interchange of information and insight that is the driving force of this process. Metadata makes it possible to extract knowledge from numbers and to share this knowledge with others. At the same time the conversation around the data and the various layers of knowledge products that derive from this conversation should become part of the metadata. For a secondary user of a data source it is of course important to get access to all relevant information about the data production process as provided by the data producer. However, it is also of immense value to get access to the knowledge of previous users, not only to avoid walking down analytical paths that are already fully explored, but also to learn from past experiences and to make it possible to add new approaches and new insight to the layers of already accumulated knowledge. Empirical research is one of the few arenas where it makes sense and indeed also is legitimate to stand on others shoulders.

7. The knowledge and insight acquired through the use of data is not only of value to other users, but can also be exploited by the data producers. Information about how data are used and evaluated by secondary analysts might provide important input to the data production process that on a longer term might improve data collection instruments and methods.

8. The communication perspective on metadata is summarised in the following diagram that uses feedback loops to illustrate how important information derived from the process is fed back into the


The Lifetime of Data
The statistical production process
(Secondary) use of statistical data
The statistical life-cycle
Statistics as production
of knowledge

system.
*Figure 1: Metadata in the statistical life-cycle*

9. The perspective leads to an *extended* metadata concept where not only descriptions of the data are relevant information, but also various types of knowledge products (formal as

well as informal) deriving from their use. It is also implying a *dynamic* concept where metadata is seen as a collection of information that is developed and enriched all the way through the life cycle of the dataset and not something that can be created and published once and for all. Finally, the perspective is leading to a concept where a broad spectre of actors is seen as legitimate contributors to the metadata holdings. Whereas the core metadata still are developed by the data producers as part of the data production and publishing process, further layers of metadata will be provided by others as an ongoing activity lasting for many years after the data themselves have left the production line.

## III. THE METADATA MEDIUM IS THE METADATA MESSAGE

10. The Web is about to become the dominant media for publication and dissemination of statistical information products, gradually replacing paper-based publications as well as other more static digital products (like CD-roms etc.). The move has initiated a discussion (within this group and elsewhere) on how statistical metadata should be designed, packaged and optimised to fit the new format. This discussion is important and should continue. The Web is radically different from the communication technologies that we know and are leaving behind and is therefore putting new and unfamiliar requirements and constraints on the content providers. To illustrate: an important metadata concept like footnote which makes perfectly sense in the paper-based world from which it originates, might at the best serve as a vague metaphor in a Web-based publication environment.

11. However, there are much deeper and more profound changes following in the wake of the Web-revolution than new requirements on metadata design. With the aphorism "the media is the message" the Canadian communication theorist Marshall McLuhan wanted to put a stronger focus on the medium as such, and not only on the messages or content that are delivered through the various communication channels (see McLuhan, Marshall 1964 and Levinson, Paul, 1999). According to McLuhan the very fact that we are using one media as opposed to another (TV instead of radio, or the Web instead of the printed publication), has a more significant impact on the way we think, work and collaborate, than the given content of any communication. Or applied to the topic of this paper: moving the dissemination of statistical information to the Web is not only affecting the structuring and design of the accompanying metadata. The move to the Web is changing the way statistical information is perceived and used in society and is thus altering the fundamental concept and function of metadata. In other words, the predominant communication technology has a deep impact on the structuring of the knowledge production process and is consequently affecting the way numbers and statistical evidence are linked into this process.

12. So, what are the significant features of the Web, which hold the potential of changing the very concept and function of metadata. At least the following should be mentioned:

❑ *From "one to many" to "many to many"*: The Internet and the Web is the first mass media that really challenges the traditional "from one sender to many receivers" model. The costs and skill requirements needed to provide content are reduced to a minimum allowing more or less everybody to become a "publisher". The Web has also several layers of formality allowing "quick-publishing" and informal exchange of ideas to live side by side with more formal contributions. Belonging to this picture is also the interactivity of the Web, which encourages the user to participate, not only to consume.

❑ *From publishing to collaboration*: The Web is gradually changing from a publishing media to an arena for collaboration totally independent of time and space. Tim Berners-

Lee (the only person that legitimately can name himself the initiator of the Web-revolution) has always seen the Web as an environment for collaboration, but admits that it has taken longer time than expected to reach this aim. However, the direction is unmistakably correct (see Berners-Lee, Tim, 1999).

- *From several local to a global hypertext space*: The Web is *hyperlinked*. It makes it possible to link one piece of information to another, more or less in the same way as the human brain snaps from one idea to the next by means of associations. It is also constituting *a global hypertext space* (breaking the confines of prior hypertext technologies), allowing information objects, totally independent of location or content, to be inter-linked. Moreover, there is no such ting as a linking authority weaving the Web on our behalf. Everybody can link other resources to their own or create resource-pages or portals that bring together information that according to the creator are related. In this way the Web should be seen as a collective effort growing like a "global brain". (for an interesting view on this aspect, see http://pespmc1.vub.ac.be/.)

- *The Web is genuinely multi-media*: The Web has taken as its content more or less all existing media. Text, pictures, animations, sounds, moving pictures, games or software tools - you name it - they are all available and inter-linked through a single interface, the ever-present Web-browser (note, that it is the browser and not the PC that is the integrating element - through WAP and other emerging technologies the Web functionality is about to break the confines of the computer screen).

- *The Web has memory*: The Web has the ability to remember artefacts as well as interactions and activities. Part of this memory is private (like local mailboxes, bookmarks- and history-lists), but the major part is public, constituting a huge public archive of "historic" documents as well as information exchange (like organised contributions to public news-groups and mail-lists etc.).

- *The Web has the" right" amount of standardisation*: W3C (The World Wide Web Consortium, which is the closest that we come to a Web authority) has taken care not to over-standardise the Web. W3C is developing and recommending the basic protocols (like HTTP), and the general and "all purpose" languages (like HTML, XML, RDF etc.), but these are only providing the necessary level of stability and interoperability to allow the industry as well as the domains (like the statistical community) to develop their own higher-level domain-specific standards. By using languages like XML or RDF to represent standardisation efforts, there is at least a chance that standards and schemas developed within different domains are able to "talk to each other".

13. The list of features could have been made longer. However, the intention of this paper is not to describe the Web in all its exciting details, but to discuss the relevance of some of the more basic traits for our understanding of statistical metadata.

## IV. METADATA IN THE AGE OF THE WEB

14. It should not come as a surprise that the communication perspective on metadata outlined above is pretty well served by Web technology. The Web is providing a communication platform that will allow us to establish a metadata-based conversation and to feed the important knowledge products deriving from this conversation back into the system. It is also a system that allows us to link various types of information objects like:

- data descriptions (traditional metadata),
- producer provided knowledge products like reports and fact sheets,
- user provided knowledge products like research notes, discussions, papers and articles,
- people (data collectors, domain experts, external researchers that have used a particular data source etc.)

15. The various elements will create a metadata-space, with layers of hyperlinked information. Closest to the data-core we find the traditional metadata elements developed by the data producers to allow the users to *find*, *understand* and *assess* the described data. Further from the core are objects that provide important contextual knowledge for secondary analysts, but which have been developed for other purposes than to serve as metadata. In the outer circles we will also find information provided by users, as opposed to producers and information that are of a more informal nature than the formal and structured information closer to the core (see Figure 2).

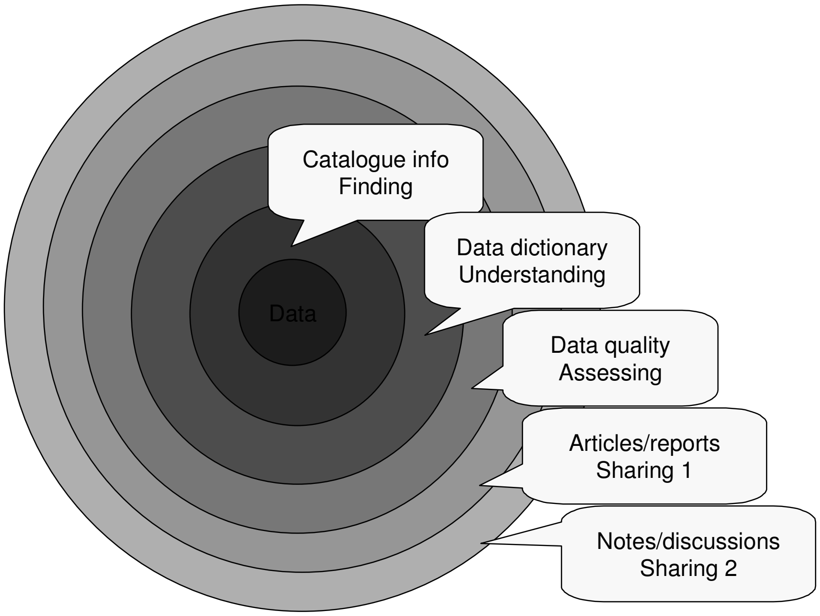


*Figure 2: Elements of a hyperlinked metadata-space*

16. A network of hyperlinks, which makes it easy to jump from one piece of information to the next, interweaves the entire metadata space. The links are not only providing an efficient navigation network. By associating conceptually related but physically dispersed information they are also adding knowledge to the system. (Part of) the knowledge are in the links.

*17.* *Scenario 1*: A user that is looking at a data source to test a theory might be alerted to other studies based on the same data or to comments on data quality provided by previous users. She might also be allowed to inspect the relevant contribution to a mail-list, which contains a discussion on the validity of the data for the type or research that she is doing and even to give her own contribution to this discussion. Further down the line she will be allowed to enter a link to her completed research paper to allow future users to take her work into account before they proceed on their own.

*18.* *Scenario 2*: A user that is reading an article in an on-line journal finds a link that connects him to the data that was used by the author to underpin the arguments. The link allows the reader to rerun the analysis, and also to dig deeper into the same data source. Through the metadata he is even made aware of several other sources that might be used to shed light on the phenomena and which might give another answer than the source used by the author. References or links to papers and articles based on these additional sources are of course also available.

19. Both scenarios are blurring the traditional division of roles between providers and users, authors and readers. In the first scenario a data user becomes a partner in the creation of metadata. In the second scenario a reader becomes a participant in the analysis of the data source. Both scenarios are facilitating cumulative research and both are made possible through hyperlinked metadata.

20. It might be argued that the above description is leading to a vague metadata concept that encompasses more or less all information that can be of any relevance to a user of a dataset. This is absolutely true, and we should even add "over the entire life-span of the data." For the users of a dataset, the "borders" of the metadata concept are totally irrelevant. We know that as we move away from the core metadata as described in Figure 2, we will sooner or later reach a territory, which hardly have anything to do with metadata. But to draw that border between metadata and non-metadata is of no more value than to draw the border between hot and cold water in the Atlantic Ocean. It is not the object as such that defines it as metadata or not, but the role it can play for potential users.

21. The Web is providing us with the technology and models, which might allow such metadata-spaces to grow. The role of the data producer is not to develop the entire space, but:

- ❑ to initiate the process and to publish the core metadata that provides the basic foundation for the metadata conversation
- ❑ to use metadata standards that are able to interoperate on the Web and are open-ended enough to include hooks to external information,
- ❑ to provide feedback systems that allow users to append or link information objects to the core

The rest of the process should be left to the users and to the collective energy that characterises the Web. It is indeed exactly in this way the entire "Web project" has grown and prospered over the last decade.

22. There is obviously not much left of the traditional monolithic view on metadata in this vision. Metadata is not seen as a coherent and centralised collection of information with clearly defined boundaries, provided by a single authority for a defined community of users. Rather we are looking at a multifaceted collection of information, distributed in space and constantly growing over time, created by a loosely connected network of contributors who are doing it for themselves as much as for any other potential participant.

## V. BAZAARS VERSUR CATHEDRALS

23. The approach to metadata development described above has several parallels to a phenomenon that has existed for quite a while, but which first caught attention outside its own narrow tribe of hackers with the release of Linux - the first operating system to challenge Microsoft on it's home ground. Linux is not developed according to the standard models and procedures of the software industry, but is the joint product of thousands of volunteer programmers collaborating over the Internet. By building a software system as complex and vulnerable as an operating system, the loosely connected Linux confederation managed to overturn many of the old-standing truths of the software industry. Overnight "the open source movement", of which the Linux tribe was a prominent member, reached the headlines of the computer press and convinced giants like Netscape, Oracle and IBM to join the party in order to keep up the innovation.

24. Eric S. Raymond, the self appointed participant anthropologist of the open source movement has described the new style of development as a "great babbling bazaar of differing agendas and approaches". In his now famous essay, published on the Web in 1998 (see Raymond E.S., 1998), Raymond compared this Bazaar style of development with the centralised and carefully planned Cathedral-building models of the software industry and he pointed to several distinguishing factors that may explain it's success.

25. One obvious factor is the total *openness* of the Bazaar model. The initiator of an open-source project makes all the source-code and documentation publicly available and invites others to criticise the approach and to come up with improvements. All the way trough the lifetime of the project all developed artefacts are open for inspection and discussion, and superior approaches are gradually replacing those that fail to convince. Open source projects thus resemble a *conversation* gradually moving foreword through exchange and evaluation of arguments.

26. Another important factor is the recognition of the value of different types of knowledge and the simple fact that many heads inevitably are better than a few. By excluding no one from the party, intelligently managed open source projects can master a pool of talent and profit from a constant flow of new perspectives and ideas that can boost innovation as well as quality.

27. According to Raymond, a third and very decisive factor is the belief in the value of the users. Users are treated as co-developers and not as receivers of a finalised commodity. By allowing users to make proposals or even to take part in the discussions among the developers, users can add value to the product. Or to site Raymond: “The next good thing to having good ideas is recognizing good ideas from your users”. Feedback is a key to success. As in our metadata-scenarios (see paragraph 17 and 18 above) the division of roles between providers and users, authors and reader is blurred – and again to the profit of the knowledge production process.

28. The real challenge of the open source model is to understand what drives thousands of smart programmers to spend parts of their highly valued time to develop code and (even to document the code so that everybody else can understand it) for free. What is the motivation? Again according to Raymond, open source hackers are not genuine altruists. The driving force behind their voluntary efforts is maximisation of reputation and status within the group. To invent the killer algorithm or to locate the faulty line of code that have caused the system to crash for no obvious reason allows the contributor to climb in the culture’s status hierarchy (see Raymond E.S., 1999 and Kuwabara, K., 2000).

29. Primitive as it may sound, the motivating force is not very different from the one we find in any non-profit academic research environment, including empirical social science that we are addressing in this paper. Merits are measured according to intellectual contributions and status allocation resides on a system of peer reviews.

30. Developing software and metadata are obviously two different things, so the parallel should not be over-stretched. However, the important thing to notice is that there are alternatives to cathedral building in software as well as metadata and that this alternative resides on a system of interaction, feedback and involvement of a variety of actors, many of whom we traditionally have classified as users.

## VI. TOWARDS THE DATA-WEB

31. Many of the ideas presented in this paper are implemented or about to be implemented in a Web-based data access and dissemination system called NESSTAR (Networked Social Science Tools and Resources), developed by the UK Data Archive, the Danish Data Archive (DDA) and the Norwegian Social Science Data Services (NSD). The basic philosophy of NESSTAR is to provide a software system that will allow producers and disseminators of statistical data to publish their resources on the Internet, either as standalone service to a limited group of identified users, or as an open offering connected to a distributed virtual data library. For the end users of statistical data, the system will provide a flexible interface that will allow them to search for data across the holdings of a broad range of data publishers, to brows detailed descriptions of the data (metadata), to visualise and analyse data on-line and to download data in a variety of formats ready for further local processing. (Musgrave, S. and Ryssevik, J. 1999 and 2000, more information are available at www.nesstar.org and www.faster-data.org)

32. NESSTAR is building upon an XML-based metadata standard developed by an international committee of data archives and data producers called the Data Documentation Initiative (DDI). This is a very end-user oriented metadata standard, allowing a rich amount of semi-structured information to travel along with the data on their way from the production line to the secondary analysts. One of the most important features of the DDI-standard is the ability to embed Web hyperlinks (URIs) in every metadata-element allowing external resources to be referenced or linked. This might include references to external knowledge products provided by the data publisher, as well as products and information provided by others (Ryssevik, J., 1999).

33. NESSTAR is supporting this feature of the DDI-standard. Moreover, the entire communication protocol of the NESSTAR system is based on messages composed as URIs. This is allowing every information object on a NESSTAR server to be hyperlinked or bookmarked, from within the NESSTAR client as well as from external Web objects. Any search for data, any dataset or any table or analysis derived from stored data can thus be described as a URI and activated from any other resource that are stored on the Web. NESSTAR is therefor providing a framework for bringing live data into on-line texts, as well as a framework for linking on-line scientific texts into the metadata body of a data material.

34. Both of the metadata scenarios described above (see paragraph 17 and 18) are consequently feasible in a NESSTAR environment. A research paper or report published on the Web as an HTML or PDF document can perfectly well include a hyperlink (URI) that gives the reader direct and live access to the underlying data stored on a NESSTAR server. The reader will be allowed to rerun the analysis, bring in new variables or even use the active

data as a springboard to find similar data sources that can throw light on the research topic in question. Starting from the other end - the data source - users will also be allowed to find relevant reports, papers or other documents that are based on the particular data.

35. What is lacking in the current NESSTAR system is the feedback loop that gives external users the chance to link their contributions directly into the metadata. In the current scenario links to external resources must be created by the data publisher - a procedure far to rigid to support the dynamic and open-ended metadata conversation that we have argued for in this paper. However, this technology is underway. What is aimed for is a system that allows the user of a dataset to create links from the metadata to reports and other products of the research process, as well as a system where users can append comments, advises, warnings or proposals to the core body of the metadata.

36. NESSTAR is one of several projects engaged in the development of what gradually has become known as the "data Web". Other relevant projects on this arena are the Virtual Data Centre currently under development at Harvard-MIT (King 1998) and the Ferret system developed by the U.S. Census Bureau (http://dataferrett.census.gov/). The common goal of all of these projects is to use open standards to build a true "data Web" where the models, technologies and collective energy of the Web is brought to the world of statistics.

37. Metadata specification languages like XML and RDF, developed by W3C, are providing important building blocks for this endeavour. However, to succeed we might be forced to rethink what we really mean by metadata standards and how we organise their development. In our current statistical information systems even standards are Cathedrals - they take ages to build and if ever completed they are literally "cut in stone". In the environment of the Web, development of new standards is normally measured in months, not in years. Any standard that takes more than a couple of years to develop bears the risk of becoming obsolete before it is even published.

38. Current metadata standards, including the DDI, are also static - they need to be revised in order to support a new requirement. What we will need in order to build the "data Web" are metadata standards that are flexible enough to evolve without initiating a costly and time-consuming revision process. In addition to an agreements on the key concepts and their relationships, the new generation of standards should include extensibility mechanisms that make it possible to add new concepts and relationships by building on what is already known.

39. Cathedral standards are based on an assumption that it is possible to reach global agreement on every little detail of a complex construction. There are few real-life examples to support this assumption. Following Tim Berners-Lee's vision of the Semantic Web we might be satisfied with "partial understandings", that is agreements on the key concepts and a common logical framework to express the local variations (Berners-Lee, 1999, pp. 201 - ).

## 1.1.1.1.1.1.1 REFERENCES

Berners-Lee, Tim (1999), "Weaving the Web - The Past, Present and Future of the World Wide Web by its Inventor", Orion Business Books, Great Britain.

King, Gary et.al. (1998) “An Operational Social Science Digital Data Library”, Proposal responding to NSF 98-63 Digital Data Library Phase II Program, Harvard University, Cambridge 1998, avialable at http://thedata.org/harum.pdf

Kuwabara, Ko (2000), "Linux: A Bazaar at the Edge of Chaos", published in the Web-Journal First Monday Volume 5, Number 3 - March 6th 2000, available at http://www.firstmonday.dk/issues/issue5_3/kuwabara/index.html

Levinson, Paul (1999), "Digital McLuhan - a Guide to the Information Millennium", Routledge, London

McLuhan, Marshall (1964), "Understanding Media - The Extensions of Man", MIT Press Edition, 1994.

Musgrave, S. and Ryssevik, J. (1999) “The Social Science Dream Machine. Resource Discovery, Analysis and Delivery on the Web”. Paper presented at IASSIST Conference "Building bridges, breaking barriers: the future of data in the global network", Toronto, May 1999. Available at http://www.nesstar.org/M_Paper.shtml

Musgrave, S. and Ryssevik, J. (2000) "Beyond NESSTAR: FASTER Access to Data" Paper presented at IASSIST Conference , Chicago, June 2000. Available at: http://www.faster-data.org/FASTER.doc.

Raymond, E.S. (1998), "The Cathedral and the Bazaar", published in the Web-Journal First Monday Vol.3 No.3 - March 2nd. 1998, available at http://www.firstmonday.dk/issues/issue3_3/raymond/index.html

Raymond, E.S. (1999), "Homesteading the Noosphere" published in the Web-Journal First Monday Vol.3 No. 10- October 5th. 1998, available at http://www.firstmonday.dk/issues/issue3_10/raymond/index.html

Ryssevik, J. (1999), "Providing Global Access to Distributed Data through Metadata Standardisation –
The Parallel Stories of NESSTAR and the DDI", Working paper no. 10 from the UN/ECE Work Session on Statistical Metadata, Geneva, September 1999.