CLUSTER OF SYSTEMS OF METADATA FOR OFFICIAL STATISTICS

EPROS Project Number	**IST-2000-26050**
Deliverable	**D12**
Workpackage	**1**
Title	**Final Report – v2**
Authors	**Joanne Lamb**
Date	**July 2003**
Contact	**UEDIN**
Project web address	**http://www.epros.ed.ac.uk/COSMOS**

PROJECT MANAGER

UEDIN Centre for Educational Sociology
Department of Education and Society, University of Edinburgh
St John's Land, Holyrood Road, Edinburgh EH8 8AQ, Scotland
Tel: +44 (0) 131 651 6276 Fax: +44 (0) 131 651 6239
Contact: Dr Joanne Lamb (J.M.Lamb@ed.ac.uk)

PROJECT PARTNERS

UKDA The Data Archive, University of Essex
Wivenhoe Park, COLCHESTER, CO4 3SQ, England
Tel: +44 1206 872321 Fax: +44 1206 872003
Contact: Simon Musgrave (simon@essex.ac.uk)

UoA Department of Mathematics, University of Athens
Panepistemiopolis, 15784 Athens, Greece
Tel: +30-1-72 74610 Fax: ++30-1-72 74119
Contact: Haralambos Papageorgiou (hpapageo@cc.uoa.gr)

DESAN DESAN Marktonderzoek BV
Raadhuisstraat 46, Postbus 10288, NL-1001 EG Amsterdam
Tel: +31 (0) 20 5207128 Fax: +31 (0) 6387229
Contact: Hans Rutjes (rutjes@mail.desan.nl)

SCB Statistics Sweden
PO Box 24 300, Karlavaegen 100, 104 51 Stockholm, Sweden
Tel: +46-8 5069 4148 Fax: +46-8 6615261
Contact: Prof Bo Sundgren (bo.sundgren@scb.se)

UU School of Information and Software Engineering, University of Ulster
Crowmore Rd, Coleraine, Co. Londonderry, BT52 1SA, Northern Ireland
Tel: +44 (0) 1265 324602 Fax: +44 (0) 1265 324916
Contact: Prof Sally McClean (SI.McClean@ulst.ac.uk)

WSEL World System (Europe) Limited
2 rue Alber Borchette, L-1246, Luxembourg-Kirchberg
Tel: +352 423113438 Fax: +352 424608
Contact: Yao Chen (e.c.y.chen@wsel.lu)

Dim Dimension EDI Ltd
High Trees, Elmbridge Road, CRANLEIGH, GU6 8JX, England
Tel: +44 1483 271443 Fax: +44 1483 271443
Contact: Chris Nelson (chris@dimension-edi.com)

Introduction

This report describes the COSMOS Cluster. In the Chapter 1, Objectives, we explain what a Cluster is, how the Cluster came about, and what its aims are. Next, in Chapter 2, we describe the five projects that make up the Cluster. In Chapter 3 we describe the process of bringing projects together, to identify the 'common core model'. Chapter 4 deals with the demonstrations that we achieved which illustrate the benefits of our approach. However, technical solutions are only part of the problem, so Chapter 5 discusses other issues that the cluster identified during its work. Chapter 6 reflects on what we have learned, and what we have achieved.. Chapter 7 relates COSMOS more directly with the projects from which it came, discussing a two-way influence between COSMOS and the projects. Finally, Chapter 8 looks at how the COSMOS work might be taken forward in the future, and draws some conclusions.

Contents

1. Objectives

COSMOS stands for Cluster of Systems of Metadata for Official Statistics. It is a collaboration between five research and development projects that have been funded by the European Union Fifth framework programme. In this research environment, a Cluster has a specific meaning, and should have specific aims.

'A cluster is a defined group of RTD activities. Its aim is to guarantee complementarity among projects, to maximise European added value within a given field and to establish a critical mass of resources at the European level.

An integrated approach towards research fields and projects financed is needed to solve complex multidisciplinary problems effectively. Clusters reflect this **problem-solving approach.** Indeed, in a cluster, projects are joined together because they complement each other in addressing major objectives in the context of a key action or a generic activity (sometimes even across different key actions or Specific Programmes). Clusters are expected to optimise scientific networking, management, co-ordination, monitoring, the exchange of information and, on voluntary basis, the exploitation and dissemination activities. The cluster may thus become a natural process to generate European added value, wherever it makes sense, beyond the limited resources of an isolated project. ….

Given the integrated nature of the programme, projects are encouraged to work together, to pool and to collectively build on their individual results whenever it makes sense to do so. Project clusters, each with their own specific objective, will be actively supported and encouraged in so far as they add value to the results of the IST programme seen as a whole. Whilst remaining a voluntary activity, it is anticipated that projects will find it to be in their own interest - and so worthwhile - to actively contribute to the work of specific clusters.' (IST, 1999)

The five project of the cluster already had a number of common partners when an informal meeting of the co-ordinators was held during the Eurostat Metadata Workshop, Workshop, held in Luxembourg in February 2000. At this point it was agreed to put together a proposal that aimed to bring the projects together, share experiences, and demonstrate the interoperability of the systems. We knew that all the projects were developing system to support statistical information processing and dissemination, and that each had a metadata model underlying the system. It seemed evident that, by brining these models together, we should obtain a better understanding of metadata and its uses. As is evident from the conference that we were attending, we knew of many initiatives on developing metadata and harmonising metadata ideas, and determined that our approach would be practical, with some software development to support our ideas.

The technical annexe of the successful cluster had the following objectives:

- to build better metadata repositories by exchanging ideas and experiences in using metadata systems for the individual projects;
- to identify a common set of metadata objects, with agreed definitions, attributes and methods;
- to implement a demonstration subset of these objects to show interoperability of the developed systems; and
- to define a methodology for further developing this interoperability.

We can identify a number of ideas behind these objectives. First, all the projects had some kind of idea of a repository of metadata: a store where metadata could be made available to

all software that could access the repository, and which shared an understanding of the semantics and structure of the metadata held there. We shall see later how this idea evolved. The second ideal was that there must be a common (sub)set of metadata where we could easily draw up a list of agreed terms and relationships. We shall also see that this was not as easy as anticipated. The third objective re-iterated our idea that there should be demonstrable software; that we should go further than conceptual or logical models. We feel this was the right decision, as it gave us insight and a focus, but the resources to do such an implementation were greater than anticipated.

The fourth objective was to keep a record of our activities, as the process of bringing the projects together was just as important as the results for these particular five projects.

With these objectives in mind, we devised a plan that had the following aims:

- To rapidly bring together the technical personnel on the projects, so that they could develop a shared understanding
- To ensure that all members of all projects (not just the key players in the cluster) should have at least one opportunity to meet and exchange ideas
- To finish with a conference to which external experts were invited, at which the developed software was demonstrated.

Formalising our sense of all working in the same are, we wrote the following in the proposal. 'The context of this proposal is the common goals and therefore potentially overlapping software developments being carried out by the projects involved in the Cluster. The aim of the Cluster is to identify the synergies in one particular aspect - metadata tools and repositories - and to map the various tools of the individual projects onto this common subset so that they can demonstrate interoperability using a core set of metadata objects. The relationship between the projects in the context of the statistical lifecycle of raw data collection, production, and dissemination [can be shown diagrammatically].'

The diagram (reproduced as Figure 1 in Section 3.6) showed how the different project addressed different parts of the statistical lifecycle, and this was to have a major bearing on our future findings.

2. The projects

Before going further, it is necessary to describe the five projects of the Cluster. This is done in three stages. First a table giving the basic details of the five projects is given. Then each project is described briefly: its objectives, partnership and results. Finally a comparison is done of the five projects. This was the first stage of the project, to allow us to understand each other's point of view and approach.

2.1 Basic details

Acronym	Full name	Project number	Start date	End date
FASTER	Flexible Access to Statistics, Tables and Electronic Resources	IST-1999-11791	1st Jan 2000	31st Mar 2002
IPIS	Integration of Public Information Systems and Statistical Services	IST-1999-12272	1st Feb. 2000	31st Jan 2003
IQML	A Software Suite and Extended Mark-up Language (XML) Standard for Intelligent Questionnaires	IST-1999-10338	1st Feb 2000	30th April 2003
METAWARE	Statistical Metadata Support for Data Warehouses	IST-1999-12583	1st Feb 2000	31st Jan 2003
MISSION	Multi-agent Integration of Shared Statistical Information over the (inter)Net	IST-1999-10655	1st Jan 2000	31st March 2003

2.2 The Five Projects

Each project description is leaded by details of the Coordinator and its website. A full list of partners is give, with participants in COSMOS given in italics

2.2.1 FASTER

COORDINATOR: University of Essex, The Data Archive
WEB SITE: http://www.faster-data.org/

PARTNERS:

Organisation	Country
University of Essex	UNITED KINGDOM
Universita degli Studi di Milano	ITALY
Central Statistics Office	IRELAND
Centre National de la Recherche Scientifique	FRANCE
Norwegian Social Science Data Services	NORWAY
DANSK DATA ARKIV	DENMARK
Centraal Bureau voor de Statistiek	NETHERLANDS
Statistisk Sentralbyraa	NORWAY

BRIEF DESCRIPTION:

FASTER aimed to develop a flexible metadata interface that would accelerate access to all types of statistical data.. A flexible user environment was established to enable users to identify and locate data resources regardless of physical location, and to browse the data and metadata in appropriate client applications. The objective was to develop an open architecture for the dissemination and use of statistics, based on XML, and to develop a configurable user environment, as a Web based client application or applet, in which the user was able to personalise their environment.

RESULTS
<input UKDA>

2.2.2 IPIS

COORDINATOR: Quality & Reliability International S.A.
WEB SITE: http://www.instore.gr/ipis/

PARTNERS:

Organisation	**Country**
Quality & Reliability International S.A.	LUXEMBOURG
National Statistical Service of Greece	GREECE
Ministry of Finance	GREECE
Université des Sciences et Technologies de Lille	FRANCE
Organisation for Vocational Education and Training	GREECE
University of Athens	GREECE
Centro de Formacao Profissional para o Comercio e Afins	PORTUGAL
Centre d'Etudes de Populations, de Pauvreté et de Politiques Socio-Economiques	LUXEMBOURG

BRIEF DESCRIPTION:

The general objective of IPIS was to develop new tools and services to enable public administrations to design, organise, develop and disseminate Public Information Systems (PIS) in a pre-harmonised and standardised way. It identified two areas of policy-making and their pertinent institutions. The scientific aim was to do an analysis of already existing software tools and services and then develop a new metadata- based system to assist Public Administration. in utilising and handling a Distributed Information System.

RESULTS
<input UoA>

2.2.3 IQML

COORDINATOR: The University of Edinburgh, Centre for Educational Sociology
WEB SITE: http://www.epros.ed.ac.uk/iqml

PARTNERS:

Organisation	Country
University of Edinburgh	UNITED KINGDOM
DESAN Marktonderzoek B.V.	NETHERLANDS
Comfact Ab	SWEDEN
National Technical University of Athens	GREECE
Dimension Edi Ltd	UNITED KINGDOM
Statistisk Sentralbyraa	NORWAY
Central Statistics Office	IRELAND

BRIEF DESCRIPTION:
IQML is aimed at supporting the collection of statistical data in an efficient way that also reduces the burden on respondents. It aimed to utilise metadata models and interchange standards that were being developed at the international level by the software industry. It sought to influence the development of these and to implement a solution for intelligent questionnaires.

RESULTS
<INPPUT UEDIN>

2.2.4 Metaware

COORDINATOR: Statistics Sweden
WEB SITE: http://metaware.wsel.lu/

PARTNERS:

Organisation	Country
Statistics Sweden	SWEDEN
World Systems (Europe) Limited	LUXEMBOURG
Hungarian Central Statistical Office	HUNGARY
Uniwersytet Warszawski	POLAND
Statistisk Sentralbyraa	NORWAY
Instituto Nacional de Estatistica	PORTUGAL
Statistics Denmark	DENMARK

BRIEF DESCRIPTION:

The implementation of data warehouse technologies in national statistical offices require a high degree of metadata support, in particular when data warehouses will be accessible for the public sector. The commercial packages on the market were not regarded as sufficient in this respect. The objectives of the project are the development of a standard metadata repository for data warehouses and standard interfaces to exchange metadata between data warehouses and the basic statistical production system. The aim is to make statistical data warehouse technologies more user-friendly for user access by the public sector.

RESULTS
<input SSB>

2.2.5 Mission

COORDINATOR: The University of Edinburgh, Centre for Educational Sociology
WEB SITE**:** http://www.epros.ed.ac.uk/mission

PARTNERS:

Organisation	**Country**
University of Edinburgh	UNITED KINGDOM
University of Ulster	UNITED KINGDOM
DESAN Marktonderzoek B.V.	NETHERLANDS
Central Statistics Office	IRELAND
Statistics Finland	FINLAND
University of Athens	GREECE
Office of National Statistics	UNITED KINGDOM

BRIEF DESCRIPTION:

The Web is perceived to have had a profound impact on the way National Statistical Institutes publish data. MISSION aimed to provide a software solution that will address the issues raised by the context of a global market to which Europe is moving. This software is aimed at allowing statistical data providers to publish data on the Web. Certain characteristics are identified, including ideas of supplying a flexible architecture that allows third parties to act as Independent Metadata Providers; allowing users to make requests in a declarative manner, and combine data from different sources; the tailoring of the users' environment, and the sharing of metadata.

RESULTS
<input Uedin>

2.3 Comparing the projects

The process of finding the common model was more complex than had been anticipated. An analysis by the University of Athens (Papageorgiou, 2002) compared each of the projects according to a number of headings. First each project was described in more detail than has been given in section 2.2. The headings for comparison were

- Main objectives
- Metadata
- Architecture

Other headings of interests to particular projects, such as *users*, *access control*, and *data processing*, were also included.
From the analysis it was shown that all five project supported the following objectives, directly or indirectly.

- Development of a standard metadata repository
- Use of the web for data dissemination
- Metadata collection and manipulation
- Use of a metadata model
- Support for Micro and macro-data
- Use of current state-of-the art technologies

The difference in approach were identified under the following headings:

1. Areas of Application
2. User communities supported
3. User services – Data collection
4. User services – data manipulation
5. User services – data import
6. User services –data export
7. User services – data browsing
8. User services – multilingual support
9. User services – visualizing
10. User services – harmonization
11. User services –other
12. Architecture
13. other issues

The results of the analysis are given below. The projects are identified by letter (F=Faster, P=IPIS, Q=IQML, W=Metaware, S=Mission)

(1) We found that two projects (F, P) addressed specific areas of application, which the other three took a generic approach, although all had some area of application for demonstration.
(2) Descriptions of the user communities differed, although we can identify three main groups: providers of statistical services; 'expert' end users (experienced either in statistics or the area of enquiry; and 'casual ' end users (not experienced in either statistics more the area of enquiry). All project supported the providers of statistical services, but the emphasis on whether this was for internal use or for dissemination varied.
(3) Two projects (P, Q) supported data collection.
(4) Three projects supported metadata manipulation (P, W, S), and four supported data manipulation (F, P, W, S)
(5) Four supported data import (F, P, Q, S)
(6) The same four supported metadata export
(7) These four also supported searching and browsing to a grater or lesser extent
(8) Two projects (Q, W) claimed to support all languages; one (F) distinguished between the contents of the system and the language of the system itself – the contents could be in native language. One (P) supported the four languages of the partners, and the other (S) identified only English as the language of the project.
(9) The responses to the questions on visualisation were varied, indicating a difference in interpretation of the question. The main responses were HTML, Java GUI, Excel, and not applicable.
(10) Two project (P, S) harmonised results, and four (P, Q, W, S) supported transformations.
(11) Other possible user services included access control (F, Q, W, S)) statistical disclosure control (one project (F) was in touch with the CASC project (CASC, 2001))
(12) All architecture was 3-tiered; three systems (F, W, S) were distributed and two centralised. Only one project (F) used a standard metadata model, although another had based its architecture on a standard domain model (P). Interchange mechanisms were XML, DDI, HTTP and SOAP. One project (S) used agent technology.

This analysis has shown how the projects differed in their approach and orientation. Some differences are fundamental to the project: its objectives and its user requirements. Other

may be more arbitrary, depending on the experiences and preferences of the organisations and individuals involved in the projects. This theme is pursued further in the next section.

3. Finding the common model

3.4 Introduction

One of the main objectives of the COSMOS project has been to develop a common model as a basis for interchange of metadata between cluster projects with different metadata models. Equally important has been to learn from the process of developing and implementing that model, in order to further the understanding of how to harmonize metadata from different sources and to learn how to build better metadata repositories.

The following text gives an inside view of the different steps and decisions that contributed to the scope and content of the final model and analysis those steps in terms of system analysis, system design and system development.

3.5 Comparative system analysis

3.5.1 Objectives and scope

The first step in the COSMOS project was to describe the different cluster projects and compare their general objectives and scope. This resulted in the so-called COSMOS projects profile and constituted a starting point for understanding the different projects on a general level. That work was summarized in two graphs, which help give an overview of differences and similarities between the cluster projects. The first graph (Figure 1)is based on a general input-process-output model of the statistical production process, including both archiving and design (questionnaire).

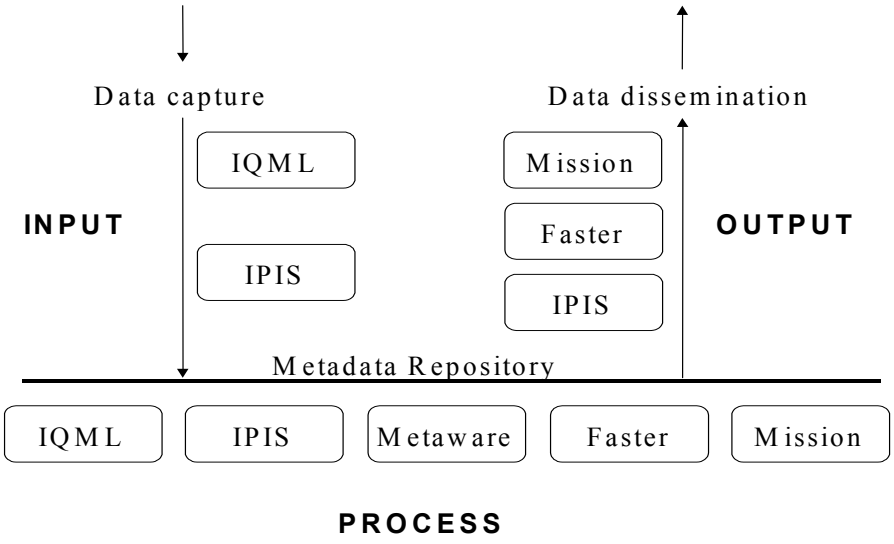
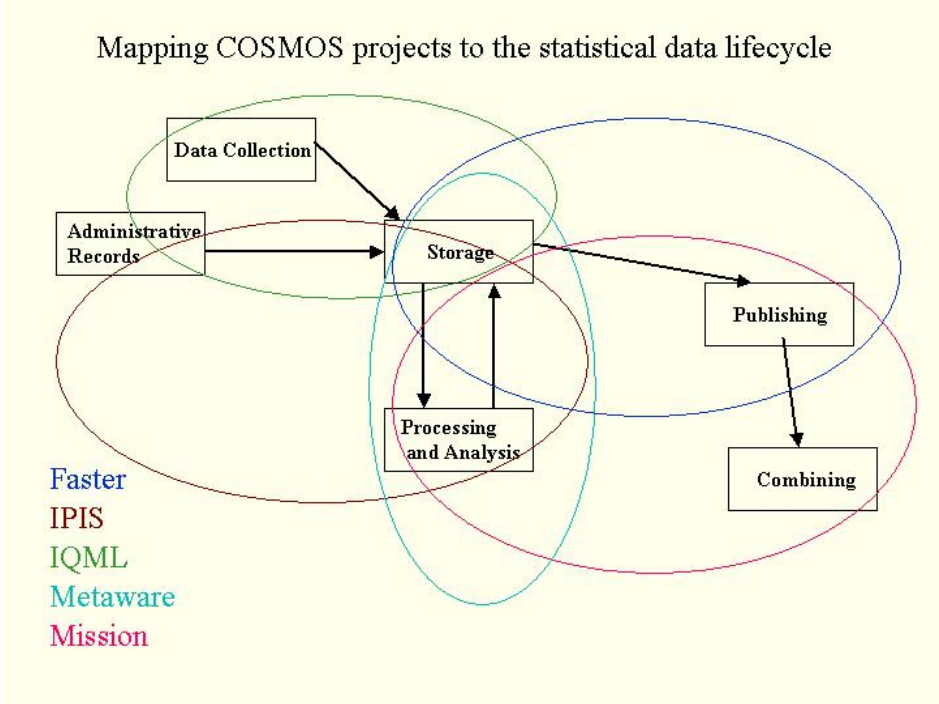


Figure 1:COSMOS projects in a general input-output model

IQML is purely design and input oriented. Faster is the project with the strongest focus on archiving. The other three projects are output-oriented, and have in common statistical data warehouse-oriented functionality. Thus Metaware, IPIS and MISSION are all oriented towards cubes, which can be manipulated dynamically, with operations such as selections and aggregations. The actual production processes (mainly editing and creation of final observations with derived micro variables) are as a rule not present in the cluster systems, even if both conceptual and process metadata from such processes may be included.

The second analytical model is based on different sub-processes in the so called statistical data lifecycle, which include different data manipulation processes and two different sub-types for data collection.



The graph clearly shows that some of the projects have unique domains of functionality or do not handle one or more of the processes. Notably, IQML is the only project that supports data collection; Mission the only one that supports combining and Faster and IQML are the only projects that do not include processing and analysis.

3.5.2 Domain comparisons

The next step in the system analysis, the purpose of which was to form a basis for the elaboration of a large federated common model, was a more detailed comparative analysis based on a number of domains for comparison. These tasks were divided between different members of the project. The domains for comparison were.

- Data input, questionnaire, survey
- Dataset/cubes
- Access control
- Transformations
- Classifications to support analysis
- Definition of indicators and standards
- Terminology/thesaurus
- Quality/footnotes

This list, in itself, constitutes a model of sorts of what the most important aspects/contents of a statistical metadata model would be. As such it also defines a methodology for more detailed comparative analysis of statistical information systems. Part of the COSMOS learning process was to define such a model and learn from the use of the model. The results of the comparisons were presented at the COSMOS main conference.

In hindsight most of the domains were relevant but the amount of work needed on each domain was perhaps not distributed in the optimal way. For example, the domain dataset/cubes is probably much more challenging than terminology/thesaurus and would therefore require relatively more resources. Other findings were that project documentation was often produced for internal project use and was therefore difficult to penetrate for those who were responsible for the comparisons. Understanding systems and models turns out to be a time-consuming task in which frequent face-to-face meetings with experts often is necessary to fully understand the thoughts on which concepts, design choices, functionality, etc, are based, even when the documentation is as transparent as possible. A further complication was that several of the projects, at the time, were research-oriented and therefore in the middle of a process of clarifying to themselves the exact meaning and usage of there systems and their different metadata objects.

From a methodological point of view the lessons learned could be summarized as follows.

Decide on a common format for system documentation. The Metaware Neuchatel document as well as the IPIS documentation could serve as the basis for such a format. The Neuchatel conceptual model is an example of an implementation independent model with a strict focus on a clear, precise and coherent natural language definition of user-oriented concepts. Such a UOC-model was used to document the COSMOS model. On of the most useful aspects of the IPIS project documentation were articles directed to statisticians that described the theoretic and statistical underpinnings of the different parts of their system and model. A natural modelling language for the logical model was UML class-diagrams. All projects had UML-class diagrams. The use of full UML use-cases, was not part of the documentation of any of the systems, however, such use-cases were described for the COSMOS exchange scenarios. It is questionable if fully documented UML use-cases – with actors and sequence diagrams - always are needed or even appropriate for documentation of this sort of systems. Often, the statistical underpinnings are more important to document as well as, perhaps, more generalized usage scenarios. However, since such documentation did not exist it is difficult to ascertain if it would have made a difference.

The domains chosen and the order in which the comparisons are made can perhaps be changed somewhat. Preferably the first and most important domains would be Survey and datasets. As a corollary to datasets there should be a variable/indicator domain. Furthermore there should be a value domain/classifications domain. Domain comparisons should first focus on structures and then on processes/quality/footnotes for each domain. This work would then have to be brought together and especially the quality/footnotes area would have to be re-examined in the light of the individual comparisons to create an overview perspective. The questionnaire domain comparison should have as a starting point the results of the dataset and variables comparisons. Other, final, areas could be thesaurus and other search-oriented and cataloguing aspects as well as administrative functionality, such as owners, contact persons and access control. This would be the first step and would create a more detailed understanding of similarities and differences. In order to consolidate these comparisons the different domain comparisons should result in both detailed conceptual mappings and concrete suggestions for common domain models.

3.6 The process of finding a common model

3.6.1 Methodology

One of the first issues to be solved was that of terminology. From experience it was known that it could be very difficult to agree on a common terminology for a metadata model, but more easy to agree on the underlying concepts. Therefore, the decision was taken to use the IPIS model terminology as a starting point. The reason for this was that it was the most

comprehensive, well-documented and scientifically annotated model. It also had a distinct statistical profile. The idea was that the terms are not interesting in themselves, the only serve as labels for concepts and as soon as the meaning of the concepts have been made clear and shared within the project the terms can easily be substituted for more relevant ones. This was also done gradually. As the model evolved some terms were first changed to terms that were more frequent among the cluster projects and then a final overview of the terms was done based on DDI (DDI, 2003) and SDMX (Pellegrino, 2002) as reference terminologies.

The next phase in the discussions was about the relation between the final demonstration and the common model. From the beginning the idea was to create one large federated common model and then implement a subset of it to demonstrate interoperability between cluster projects. However, at the main conference a developing standard for web-based registries was presented. This turned out to be pivotal for the future development of the project since it gave a more precise definition of the COSMOS system. The vision of a web-based registry formed an important basis for the future discussions as well as the development of the common model; it helped "gel" the project. The web-registry paradigm partly meant an expansion of the original ambitions of the project. At the same time it meant that the idea of a federated common model was substituted by a more demo and use scenario orientation for the future work with the common model. This also meant that the model development strategy switched from being concept driven to being use case driven. This also turned out to be practical because no federated model had resulted from the domain comparisons or could be based solely on the results of those comparisons.

At the same time the idea of a core model was introduced. The development of the model would start with the definition of a common core, to which further objects and attributes could be added or to which whole extensions could be added for different purposes.

The next methodological decision was to start with two project metadata models as first input for the core model and then gradually adapt it based on input from experts on the other models. The models that were chosen were IPIS and Metaware.

3.6.2 Differences with an impact on the commonality of core objects

When attempting to define a common core model it was quickly clear that no common core existed for the projects, not even between the three "main" projects, namely Metaware, IPIS and MISSION. The two commonalities sought were Survey and Dataset. Metaware has a statistical activity object, which is similar but not identical to a Survey. MISSION does not have a Survey as such because it is focused on subject matter and content. In IPIS Survey and Dataset are one and the same, which means that one Survey cannot result in several datasets, covering the whole production process. The greatest similarities and the most general model could be found by combining Faster/DDI with Metaware thus having a Survey, which can result in one or many Datasets.

As a complement to the initial system comparisons the more detailed and concrete analysis of the different models yielded a new set of differences with specific impact on the possibility of defining common objects. These could be divided into on the one hand general design principles/ model structure and on the other impact of specific objectives on the design of the individual models.

The general design principles/ model structure were:

- The degree of separation between the conceptual model and the logical model
- The degree of separation between structures and processes in the logical model

- The degree of separation between the logical layer and the physical layer.

The system objectives were:

- IPIS and Metaware are centralised systems for production and publication by a statistical office. Mission, on the other hand, is a distributed system for remote integration of data and metadata from different statistical offices in different member states
- Faster has a strong focus on evolving metadata standards related to the Internet, for archiving, search and retrieval
- IPIS and Mission are fully proprietary systems. Metaware, on the other hand, combines commercial Data Warehouse (DWH) software and related object models with a proprietary metadata repository and a proprietary metadata object model.
- IPIS and Mission are strong in the areas of end-user functionality based on concurrent processing of data and metadata. Metaware, on the other hand, has a strong focus on concepts, structures and technical DWH processing issues.

3.6.3 Design decisions and important design issues

The road taken was to define a Survey object, which could have several datasets. Each dataset could then have one or more variables. Different basic objects were then added to this core based on input from experts and the definition of the needs for the final demo use scenarios. The actual extension models were developed by the groups responsible for the individual use scenarios and were then harmonized with the core model.

One of the more important limitations of the model has been the lack of objects to group other objects over time. For instance, the only way to determine what constitutes a series of consecutive Surveys would be that the time independent part of their names are identical. The same is true for code lists variable versions, etc.

Some further issues of importance to the development of the common model were

- Collection versus dissemination – datasets differ and it can be difficult to create a model, which covers micro, macro and raw data.
- Structure versus process – the dynamic dimension will often have a mayor influence on system design and the metadata model. The specific needs of dynamic processing will tend to affect the basic object structure
- Free text versus formalized objects for quality – systems differ widely in the degree of formalisation of textual information, especially quality but also general footnotes. From just a link from the Survey object to a text document to a full-blown "footnotology" on the level of individual objects and attributes, even individual cells in a dataset
- The registry paradigm made a difference between a classical integral database approach and a batch like catalogue, this influences model constraints in the actual registry, i.e. such constraints can not exist in the RDBMS sense because object instances from different sources can not be subjected to referential integrity. This is because the registry is purely for discovery and retrieval and not for updating.

3.6.4 Generalists, contextualists and incrementalists

When discussing the common model at the main conference it was evident that there were different ways of approaching how the model could be developed. These differences were in turn dependant on, perhaps, slightly different ways of understanding the nature of metadata

and metadata systems. One could speak of “generalists”, “contextualists” and “incrementalists”

The term metadata is often difficult to explain. What is “metadata”? The most common answer given is that it is “data about data”. This, however, does not always give a clear picture of what, more precisely, is meant in a specific context. The term has to be narrowed done, by talking about, for example, “statistical metadata”, “semantic” metadata, “quality” or a specific sub-area such as “classifications”.

A generalist would claim that there is one universally valid statistical metadata model and that the goal of all statistical metadata efforts is to produce this one “reference model”. The model would then serve all possible needs and all metadata dependant processes and systems.

A contextualist would claim that the universe covered by “statistical metadata” is so large and complex that it not practically feasible to design one model that fits all needs. In practice, a choice has to be made, and a model will be developed based on a choice of system objectives in a concrete development project. When new requirements are added the system will have to be more or less reworked from scratch. Perhaps in the long run one may approximate an all-encompassing “reference” model, the contextualist may concede, but to develop such a model would not be a viable strategy in the short term.

An incrementalist would recognize the difficulties in a one-off development effort for a reference model/ complete statistical metadata system but claim that as the experience accumulates there is not a need to completely rework the model / metadata system. Instead, one can start with a core model and then gradually add extensions. The core may be defined in different ways. Either as a core set of objects or as a pareto (the 20% of objects and attributes which are used 80% of the time) set of “most-used” objects, and the increments can be added organically by linking in different ways to existing objects or by some expansion of hierarchies.

The experience of designing a common model for COSMOS seems to support the contextualist point of view. One consequence of this is that a stronger basis for interoperability would be both a common, more generic, way of designing individual systems and, in the long run, the development of and gradual harmonization of international standards with a view to some sort of modular approach to such models where structure is separated from process as far as possible.

3.7 Summary

The lesson learnt in the development of the common model can be summarized as follows. The focus is on creating metadata models and statistical information system (SIS), which have an increased potential for both gradual expansion and interoperability.

1. When building a SIS try to make the construction as general as possible, even if this is not always necessary to achieve the specific project goals.
2. The basis for all SIS should be well-documented statistical theory, which states the theoretical and user-oriented underpinnings of the system.
3. Start the modelling with a strong focus on structures and user-oriented concepts. Create an implementation independent UOC-model (user-oriented concepts) with structural concepts and characteristics.
4. Continue with the logical layer and be careful to again separate structure from process and also the logical model from the physical model, as far as possible.

5. If you decide on extensive and detailed formalisation of process metadata try to make it possible to easily consolidate that information into a free text document on the survey level.
6. Use UML class diagrams
7. Describe the main use scenarios for the system in a more general fashion and maybe also in a more practical hands-on user guide fashion.
8. Take a look at existing or developing international standards and try to harmonize with them.
9. You may want to try a core-extension approach or some sort of modular approach to the system design and modelling.
10. Compare your model with the COSMOS model and other models based on broad comparisons and/or a more comprehensive approach to see if your model covers all relevant domains.
11. When attempting interoperability between different systems, first define the system, which will be used to achieve the interoperability.
12. Begin by defining and implementing a common format to document the different systems.
13. Use a sufficiently detailed domain model of a SIS metadata model and compare the different domains in a certain order, starting with surveys and datasets and ending with questionnaire and an overview of quality and footnotes.
14. Arrange frequent face-to-face meeting between experts.
15. Define a core and then add extensions.

4. The Final Model

The Final model was also influenced by decisions taken at the Main Conference of the COSMOS project (Papageorgiou, 2002a), namely

- The IPIS model (IPIS, 2001) should be taken as the basis, as it was the most rigorously documented
- The common core model should be an intersection, rather than a union of the projects.
- The definition of this model would be driven by use cases required for the final demonstration
- The core model would allow extensions, where projects working on a use case could refine the model to allow for their particular needs.

The final model final had 14 objects, which were described using the common structure given below:

Object name

Concept

Extensive definition of the object

Properties

<u>Property name (underlined if mandatory)</u>

Property description

Property name (not underlined if optional)

Property description

Object relationships

Definition of the relationships the object is a part of

A full description of the model can be found in (Abelin, 2003).

5. The Architecture

The technical architecture and plan for the inter-operability demonstrations were devised by a technical sub-group of the COSMOS project, set up at the main conference. Of the main objectives of the Cosmos cluster, the technical group focused on the following two:

- to implement a demonstration subset of these objects to show interoperability of the developed systems
- to define a methodology for further developing this interoperability.

The group tried to design a system and define a methodology and framework that are independent of the contents of the common model, so that the system will be able to accommodate any changes or enhancements that may occur in the future.

In the next sections we give an overview of the architecture of the system and its functionality, together with some descriptions of the technological choices we have made for implementation purposes. The design was driven by use cases, and examples based on these use cases are given.

5.1 Overview of the system and its functionality

One of the basic aims of the COSMOS cluster design was the effort to keep the system fully distributed and scalable. Although there are currently five participating applications, the overall system should remain open and with a well-defined API, so that any prospect data publisher may effortlessly join in and make their data accessible to the cluster.

A project that wishes to make its data and metadata available to the other members of the cluster via the COSMOS core model and architecture is known as a COSMOS publisher. The main idea is that each COSMOS publisher will make their data available to others by publishing it on a local web server in a predefined format (conforming to the common model). Web technologies and protocols will be used for the exchange of information, and for more efficient search capabilities, an intermediate level, representing an index or a yellow pages service is also used.

This intermediate level consists of a Registry. This will maintain information related to the participants, the kind of information provided, and the way this information may be accessed. Since it only acts as an index, it does not replicate data kept at the publishers' sites unnecessarily, but only maintains pointers to it. In our initial implementation, there will only be a Registry, but federated registries could be considered in the future.

The following procedures relate to the system functionality:

(a) publishing of the information (in form of objects) to the cluster
(b) registration of a participating publisher to the cluster
(c) registration of the published information to the registry
(d) query the Registry
(e) information exchange between participating applications.

Items (b), (c) and (d) relate directly to the Registry, where items (b) and (e) relate to the common API.

Initially, an individual COSMOS publisher transforms the information it wants to make available to the cluster from its own proprietary model to the common one. It then publishes it to the cluster by both making it available at its web site and by registering it to the Registry. For this to happen, knowledge of the contents of the common model is necessary. Although the COSMOS publishers will put out all the relevant information, they only register with the registry their 'root' objects.

The cluster system, together with the Registry and its links, is now set up and ready to service requests. As mentioned before, the Registry works as an index to the published information and therefore its replies consist of:

- which participating publisher can service the request
- what kind of information the publisher can provide the requestor
- how the particular service can be accessed

At the information exchange stage, the Registry is not used at all. The publishers interact directly with each other, through the knowledge of the common model and the framework of the common API.

5.2 System architecture description

Two main issues were considered during the design process. The first was related to the independence of the system architecture and the various technologies chosen, from the contents of the common model. We tried to keep them apart, so that possible extensions or modifications to one would not affect the other. The second issue was related to the use of both emerging and state of the art technologies. Standard and established web technologies were chosen for their common and wide use. Newer, related technologies were a challenge, since it was required to find the way to apply them to the statistics domain.

The architecture design reflects the above considerations. The top layer, the most abstract level, describes the common metadata model. The next layer is the common API and describes both the format for representing objects and the mechanisms of communications that take place among COSMOS publishers. We also refer to this layer as the *Network Layer.* The third layer, called the *application layer*, deals with the publishing applications' metadata model transformations to and from the common model, the realisation of these transformations in an agreed format, and their interaction to the registry.

In addition, we wished to take advantage of the experience of the participating projects, in particular utilising

- RDF and DDI based representations inherited from Faster
- Use of the Registry concept inherited from IQML

Figure 2 gives a basic overview of the architecture.

5.2.1 The Common Metadata Model Layer

This layer describes the common metadata model, as described above and in (Abelin, 2003). It is the conceptual means of communication for any kind of information exchange among the publishers. Each COSMOS publisher must have a complete knowledge of the contents of the common model in order to import/export any information to or from the other applications.

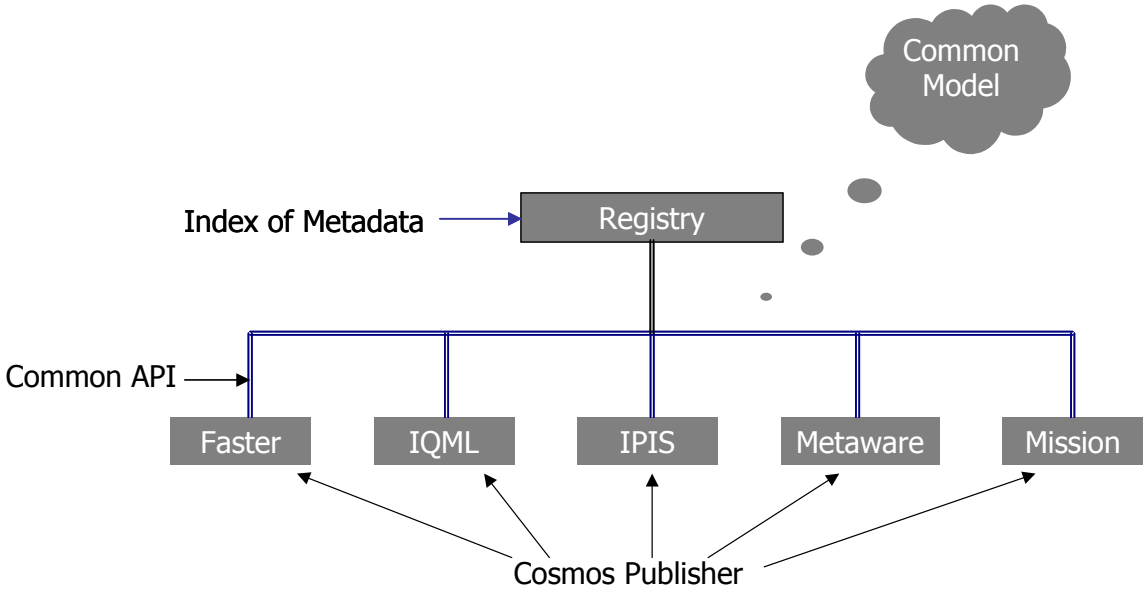


Figure 2:Overview of Architecture

5.2.2 Common API, Network Layer

This is the layer that specifies the common framework for

- the format in which the exchanged information is expressed
- the mechanism of communication between the apps and information exchange

For the format we chose either RDF (a W3C standard) or DDI files. For the means of communication we chose the Internet
This layer consists of the COSMOS publishers, the proposed Registry and the interactions between them.

5.2.3 Application Layer Registry

As it has already been stated, the Registry is an index of the published data, and it is used as an intermediate step in finding information in the COSMOS domain. We propose to use an existing Registry framework, called ***ebXML***, which contains some additional internal structures for more efficient and intelligent searches. EbXML is becoming a standard in the e-business world, and is supported by large organisations and standardisation groups. In addition, IQML, one of the participating projects, is using this technology, which allows us to take advantage of the accumulated experience.

The specifications of ebXML are by the OASIS[1] group. For the pilot project we are using an open-source implementation of the registry hosted at sourceforge: *http://ebxmlrr.sourceforge.net*.

For a COSMOS object to be discovered in the Registry, certain basic classification fields are needed. The fields chosen are:

- Origin (Cosmos Publisher)
- Type of Metadata (classification, dataset, etc.)
- Topic (standard classifications)
- Country (the published object is referring to)

[1] Organization for the Advancement of Structured Information Standards: http://www.oasis-open.org/home/index.php

- Format of published data (RDF, DDI)

5.3 Use Cases

The architecture and demonstrations are driven by three use cases:

- Search Portal
- Exchange of Information between publishers
- Dynamic Exchange of Information and statistical processing

The search portal is the basic use case, and is utilised by the other two. For each Use case we describe the sequence of events needed to achieve the Use case

5.3.1 The Search Portal

For this there are four steps:

1. Creation of the metadata object in a common format: Conceptual mapping of the Cosmos Publisher's metadata model to the common one; build and use a tool that accesses the Publisher's metadata and transforms it to the common model by converting the metadata to a common format (RDF or DDI)
2. Publication of the metadata object on the Web: Put the common format files onto a web server; create a catalogue file, which contains the metadata classification information used to enter the relevant information into the Registry
3. Search for the metadata in the registry via a browser: We have chosen one of NESSTAR's internal applications from the FASTER project to act as an RDF browser.
4. Display the metadata in the browser: the NESSTAR browser displays RDF files.

5.3.2 Exchange of Information between Publishers

This use case is concerned with the exchange of objects between two publishers. A requesting publisher sends a query to the Registry enquiring about particular types of objects (e.g. *datasets, classifications, variables*) that are published in the cluster. The Registry returns a result giving back the URIs of the objects that satisfy the query, together with the way of accessing these objects. The enquiring publisher then uses this information to transfer, import and finally use the actual objects in its own environment via the common API.

More specifically we focused on two sub-cases:

a) Importing a classification from METAWARE to MISSION

b) Importing Questionnaire data from IQML to FASTER.

5.3.3 Exchange of Information to support statistical processes

This is an extension to the previous use case, where dynamic exchange of information is required. Before the actual object transfer takes place, statistical operations on that object are performed. A COSMOS publisher makes an enquiry on some particular published information, but only wants to access parts of the data represented. In other words, it uses the published metadata to perform a query on another cluster publisher. An example illustrating the above could be a *Survey* that is taken over multiple countries and years, but only parts of it satisfy the initial query criteria.

For a realisation of this use case, we need to extend the core common model so that certain classes may support operations (class methods). These method calls will provide the publishers the mechanism of invoking their internal statistical processes.

More specifically we concentrated on the use case when a query is initiated in MISSION and redirected to IPIS.

A full description of the COSMOS architecture can be found in (Allman, 2003).

6. The Demonstrations

The previous section has described the architecture of COSMOS and the use cases it supports. This section will describe the implementation of these use cases as demonstrated at the Final Conference (<ref>) in June 2003. The two demonstrations that were implemented were the registration of four COSMOS projects in the Registry, and the export of metadata from IQML to Faster. These two demonstrations were chosen for a number of reasons. The first demonstrated the general principle publishing and discovering metadata objects via the registry. This step is a precursor to any other use case. The second shows how two projects can interact via the registry to transfer metadata. The IQML and FASTER use case was chosen because IQML is the only true data capture project in the cluster and FASTER supports a standard metadata model which links the question text to a variable.

Significant development work was needed to support the demonstrations, both for the defining of the mappings and the development of the software. Thus the main reason for restricting the use cases was the limited resources available to the cluster. As well as the actual demonstrations we worked on the design of other possible use cases, as will be reported in a later section.

6.4 The search portal

7. Strategic issues

8. Lessons & outcomes

9. Influences: COSMOS on project & vice versa

10. Future work/directions/research

11. Conclusions

References

(Abelin, 2003) Abelin, M and Plancq, L (2003) COSMOS Deliverable 4: Common Core Metadata Conceptual Model and Object Model

(Allman, 2003) Allman, S et al (2003) COSMOS Deliverable 6: Cosmos Architecture and API Specifications

(CASC, 2001) Computational Aspects of Statistical Confidentiality EU Fifth framework project IST-2000-25069 http://neon.vb.cbs.nl/casc/

(DDI, 2003) Data Documentation Initiative http://www.icpsr.umich.edu/DDI/

(IST, 1999) The Fifth Framework Programme, Information Society Technologies Programme, Guide For Proposers, March 1999

(IPIS, 2001) The IPIS Project: Deliverable D5: Evaluation of the present situation and integration analysis, http://www.instore.gr/ipis/deliverables.htm

(Papageorgiou, 2002) Papageorgiou, H. and M. Vardaki, M. (2002) The COSMOS Projects' profile http://www.epros.ed.ac.uk/xxx

(Papageorgiou, 2002a) Papageorgiou, H. et al (2002) COSMOS Deliverable 5: Report of the Main Conference Vouliagmeni, Athens, Greece, 1 – 3 May 2002

(Pellegrino, 2002) Pellegrino, Marco & Ward, Denis, SDMX Vocabulary of Statistical Metadata (draft), update: 28 August 2002

Managemnt annexe

Extract from contract:
The Cluster can be measured against the following assessment criteria which fall into three broad categories: (1) design and implementation; (2) dissemination; and (3) influence on standardisation:
1.1 Has the Cluster successfully identified a set of common Objects that can be used in the repositories of all (or most) of the projects, and built an object model of these common objects?
1.2 Has the Cluster identified a subset of these for which a realistic set of APIs can be developed?
1.3 Has the Cluster been able to demonstrate interoperability between outputs from two or more different projects (eg input via one set of software, and output from another)?

Has the Cluster developed a methodology which identifies how further interoperability can be achieved?
2.1 Has the Cluster communicated its findings to the wider community, via established networks, conferences and journals?
3.1 Has the Cluster made an impact on the development of International standards?
3.2 Has the Cluster made an impact on the practical working of the public bodies involved in the projects?
3.3 Are the modules and APIs that have been developed by the Cluster compatible with standardisation activities elsewhere?
3.4 Have the projects incorporated the ideas derived from the Cluster?
The first set of criteria will be measured by the output Workpackages, and by the final demonstration. The second by identifying the conferences and seminars in which the findings of the Cluster are reported. The third by the activity within standardisation bodies, SDOs and the participating projects. One output from the Cluster will be a report by each participating project identifying the impact of the Cluster on their work, and on their dissemination and exploitation activities.