MISSION (Multi-Agent Integration of Shared Statistical Information Over the [inter]Net) -The data archive perspective

by Joanne Lamb[*]

Abstract

MISSION [1] is a multi-national project funded by the European Commission. It aims to provide a modular system of software that will enable providers of official statistics to publish their data in a unified, and unifying, framework, and to allow consumers of statistics to access these data in an informed manner with minimum effort. The objective of the project is to develop an integrated set of software modules, which will

- Allow suppliers of statistics to subscribe to an integrated network of datastores via an interface to their existing data while retaining control over all aspects of access to their data: their level of involvement; the data they supply; the users who can access it; and the level of resources to commit.
- Allow users to make declarative requests, with a minimum of understanding of statistics, or the domain area, and still retrieve meaningful results from our internal routines or through an interface with external statistical packages.
- Give the user a range of options for automatic harmonisation of statistical data, with clear indication on the interpretation of the results.
- Provide audit trails of data manipulation and analysis, so that methods can be retained, re-used and published.
- Maintain libraries of metadata that can be made available to other users.
- Provide a flexible architecture that allows third parties to act as Independent Metadata Providers, thus encouraging the free exchange of knowledge.
- Allow users to build up individual profiles, accessing data and methods most relevant to their needs.
- Offer a number of independent, interoperable systems that can run on different hardware platforms and access heterogeneous data storage systems.

MISSION is a development of a fourth framework project, ADDSIA [2], but it brings a number of new initiatives to the basic ideas of that project. These are:

- The use of agent technology to optimise queries;
- The use of the Unified Modelling language (UML) in designing the system;
- The development of the concept of Metadata Libraries that are independent of data sources, and which provide a middleman service to the user;
- Tools to enable expert users to develop and share their methodology.

The MISSION Project

MISSION is a European Union R&D project, number- IST-1999-10655. The project started in January 2000 and is due to finish in December 2002. We are therefore halfway through the project. MISSION grew out of the ADDSIA project and has the same partners, who are:

University of Edinburgh, Scotland, UK (co-ordinator)
Office for National Statistics, UK
Central Statistics Office, Ireland
Tilastokeskus (Statistics Finland), Finland
University of Athens, Greece
University of Ulster, Northern Ireland, UK
Desan Marktonderzoek BV, The Netherlands

The objectives of the project can be summarised as follows: We aim to build a software suite that will allow statistical data providers to integrate their publication of data on the Web. This software will have a number of features, based on the requirements of data suppliers and data users.

For suppliers, the system will allow them to subscribe to an integrated network of datastores via an interface to their existing data. They will be able to retain control over all aspects of access to their existing data.

Data Users will be able to make requests in a declarative manner, with a minimum of understanding of statistics, or the domain area, and still retrieve meaningful results. The users will be able to tailor their environment, from simple requests to detailed in-depth analysis. They will also be able to build up individual profiles, accessing the data and methods most relevant to their needs.

A key objective is to allow methods of data manipulation and analysis to be retained, re-used and published. This will be done using Libraries of metadata and of tables, both of which can be developed in the MISSION system and then published for re-use by other users. We have in fact separated the functions of *data* providers and *metadata* providers. This approach gives Mission a very flexible architecture, which will allow third parties to act as Independent Metadata Providers.

The MISSION system will be implemented on independent, interoperable systems running on different hardware platforms and accessing heterogeneous data storage systems.

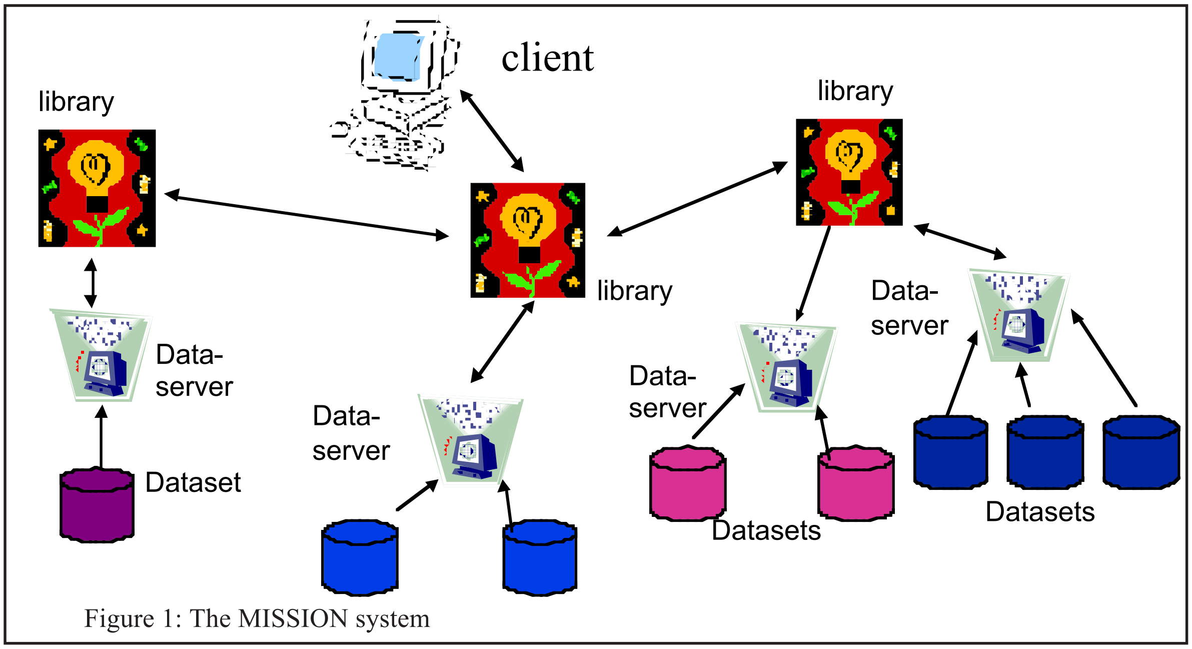


Figure 1: The MISSION system

The core of the MISSION approach is the notion of allowing a user to form a query over several datasets, using a centralised statistical engine, which would parse the request and send partial queries to different datasets held in the system. This was the key concept of ADDSIA, which has been further developed in MISSION.

To achieve this we identify three basic concepts: the *Client*, the *Library* and the *Data Server*. The Client is a piece of software that can be downloaded from a MISSION site and installed on the user's machine. It connects to a host MISSION site, and offers both the user interface and the user's workspace.

The Library is the core of the MISSION system. It is a repository for different types of statistical metadata: access, methodological and contextual. It also contains the statistical processing engine. Libraries communicate with each other via agents, and therefore once a user is connected to a host library, he or she potentially has access to all MISSION Libraries.

The Data server is the unit that gives access to the data. It provides the link from the data to the library. Data servers register with libraries, and then register their datasets. The access rights components of the data server enables the data suppliers to specify who can access which parts of their data.

Only aggregated data is sent to the Library. When a request is sent to the Data Server, the result of that query is computed, and this aggregated result is sent to the library. In this way, sensitive micro data is not sent outside the Data supplier's site, and also the amount of data transferred over the Internet is reduced, thus making the retrieval efficient.

Figure 1 gives the overall picture of the MISSION system.

Innovation in MISSION

MISSION has extended this basic idea in a variety of ways. First, we are using agent technology to enhance the system in a number of places. In the formulation of a query, the user is presented with logical variable names and descriptions, and agents will search a number of libraries to discover datasets containing these logical variables. The agents will then process the metadata for these datasets and can use a number of techniques to optimise the user's request. These agents first construct the correct query, and then plan the execution of that query. Further agents connect with the data

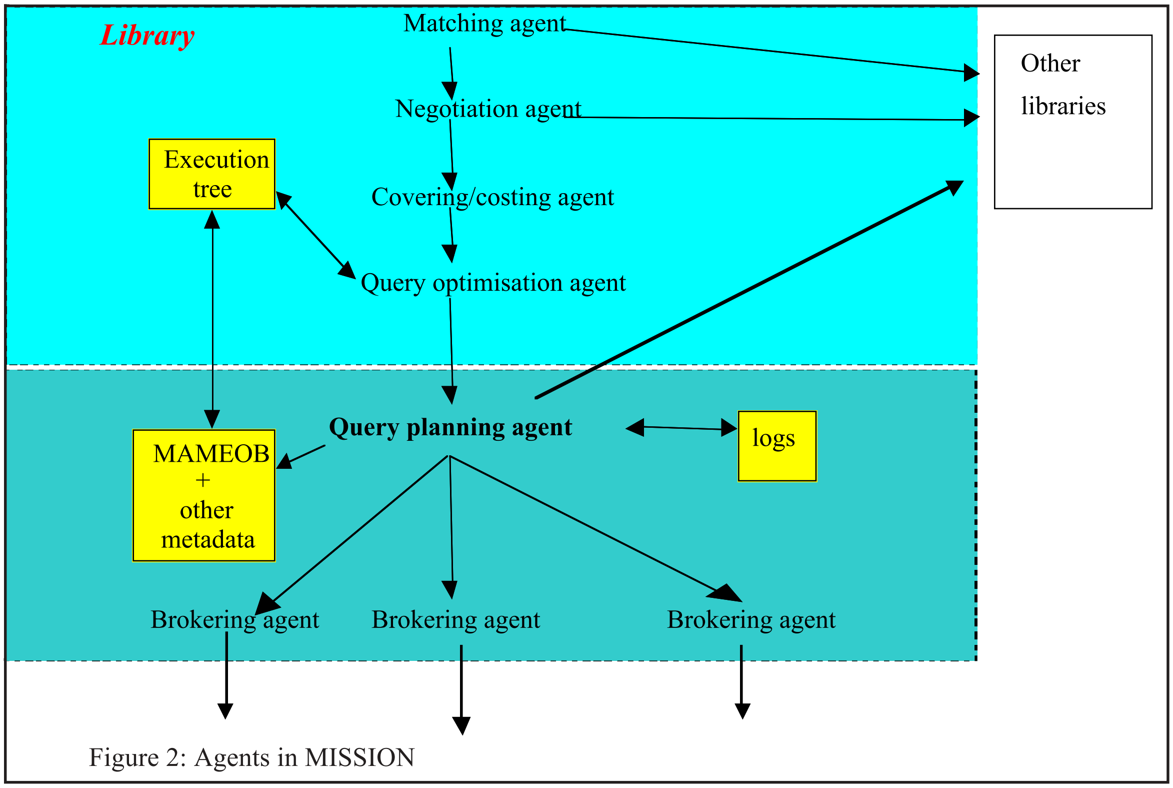
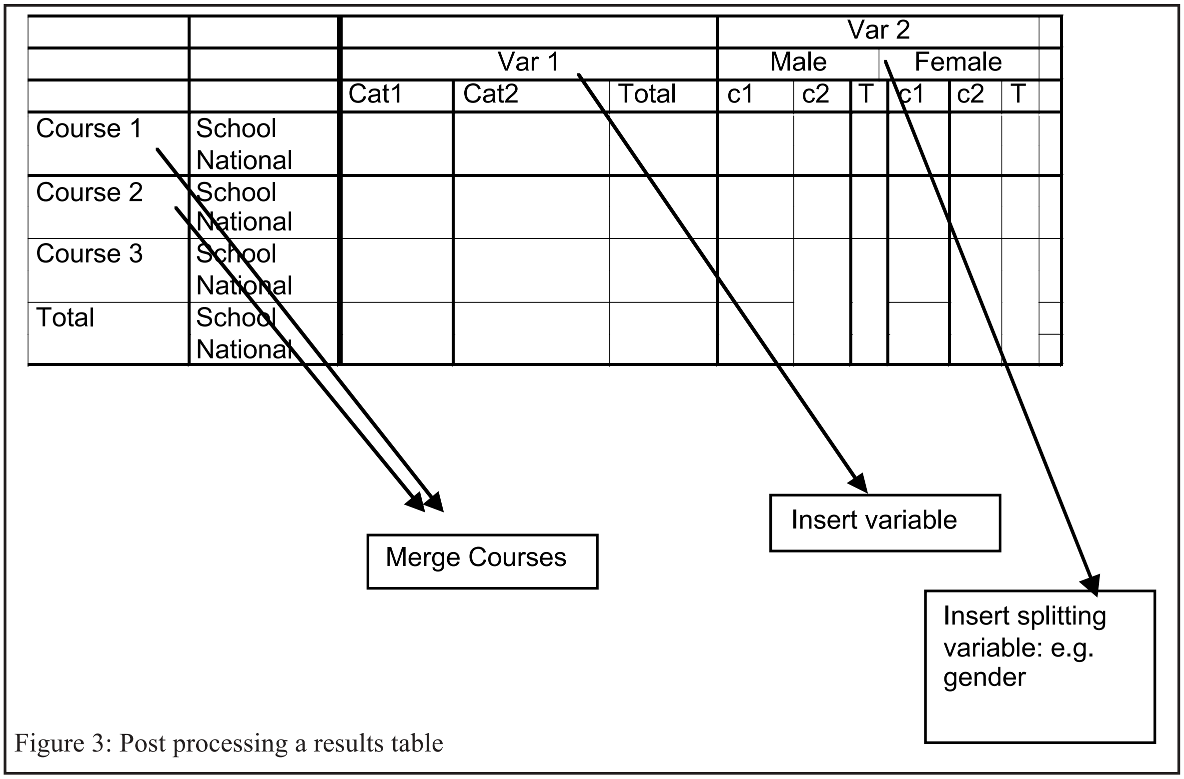


Figure 2: Agents in MISSION

servers. This use of agents is illustrated in Figure 2, where the first part of the diagram shows the query planning stage, and the second part shows the query execution phase.



		Var 1			Var 2					
					Male			Female		
		Cat1	Cat2	Total	c1	c2	T	c1	c2	T
Course 1	School National									
Course 2	School National									
Course 3	School National									
Total	School National									

Figure 3: Post processing a results table

Second, we have designed a graphical user interface, which will allow the user to specify and format his query using a table template to build up the query. This graphical specification is translated into the internal query language, and the results of the query populated the table frame that the user has built up. There will also be opportunities for the user to ‘post process’ this query, by graphically specifying the modifications, as shown in Figure 3.

A third aspect worth mentioning is the development of the MISSION model of metadata. Figure 4 shows the MIMAMED data model. MIMAMED stands for MicroMacroMetadata model, and is a model designed to

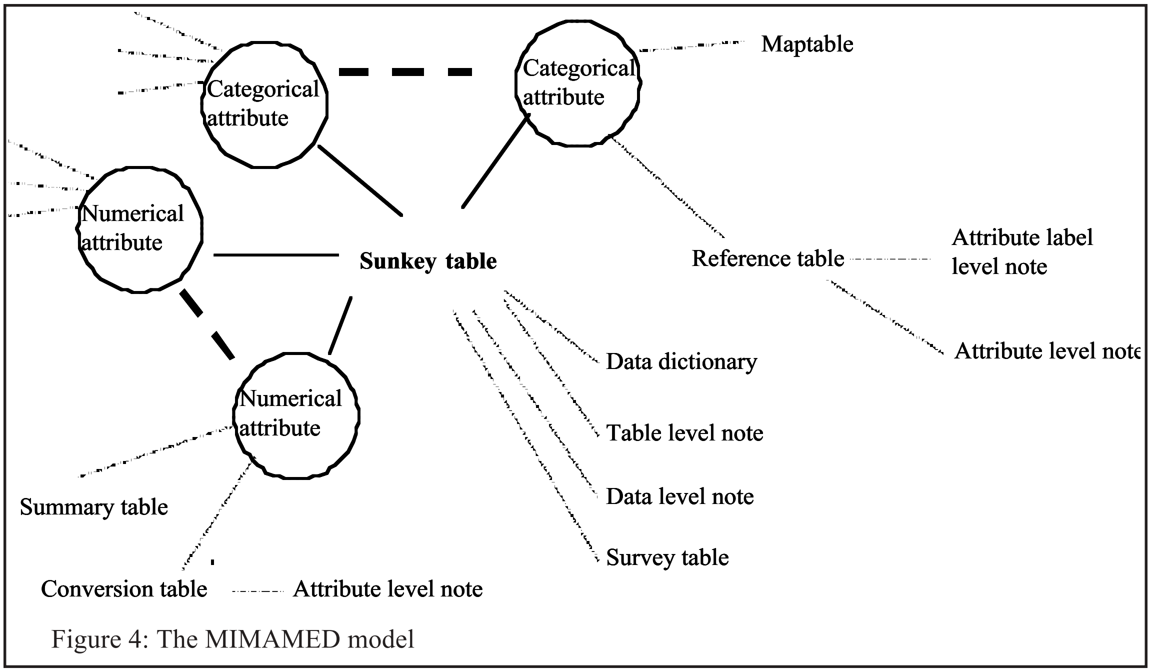


Figure 4: The MIMAMED model

describe all three types of data, and to process the data and metadata simultaneously. In this model, developed by the University of Ulster [4], the ‘sunkey table’ is a reference table for a set of particular aggregated data – the result of a query. The Summary table is the actual data, all other tables contain metadata, which, when aggregates are combined, will also be processed. In addition, we have demonstrated that this model maps to other common models, such as data warehouse structures, cubes and the CRISTAL [5] model.

Finally, the MISSION project is considering how the system will be made available at the end of the project. We have made an in principle commitment to Open Source [6] publishing of the source code, subject to this being a legal option in the setting of a European Union R&D project.

Process to date

The aspects of the system described above will feature in the first prototype, due in September 2001. In February, we successfully tested the connectivity of the system, with a Client in Ulster linking to a Library in Edinburgh, which queried three Data Servers in Athens (running on different operating systems and relational databases). We have also completed the Data Server installation package. In May 2001, we repeated the experiment with a more complete query processing, and are about to complete the Library installation package and supply the demonstration sites with these test versions.

With the first prototype in September, the user will be able to:

- Browse metadata i.e. view dimensions of an aggregated dataset;
- Produce ‘simple’ tables from one source;
- Produce ‘composite’ tables from different logically homogeneous sources.

Future Plans

In the second phase of the project we plan to provide more functionality that will enable the expert user to share methods and publish results via the system. When looking at the requirements of the system for these features, we need to consider three different aspects:

- The model for handling and displaying ‘transformations’;
- Storing MISSION tables for publishing;
- More contextual metadata.

For the first point, we will build a metadata model developed round the transformations that we have identified in Figure 5. However, practically this will be stored as XML, for two reasons:

- First, it is a principle that all metadata held in the Client workspace will not require any software except a Java environment – so the data cannot be held in a database.
- Second, we are developing a metadata user interface, which will allow the user to search any XML file.

This metadata browsing tool will give us the ability to exploit XML files from different sources. We are particularly interested in the development of the DDI for aggregated data, and also in the standards table DTD.

For importing data files, PC-AXIS, and SPSS will be the first formats that we support. We are considering which other formats are likely to be in demand, and expect to see a demand for DDI in future.

Method	Comment
Computation	uses an arithmetic expression to construct the result
Truncation	removes one or more digits from the end of a code - usually for a classification
Rule-based	a series of instructions of the form IF xxx THEN result = y
Banding	a specification that a range of values should be collapsed to a single code
Transcoding table	a table of initial and final results
Complex	a combination of methods

Figure 5: Classification of methods

The concentration for prototype 1 has been on the downloaded Client software. Users register to the Library, and have privileges to access certain datasets according to the data suppliers’ stipulation. Thus control of the use of data is left with the supplier. The degree to which data is confidential is also at the suppliers’ discretion, and we will advise caution in this area. While a single request may be easy to monitor, tracking a series of requests that may lead to disclosure is more difficult.

In contrast, the public user – i.e. a user who accesses the system via the web, without registration – has no direct access to data. Access to metadata is freely available, and also to pre-defined tables which are created through MISSION. Therefore the amount of information accessible to the public depends on the number of tables published in a library.

The second prototype is scheduled for March 2002, and the final version for December 2002.

Implications for Data Archives

We have depicted the MISSION Library as a separate entity from the Data Server, and we picture the two different modules being run by different organisations that have different expertise. The inspiration for the Independent Metadata Provider scenario came from the way in which social science currently uses quantitative data. Typically a researcher will get data from an archive, which will be well documented for secondary analysis. However, after a two or three year project, the findings are published in theoretical papers, and the modified data is destroyed for legal or economic reasons. The researcher may not legally be permitted to keep the modified data, since this would be for a purpose other than that of the original project. Alternatively, there may be a financial cost for using the data for another reason. If the researcher cannot keep the modified data for his own purposes, it is even more difficult for another researcher to pick up on this work and continue.

It has been observed that the manipulation of data prior to analysis encompasses the hypotheses of the research [7]. It is therefore important to capture not only the algorithm of the transformation, but also the reasoning behind it. If this can be presented to the analyst as a tool for aiding his own work, then the overhead of supplying this metadata is not seen as arduous. Once this reasoning has been captured, it is available to give a reasoned method for other to use.

The wider context

This section places MISSION in the wider context of statistical information systems research in CES and in relation to other EU initiatives in which we are involved.

The CES is participating in two R&D projects and three networking projects. While MISSION is concerned with data dissemination, the other project, IQML [8] is concerned with data collection.

The three networking projects will be described briefly. MetaNet - A network of excellence for harmonising and synthesising the development of statistical metadata - started in November 2000, and held its first conference in April 2001. The proceedings from the conference are due at the end of May, and will be available from the website [9].

AMRADS - Accompanying Measure to R&D in Statistics – is a project concerning issues of technology transfer from R&D projects to national statistical institutes, and between national statistical institutes. While the focus of this project is on official statistics, the issues it addresses – the transfer of new ideas and technology between research and services institutes – also has relevance for data archives. The project has six themes led by experts in the area, and CES is responsible for metadata. The first significant event in this project will be the ETK/NTTS2001 conference in Crete in June 2001, where AMRADS has had significant input to the programme.

COSMOS, a Cluster Of Systems of Metadata for Official Statistics, will start in September 2001. Clusters are specifically for EU fifth framework projects, to encourage the sharing of knowledge during the lifetime of the projects. In COSMOS, lead by CES, we have five projects, whose acronyms are:

MISSION	(see above)
IQML	A Software Suite and Extended Mark-Up Language [XML] Standard for Intelligent Questionnaires
FASTER	Flexible Access to Statistics, Tables and Electronic Resources
Metaware	Statistical Metadata Support for Data Warehouses
IPIS	Integration of Public Information Systems and Statistical

Details of the projects can be obtained from the Information Society Technology website [10]. Figure 6 shows the relationship between the projects. The two R&D projects are at the bottom of a hierarchy of generalisation. These feed into COSMOS, where the objective is to demonstrate interoperability between some components of the five projects. They, and COSMOS, feed into MetaNet, which is looking to build a conceptual framework of statistical metadata. Finally, MetaNet and metadata form one strand in the general investigation of issues concerning technology transfer from R&D in official statistics.

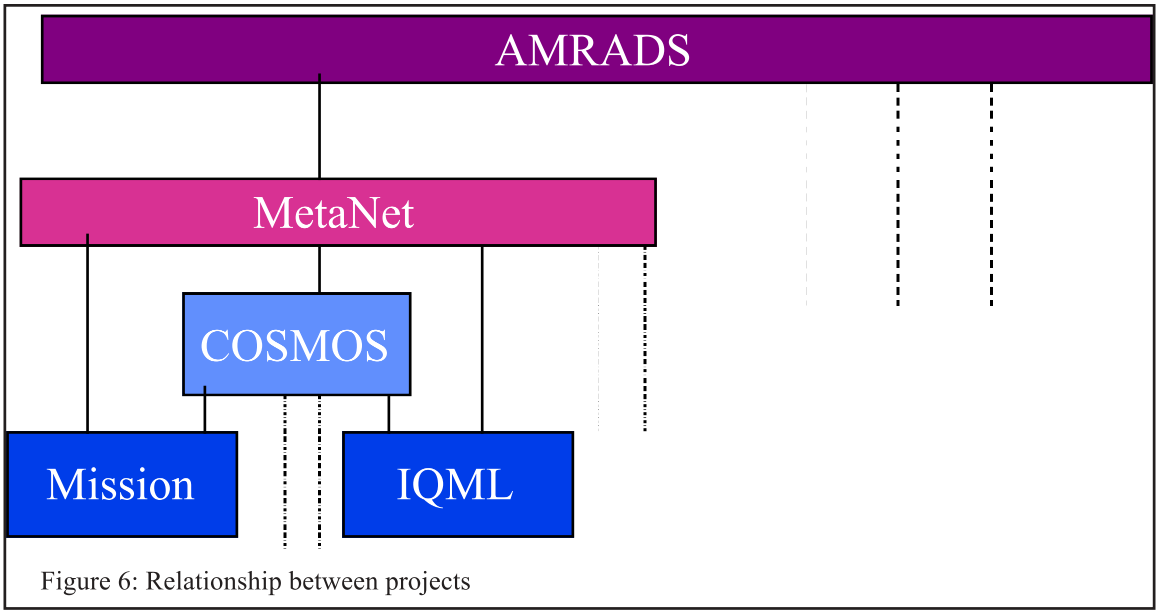


Figure 6: Relationship between projects

Conclusions

In conclusion, we would like to emphasise the following points. First, MISSION is an open system, which aims to help users exchange methodologies as well as utilise data from different sources. This gives the opportunity for data archives and data libraries to host metadata sites as well as access to data. We deliberately called these site Libraries, since we feel that much of the metadata knowledge is held at the servicing level, rather than at the data provision level.

The expertise of librarians and archivists is complementary to that of producers of (official) statistics. We expect to input the metadata from providers automatically, and are keen to utilise as many standards of metadata that it is feasible to handle.

Finally, MISSION is also contributing to other activities aimed at getting a shared understanding of statistical metadata needed for the processing, documentation and preservation of statistical data using modern technology.

References

1. http://www.epros.ed.ac.uk/mission

2. http://www.ed.ac.uk/~addsia

3. http://www.epros.ed.ac.uk

4. http://www.epros.ed.ac/metanet/conferences/proceedings.html

5.Van Bracht, E., de Jonge, E. & Kaper, E. CRISTAL Data Objects. An Object Model for Cubic, Raw, or Intermediate Statistical Data. Statistics Netherlands (March, 2000).

6.Pardue, H (2000) Open Source Software Development: A Business Model. Paper presented at the 31st Annual Meeting of the Decision Sciences Institute, Orlando Florida November 18-21, 2000.

7 Fenelon J-P, Grelet Y, Houzel Y (1997) Analysing transitions in the Labour market through individual longitudinal data: Some methodological issues. Paper presented to the fourth Transitions in Youth workshop, Dublin 1997.

8. http://www.epros.ed.ac.uk/iqml

9. http://www.epros.ed.ac.uk/metanet

10. http://www.cordis.lu/ist/projects

* Paper presented at the IASSIST/IFDO Conference 2001, Amsterdam. Joanne Lamb, Centre for Educational Sociology, University of Edinburgh, Scotland, UK