STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working Paper No. 16
English only

CONFERENCE OF EUROPEAN STATISTICIANS

Joint UNECE/EUROSTAT Work Session on Electronic Data Reporting
(Geneva, Switzerland, 13-15 February 2002)

Topic (iii): Metadata, conceptual models and standards

THE IQML MODEL OF METADATA FOR DATA CAPTURE

Submitted by University of Edinburgh, United Kingdom[1]

Invited paper

ABSTRACT

IQML (a software suite and XML standard for intelligent questionnaires) is a shared-cost project in the European Union Fifth Framework Programme (IST-1999-29093). The partners consist of three commercial companies (Dimension EDI, Comfact AB and DESAN Marktonderzoek), two National Statistics Offices (CSO, Ireland and Statistics Norway) and two universities: the National Technical University of Athens, and the University of Edinburgh, which co-ordinates the project. The project started in February 2000, and ends in January 2003.

There are five related software modules in IQML which share a common data model. In addition, certain modules utilise their own model, which maps to the common model, but which also contains features that are necessary for their particular functionality. The paper will present the common model, the related models, and the mappings between them. It will also describe the activities of the IQML group in disseminating the models to relevant standards groups, in the software development community and in the area of official statistics.

The following software modules contribute to the IQML project: the Questionnaire Presentation Tool and the Database Interrogation Tool developed by Comfact AB; the Questionnaire Design Tool developed by the University of Edinburgh, the Survey Administration Tool developed by DESAN, and the Metadata Repository developed by Dimension EDI. The Repository has two roles: it holds the central IQML metadata model, but it is also generic enough to permit the other modules to hold their own objects, although in this case the repository offers less support for the stored object. The Questionnaire Presentation Tool and the Database Interrogation Tool share a common model oriented towards data capture. The Questionnaire Design Tool has a model oriented towards the definitions of questionnaires and similar survey instruments. The Survey Administration Tool has a model for managing the administration and monitoring of questionnaires in different media (paper, e-mail, web etc). The models are designed in UML and implemented in XML. The role of the repository is to receive metadata that is shared between the modules, and also to hold metadata that is needed for other tasks in the analytical data management process, such as editing, analysis and dissemination.

While the modules provide a complete suite for capturing statistical data via questionnaires and forms, they also function as independent packages, and the Repository gives an open interface that will allow other software to interact with the IQML modules.

---

[1] Prepared by Joanne Lamb.

I. INTRODUCTION

1. IQML (a software suite and XML standard for intelligent questionnaires) is a shared-cost project in the European Union Fifth Framework Programme(IST-1999-29093). The partners consist of three commercial companies (Dimension EDI, Comfact AB and DESAN Marktonderzoek), two National Statistics Offices (CSO, Ireland and Statistics Norway) and two universities: the National Technical University of Athens, and the University of Edinburgh, which co-ordinates the project. The project started in February 2000, and ends in January 2003.

2. The software modules that contribute to the IQML project are as follows:
   - ?? The Questionnaire Presentation Tool developed by Comfact AB
   - ?? The Database Interrogation Tool developed by Comfact AB;
   - ?? The Questionnaire Design Tool developed by the University of Edinburgh;
   - ?? The Survey Administration Tool developed by DESAN;
   - ?? The Metadata Repository developed by Dimension EDI.

3. The Repository has two roles: it holds the central IQML metadata model, but it is also generic enough to permit the other modules to hold their own objects, although in this case the repository offers less support for the stored object. The Questionnaire Presentation Tool and the Database Interrogation Tool share a common model oriented towards data capture. The Questionnaire Design Tool has a model oriented towards the definitions of questionnaires and similar survey instruments. The Survey Administration Tool has a model for managing the administration and monitoring of questionnaires in different media (paper, e-mail, web etc). The models are designed in UML [1] and implemented in XML. The role of the Repository is to receive metadata that is shared between the modules, and also to hold metadata that is needed for other tasks in the analytical data management process, such as editing, analysis and dissemination.

4. While the modules provide a complete suite for capturing statistical data via questionnaires and forms, they also function as independent packages, and the Repository gives an open interface that will allow other software to interact with the IQML modules.

II. THE IQML MODEL

5. The aim of the IQML model is to capture the main aspects of the process of designing and administering a survey. Within the life cycle of the statistical process, it covers questionnaire design, survey administration, data capture and the storage of the related metadata in an accessible format. It does not cover the production of indices, analysis or the dissemination of statistical results.

6. The model is represented by a graphical language called UML, which can be realised using one of a number of software packages to aid the production and documentation of this language. The IQML project uses Rational Rose [2] for this task.

7. In this paper, for ease of reproduction, we will describe the model in a textual way. Since the Metadata Repository is the main vehicle for transferring metadata from the individual models to external system, we will begin here. Terms that are used in the IQML model will be shown in italics when first introduced.

8. The Metadata Repository (MR) has four major packages: Foundation, Classification, Question Bank, and Administration Metadata. The Foundation package has two main orientations: content and expressions. By Content, we mean the information and prompts that are contained in a questionnaire. Normally this can be regarded as text, but graphics, and, in an online questionnaire, animation and sound

also qualify as content. Within a piece of 'text' in a questionnaire, there may be a mixture of words and graphics, so an ordered set of Content can be held as a Content Group.

9. Structurally a questionnaire requires content at a number of different places, and these places are referred to generically as a Node. Types of nodes include questions, sections, variables, documents, etc. The Content Group is related to a node via a Context. The context identifies the circumstances in which a node may be placed. Examples of context include language, population definition, and delivery media.

10. The objects in the Foundation package are necessarily somewhat abstract. Figure 1 shows a diagram of these relationships.

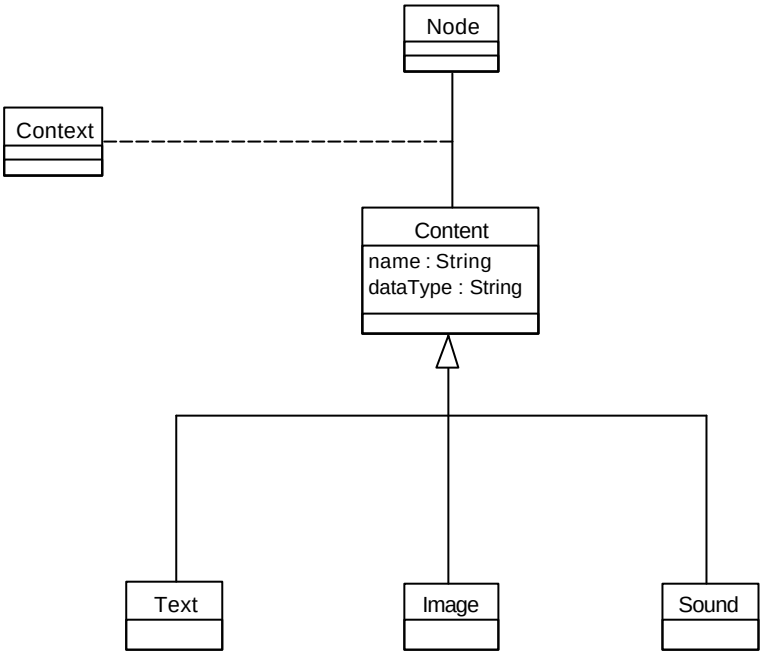

Figure 1

11. The Classification package is also a general package that supports all hierarchical structures. For this reason, the model is a recursive representation of the nodes and leafs of a hierarchical tree. Classifications are discussed in more detail in section VII.

12. The Question Bank package contains Questions and Sections that may be used in one or more questionnaires. A section is an ordered set of questions. Thus a question may be part of a section, or may stand on its own. Questions and sections are types of node. In the design of a questionnaire, it is often the case that a Concept is defined before the detailed question is developed. The question bank can hold hierarchical sets of concepts. Concepts are linked to questions, and the leaves of a tree of concepts can be related to a single Variable. A variable is also a subclass of node. These are the main ideas in the question bank, and are illustrated in Figure 2.

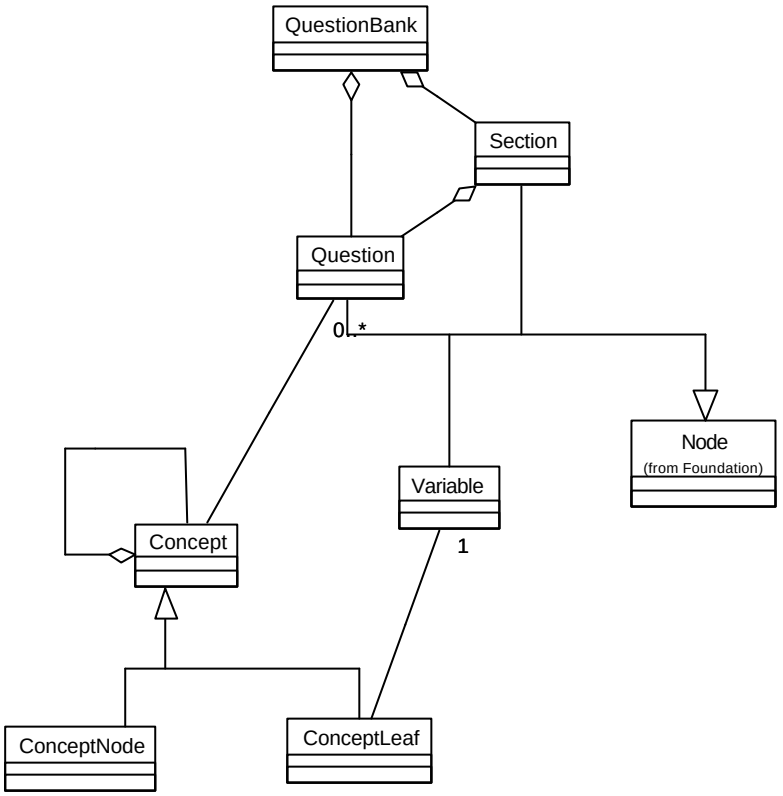

Figure 2

13. A questionnaire is constructed by referencing questions and sections in the Question Bank. A questionnaire can have a context, which determines which alternative of content should be associated with each of the questions and sections.

14. The model also represents control rules for navigation, consistency and calculation. A Process Rule is associated with a node. The Process Rule is an algorithm that yields the value TRUE or FALSE. One or more Actions are associated with a Process Rule. There are two types of action: the IfTrueAction and the IfFalseAction. The appropriate content is linked to the node, according to the specification of the action.

15. Survey Administration is concerned with details of individuals, and therefore most processing is done at the data level and not the metadata level. Nevertheless certain types of key information need to be recorded for the further processing of the statistical data. First, we identify a subset of Variables called AdminVariables. These are used in the decision-making related to survey processing. We then identify three subclasses of this variable: The AdminNumber uniquely identifies a sample member. An AdminComponent is a variable that contributes to the construction of this AdminNumber. It should be stressed that the metadata does not supply any unique identifier; it merely identifies the variable in which this is held, and specifies the algorithm to obtain this number. A residual class holds other variables that are used in the administrative process. Typically these are details of subsampling, the medium used for data capture, and other context oriented variables. Process variables such as date of despatch and reminder procedures are also included in this category, which is known as an OutputRecordComponent. The OutputRecord is the collection of all OutputRecordComponents. The OutputRecordSpecification defines an OutputRecordComponent. EventTriggers define some of the OutputRecordSpecifications, and these events can be both InputEvents and OutputEvents. These concepts are illustrated in Figure 3.

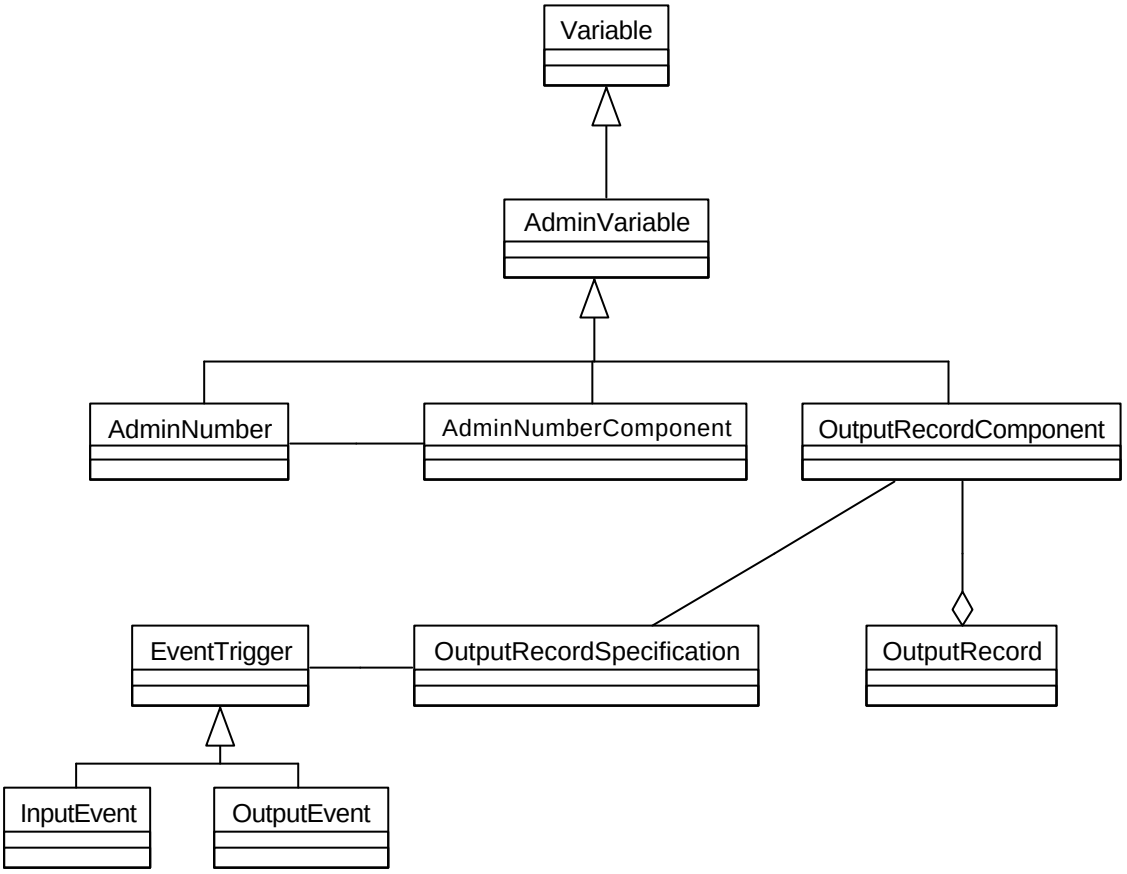

Figure 3

III. THE PRESENTATION MODEL

16. The Questionnaire Presentation Tool (QPT) is used to present XML questionnaires in web browsers on PCs for manual data entry. It supports validation, navigation and calculation. It uses available web-browser technologies such as DOM, DHTML, CSS and Java Scripts. Security and confidentiality mechanisms are integrated using off-the-shelf products. Figure 4 shows the components of the Presentation Tool.

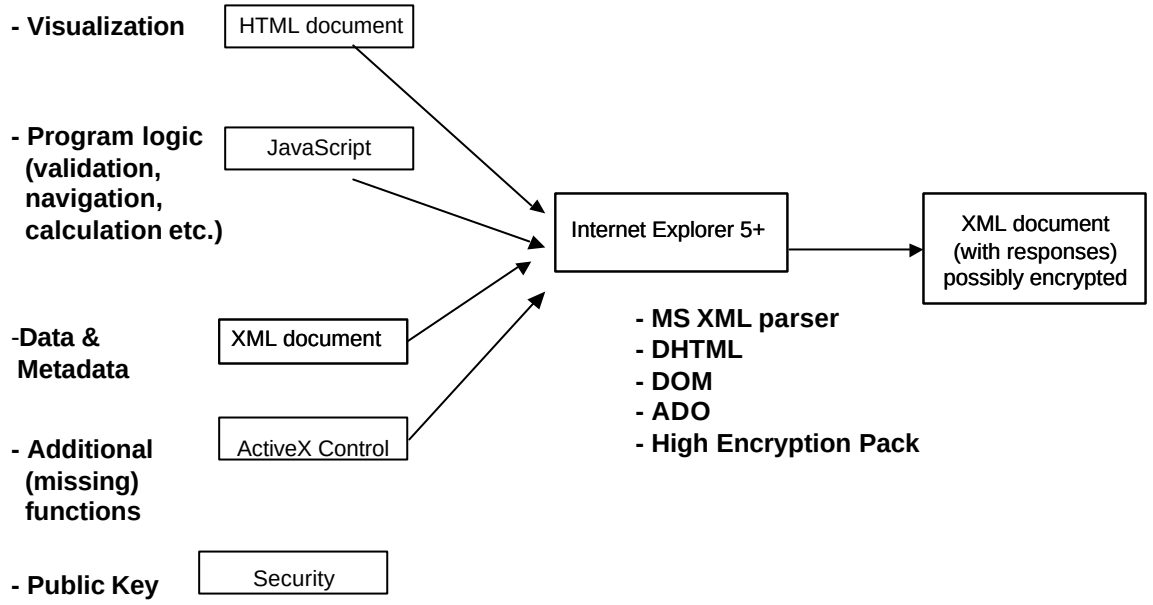

Figure 4

17. Visualisation is accomplished using HTML. The program logic, that is the control of navigation, validation and calculation is achieved using JavaScript. The data and metadata are contained in an XML document. ActiveX is used to supply the necessary extra functionality. Existing software is integrated to supply the security functionality.

18. The Database Interrogation Tool (DIT) shares the same data model as the QPT. Database Interrogation supplies support for semi-automated extraction of data from one or more ODBC compliant database. A separate XML file contains the mapping of items in a questionnaire to one or more databases. The retrieved data is saved in XML questionnaires of the same format as the QPT. It requires little or no additional software, beside a web-browser (IE5+) and makes use of technologies such as Java Scripts, ODBC, DOM, etc. It does require knowledge of the structure of the local database, and of SQL. Pre-filled information can be included.

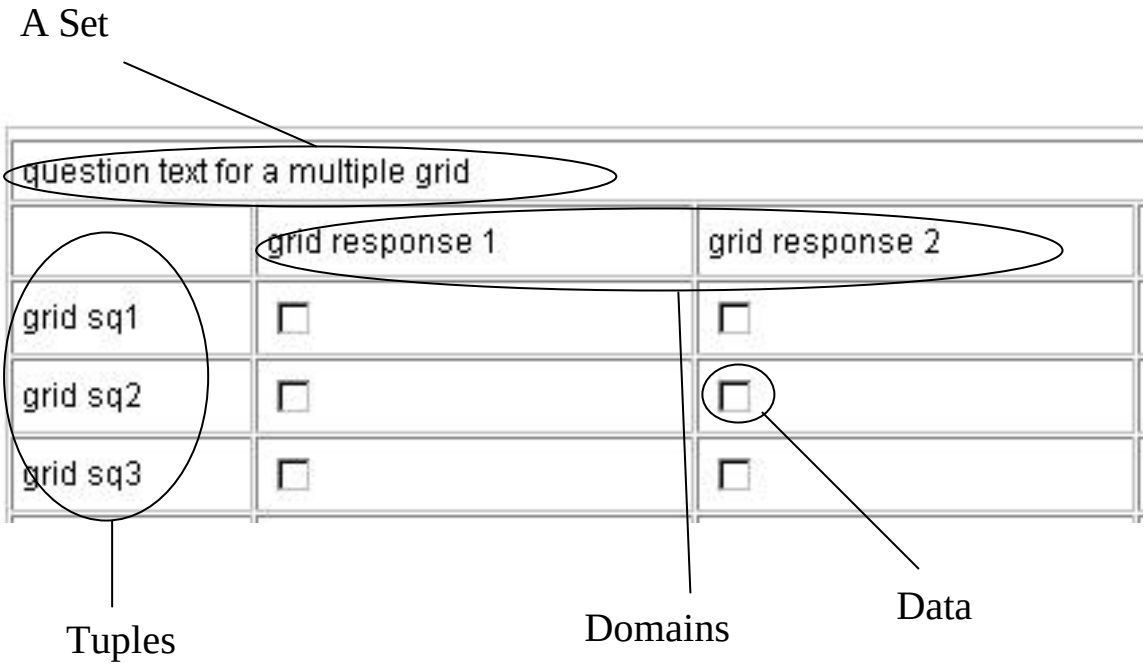

Figure 5

19. The QPT and DIT use the IQML XML definition, 1.0. The core consists of four objects: Set, Domain, Tuple and Data. These allow for flexible data identification, and are illustrated in Figure 5. In addition, metadata, such as code lists, texts (help, instructions etc.), validations, calculations, navigation, numbering are supported. Additional features include the following:

- ?? Multilingual support;
- ?? Response/Confirmation;
- ?? Requests;
- ?? Multi-dimensional time-series;
- ?? links to external sources e.g. regulations and reporting instructions;
- ?? different renderings through different style sheets;
- ?? a questionnaire can be responded to more than once.

20. Figure 6 shows the main classes in the model supporting IQML XML version 1.0. An Instance is an entire questionnaire, completed by one respondent. An XML file may hold several instances. A Set is a set of questions and answers, which can be visualised in a tabular form. A Set has Domains and Tuples, and together they identify the Data. Formats, Text, and Calculation are linked to the Data. Text can also be linked to the other objects: Domain, Set, and Instance.

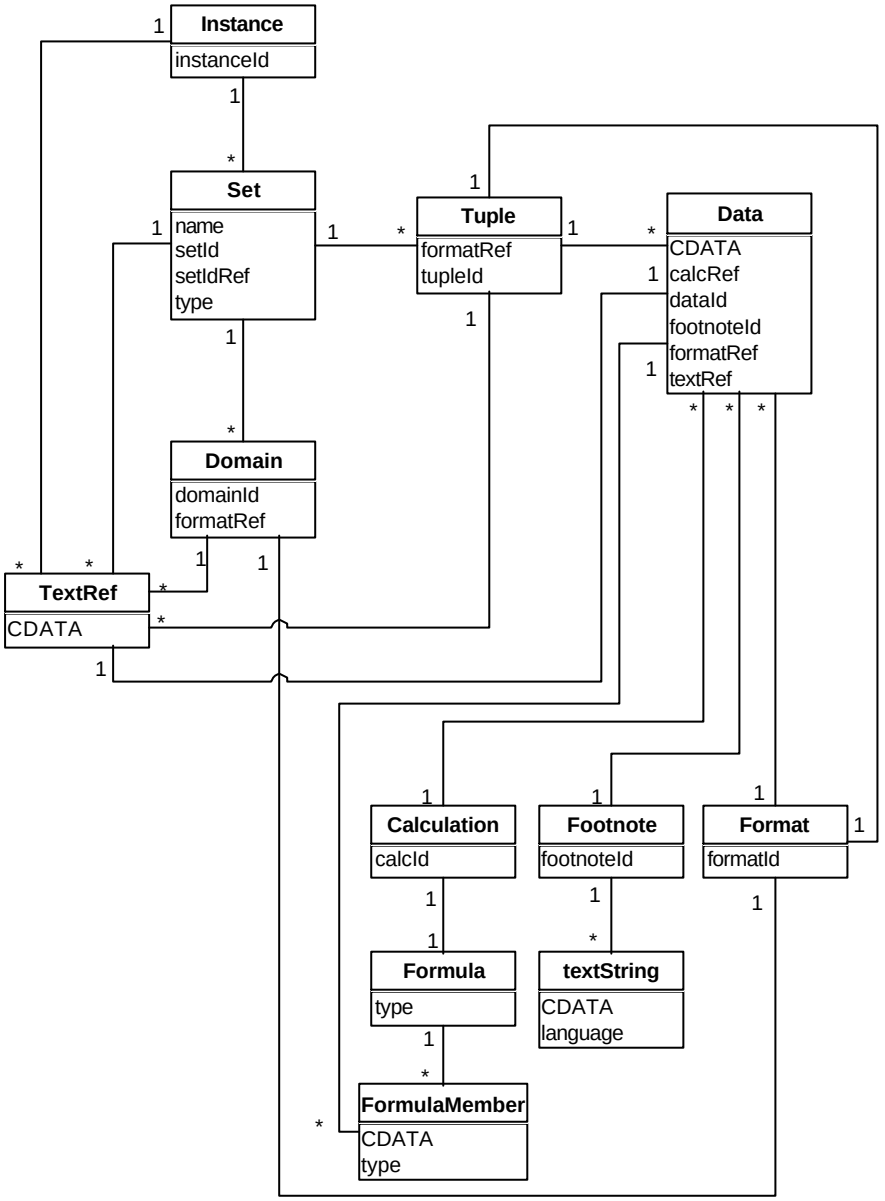

Figure 6

IV. THE QUESTIONNAIRE DESIGN MODEL

21. The objective of the questionnaire design model is to allow the designer to construct a questionnaire that will capture the required responses. Depending on the type of respondent, this may be a simple or very skilled task. We can identify three stages in the task of designing a questionnaire. First, the identification knowledge that is required of the survey, and specification of indicators. Second the defining of the questions, and the crafting of the wording. Third, the structuring of the questions into a logical flow for the respondent. These stages are repeated and refined. Different designers have different approaches to the task, so each of the three stages can be elaborated during the design process. The Questionnaire Design model has a Question Bank which holds Questions and groups of Questions (Sections) as an unordered collection. A Questionnaire is then an ordered list of references to questions and sections in the bank.

22. Associated with the bank is a QDTmeta file, which defines the metadata for the Question Bank. Two types of metadata are used to define how the questions should appear, and what type of data they should collect. The Question Type classifies a question according to three criteria: data type, response type and the use of sub-questions. These concepts are illustrated in Figure 7.

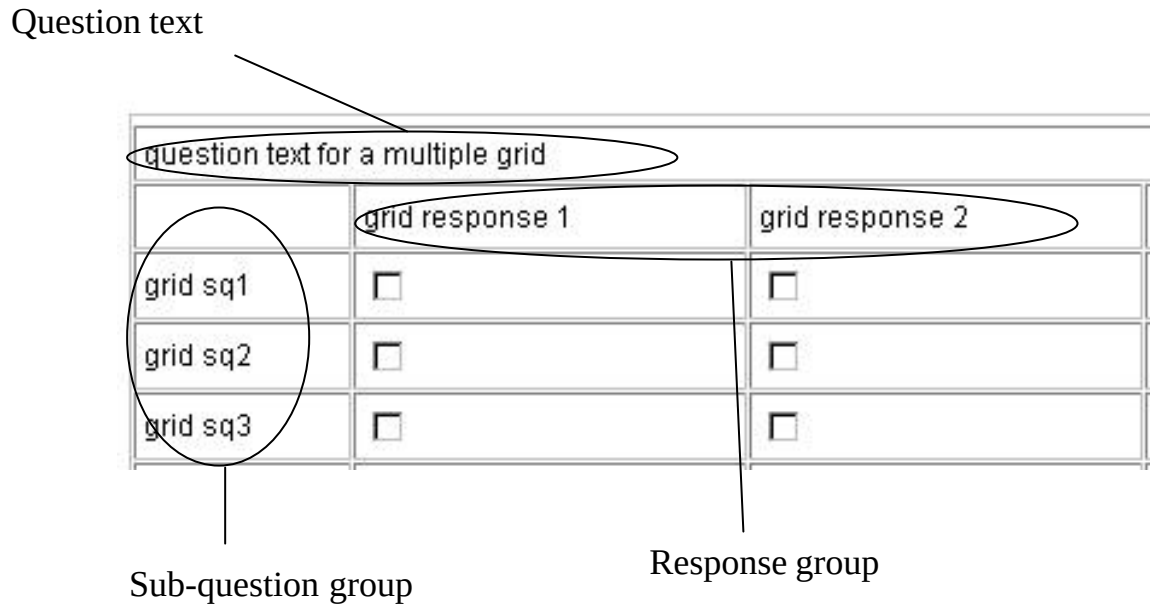

Figure 7

23. Questions are classified, depending on whether the data type is integer, text, Boolean (yes/no) etc, whether the response type is simple, single choice, multiple choice, and whether the question has "sub-questions". If so, the question is a table or grid. Question types are user defined, and so can use in house terminology. It can be seen that this mirrors the QPT scheme very closely.

24. We associate six types of Text Element with the Questionnaire: Questionnaire, Section, Question, Response, Sub-question and Supplementary. These elements can be sub-classed by the user, so that a consistent style can be applied throughout a questionnaire. For example, the following question text has two subclasses: the top line, which is in bold, and the additional text in brackets.

Total persons engaged in the enterprise in the week ending 17 June 2001
(include employees, directors, partners and family members)

V. THE SURVEY ADMINISTRATION MODEL

25. The Survey Administration module provides a tool for running the survey. It links the questionnaire from the Questionnaire Design Tool with the sample and monitors the despatch and return of the questionnaires. Figure 8 shows the main objects in the model. The objects in bold relate to survey oriented activities, and the others are held at the level of the individual. The Survey Administration Tool allows a manager to define samples and to associate them with questionnaires. The SampleUnitResponseAdmin object holds the behaviour associated with the individual (returned questionnaires, not contacted, refused etc). This information can be linked to the questionnaire data at the end of the survey.

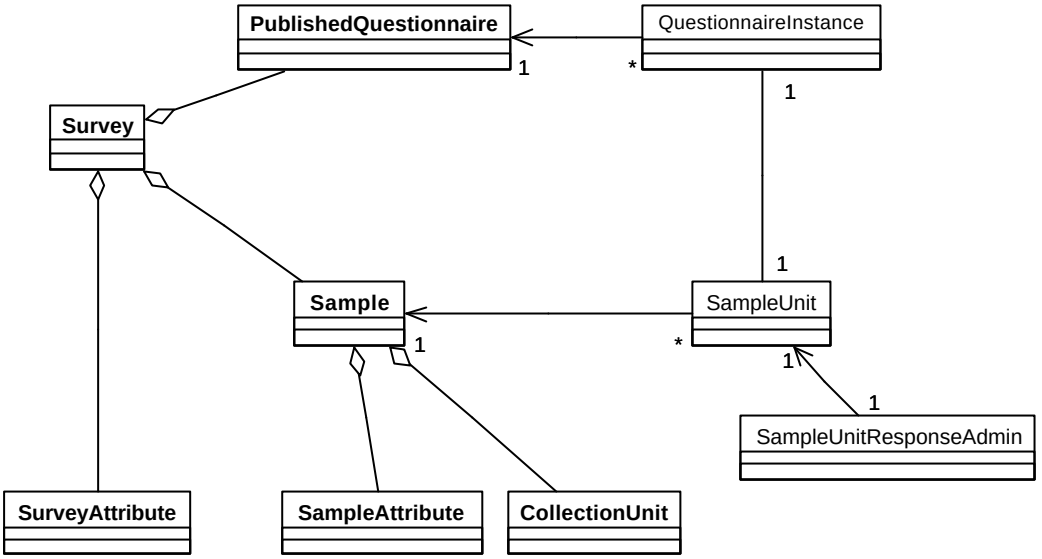

Figure 8

VI. METADATA REPOSITORIES

26. The Metadata Repository is used for holding the metadata of the IQML model, in order for the tools to exchange metadata, and for that metadata to be available for other third party tools. The Repository is independent of the model, which is defined to the Repository at start up. Thereafter, the Repository will reject objects that do not conform to the model. Figure 9 shows the objects in the Repository. The model is a simplified version of the ebXML model [3].

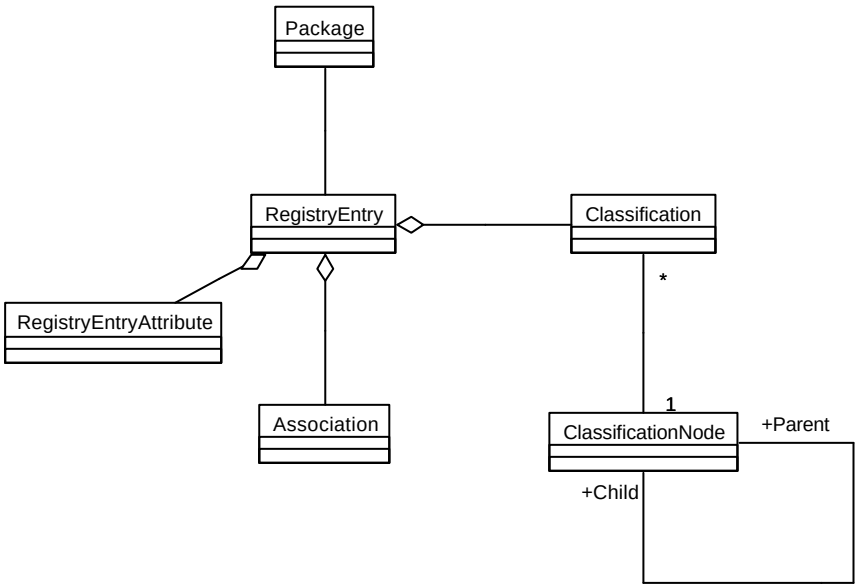

Figure 9

VII. RELATIONSHIPS AND MAPPINGS

27. Figure 10 shows the functional relationship between the modules of the project. The Repository is configured so that information can be shared between applications. From the repository it is also possible to extract the IQML XML for use by the QPT and DIT. We separate out the Application specific metadata, which is held in local storage, from the IQML model that is held in the repository. At present parts of the model are being extended, as more functionality is added.

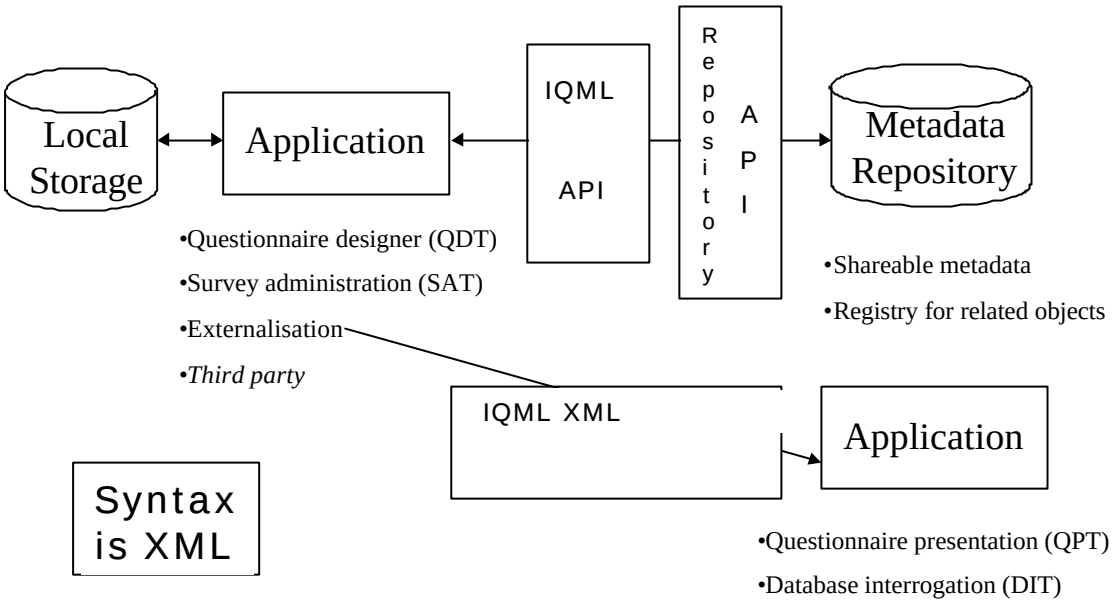

Figure 10

VIII. STANDARDS AND EXCHANGE

28. The IQML XML has been adopted by EEG6/WG4 as the basis for RDRMES/XML. XML documents conforming to this standard can be directly extracted from the metadata repository. In addition, third party software can extract metadata from the repository to conform to their own standards. We have also been developing a Request for Proposals (RFP) in the area of questionnaires and data capture as part of the Analytical Data Management special interest group of the OMG[2]. The issuing of an RFP officially starts the process of adoption a specification as standard.

29. IQML is not working in isolation, In addition to the activities described above, partners contribute to the MetaNet network of excellence [5] and the project belongs to the COSMOS cluster of IST project in the fifth framework project [6]. Hence the models we are developing will be shared with other parties who are interested in modelling statistical metadata.

IX. CONCLUSIONS

30. The IQML Project is a three-year research project ending in January 2003. At present we have developed an initial model, and implemented the first prototype of all modules (and the second prototype of the QPT). The evaluation report of the trial of the first prototype is scheduled for December 2001. During the final year of the project, the common model described in Section II will be refined and developed. IQML is making an important contribution to the development of models and software for

[2] http://www.omg.org/

raw data capture in three ways. First, we have developed an XML schema that will be used as the basis for a message for raw data exchange by EEG6/WG4. Second, we will have a complete set of software for the collection of raw data, and for the storage of the metadata associated with this task. Third we are contributing to the development of a generic reference model for statistical metadata, working with other participants in the field of data capture.

REFERENCES

[1] Booch, G et al, The Unified Modeling Language User Guide, 1999, The Addison-Wesley Object Technology Series, ISBN 0-201-57168-4

[2] Rational Rose 2000, Rational Software Corporation http://www.rational.com/index.jsp.

[3] Webber, D. and Kotok, A. ebXML: The New Global Standard for Doing Business on the Internet, 2001, New Riders Publishing, ISBN: 0735711178, http://www.newriders.com

[4] OMG, Common Warehouse Metamodel (CWM) Specification, .2000, OMG Document ad/2000-01-01, February 11, 2000

[5] Lamb, J.M. MetaNet, A Network of Excellence for Harmonising and Synthesising the Development of Statistical Metadata, 2000, CES, University of Edinburgh, http://www.epros.ed.ac.uk/metanet/

[6] Lamb, J.M. COSMOS: Cluster Of Systems of Metadata for Official Statistics, 2001, CES, University of Edinburgh, http://www.epros.ed.ac.uk/cosmos/