Experiences with Colectica

Felix Coleman, Central Statistics Office, Ireland

1. Abstract

There is currently a transformation program taking place in the Social statistics environment at the Central Statistics Office. A program is under way to develop a more process orientated approach to the production of national statistics. In order to facilitate this transformation, the social statistics questionnaire design team (QDU) is utilising the DDI metadata environment to integrate the transfer of statistical metadata through its statistical production environment. The transformation is designed to develop best practice principles around the creation and storage of statistical metadata. One of the key developments in the work is the systematic mapping of survey design in DDI to the Blaise survey instrument environment and then the "carrying through" and supplementation by Blaise of additional metadata as data exits data collection phase provided by Blaise. This short paper will describe the progress / direction being taken to date with this work and highlight some of the challenges and solutions that have been experienced with the transformation.

2. A brief context: How it was.

Traditionally CSO's household survey process has worked on the basis that its subject matter experts had the responsibility of designing and compiling the questionnaires for their statistical products. As part of this traditional approach to statistical production, the specialist statisticians applied their experience in survey design to the subject matter and produced a questionnaire specification for CSO's Blaise team to develop. This method of design has a number of benefits and drawbacks. Probably the most important perceived benefit is that the statistician responsible for the publication of a survey's results was also responsible / had a direct say over the design of the survey's questionnaire. This also meant that statisticians with the best understanding of the statistical subject matter were able to directly design questionnaires to fit the data requirements of the surveys.

Unfortunately, designing surveys in this way also has some significant drawbacks. These included

- Inconsistent survey design (e.g. the same data being requested from different statistical domains using different questions)
- Poor documentation of statistical process
    - Non standardised
    - Incomplete
- Inconsistent survey specifications provided to IT (Blaise)

So, CSO are developing solutions to some of these issues as part of their household survey transformation project while at the same time maintaining the important benefits that the more traditional approach to survey design provided in the past. To achieve this, we are moving towards the systematic storage of survey information and away from the use of standalone survey specific silos of metadata. This method of data management will allow us to automatically transfer this standardised metadata through the statistical production process to our users (analysis, dissemination, archiving etc.) The preferred method of storing this metadata is by using Data Documentation Initiative (DDI). DDI in simple terms is a standard method of tagging statistical data and metadata elements using XML. One of the leading software tools to apply this tagging is Colectica Designer. This paper explains how Colectica is helping with different parts of CSO's survey transformation project and how we hope to further develop the use of DDI into the future.

3. Survey Development: Collecting information for the Questionnaire Design Unit (QDU).

Part of the procedures being developed by the QDU within the CSO involves the development of and request for Input specification documents from upstream in the statistical process as defined by the Generic Statistical Business Process Model (GSBPM). Generally speaking the input specifications as you might expect contain the important information/metadata that a survey designer needs before they can go to work. More specifically, the lists of metadata items being requested are being drawn from Colectica Designer's design interfaces in order to simplify and highlight the relevant fields of metadata from the large and complicated DDI dictionary that exists.

The Information is being requested with the following principles.

- The metadata that is required by the survey designers is created and transferred to QDU by the persons or departments that carry the responsibility or expertise of the management of these metadata elements.
    - e.g. If a household survey's sampling methodology is being developed specifically by a statistician who is specialised in sampling procedures, then the metadata explaining the sampling procedure is passed onto the QDU by the sampler. (Subject matter specialists will invariable be involved with sample methodologist to organise the sample selection)

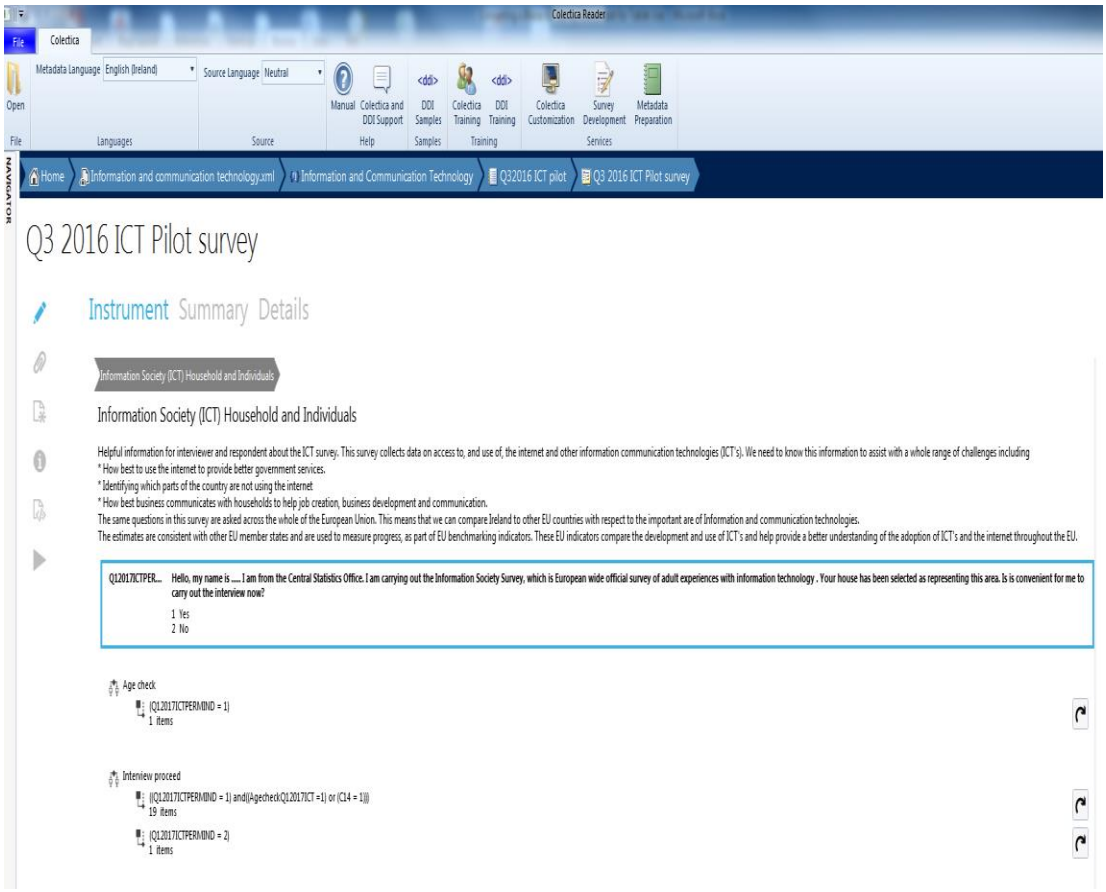
- The critical information provided is expressed in standard metadata titles/tags extracted from the DDI dictionary.
    - This introduces a more precise use of the terminology in DDI and begins the stage of structuring the metadata around survey design into DDI compatible pieces.
    - By just initially using the titles of elements and removing the DDI tagging procedures we move away from the need initially for everyone to be familiar with DDI.

An example of one of our initial input specification requirements can be found in Appendix 1.

4. Questionnaire development

Once the input specification process has been completed and submitted the QDU then begins the process of incorporating the survey metadata from the early stages of the GSBPM into the question bank provided by the Colectica software. The process of metadata documentation being captured at the correct time during the production process begins while at the same time creating data in a structured (DDI) environment. In many cases, the metadata that is associated with the early parts of the GSBPM process has been designed over a long period and is already stored in either an NSI's metadata holdings or via Eurostat's documentation

Figure 1. Colectica Reader



5. Questionnaires and Blaise

At the moment CSO is linking questionnaires stored in its Colectica software with the Blaise team using basic output functionality.

- Word / PDF documentation
- XML DDI (which can be read using Colectica Reader)
- Blaise data models created by Colectica

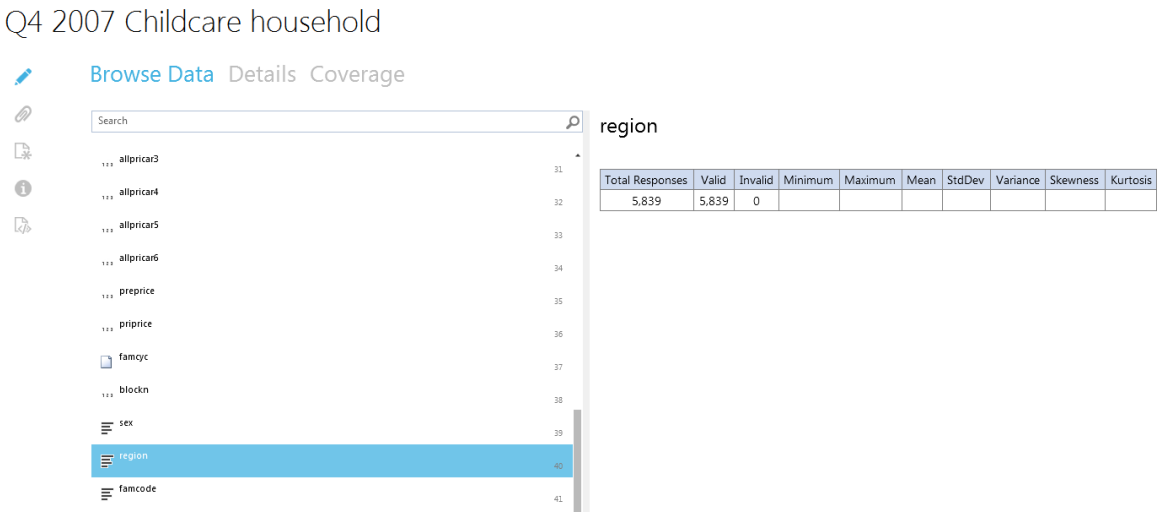
This approach initially creates a more flexible link between our DDI holdings and our Blaise survey instruments allowing the QDU space to develop more consistent metadata packages and a better/greater understanding of DDI and how to structure survey metadata in an organisational context. It is a longer term aim that the questionnaire design team will be able to develop a systematic relationship between the Blaise output from Colectica and the Blaise data models being used in our current social statistics surveys. It is important that the metadata being developed by our subject area specialists in the GSBPM's early stages can be linked systematically into the Blaise process and then output into our metadata and data servers for analysis or publication. At the moment, the Blaise data models being created by Colectica do not match the existing survey instruments we are using in Blaise however the code is useful to our Blaise team as a source for incorporating automatically generated code rather than writing programs from scratch.

6. Non questionnaire data and metadata and Quality reporting metadata

Although the survey design team's main focus is on questionnaire design (question text, answer categories, routing, etc.), it is also important that the specification provided to Blaise includes the

details of additional non question metadata that is required by analysis statisticians in order to comply with statistical regulations and quality reporting. For example, data covering a household's location or the time that interviews took place needs to be included in the metadata specifications but not necessarily the questionnaire specifications. From this point of view Colectica Designer can provide the correct creation / tagging of these elements (as an output variable rather than a question element) but it is not the initial priority of QDU to have this/these in place? We know from experience that data of this nature is being extracted from our Blaise/ survey IT environment and provided as output variables to the results and analysis statisticians. The key part of the process is to be able to incorporate these variables into the documentation process.

Figure 2. Non questionnaire data holding in Colectica Designer



7. To Blaise and beyond.

The Questionnaire Design Unit (QDU) is developing a standard output from Colectica of their work in order to.

- Improve the articulation and precision of questionnaires and specifications being provided to Blaise.
- Better define relevant metadata that is needed to fulfil statistical requirements at the right moment in the statistical production process to work towards best practice principles of documentation and quality.
- Reinforce the principle that there should only be a single source of specific metadata elements in a statistical production process that can be used and reused as required in the overall statistical environment.

In time and in consultation with our IT expertise we hope to look forwards to removing the need for paper based specifications by providing electronic link up of the Colectica data holdings with our other metadata holdings. This will enable the metadata being captured by the questionnaire design team to be used and reused our statistical products move through the production system to dissemination and beyond.

8. Challenges and future improvements

In order to fully benefit from the use of DDI and tools such as Colectica, CSO and other NSI's will need to promote the interoperability of the use of DDI. Some suggestions that could help this process would be

- Early development and sharing of DDI documentation around European statistical products
    - E.g. Model questionnaires provided in DDI rather than traditional Word / Excel.

        - (Colectica Designer allows for multiple language entries for each metadata element).
- Statistical production specialisation by member states and sharing DDI expertise and survey metadata development
- Promoting the interoperability of expertise in different stages of the statistical production process using DDI profiles.
- Eurostat's metadata server could become more of a two way system allowing NSI's to extract standardised metadata packages/ elements for statistical products.
- It would be useful to provide a source for standard classifications within DDI as a menu item. For example if a designer wishes to use a list of standard ISO countries as an answer category for a question, they currently need to import the classification. It would promote consistency if perhaps the UN classifications could be updated with the DDI tools environment.

9. Appendix 1: Input specification requirements template for Question Design Unit

Input specification documents / Input metadata requirements

- Survey NAME
- Survey area
- Creator - (RAP Statistician)
- Contributor
- Specification delivery date required
- Output variable listing
- Links to Previous specifications
- Abstract
- Purpose
- Analysis unit, (e.g. households, individuals, working individuals aged between x and x, enterprises ….)
- Subjects and keywords
- Temporal coverage (reference periods)
- Spatial coverage
- Sample population
- Sample size
- Sample procedure
- Intended frequency
- Mode of collection
- Required response rate
- Related legal regulation
- Testing requirement (quantities, cognitive)
- Cost restraints

10. References

Data Documentation Initiative (DDI) http://www.ddialliance.org/

Colectica http://www.colectica.com/