WP. 18
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD) STATISTICS DIRECTORATE

Joint UNECE/Eurostat/OECD workshop on statistical metadata (METIS)
(Lisbon, 11-13 March 2009)

Session 4 Metadata case studies

UPDATE ON EXISTING CASE STUDY

Submitted by Australian Bureau of Statistics[1]

I. Persistence and strengthening of the "worthy, but not compelling" organisational view of new proposals in regard to metadata management

1. In the last couple of years, major proposed work programs have generally been agreed to be "worthy" by the organisation but not "compelling". "Compelling" appears to have become a higher bar than once was the case, partly because resources for investment in local and corporate development projects were much more limited across within the organisation as a whole and partly because past investment in improving metadata management has generally not realised the anticipated business benefits as readily, quickly and fully as anticipated.

2. Experience over the past year has confirmed this pattern as a "trend" (which we continue to work to turn around) rather than being one or two "irregular" events.

1 Prepared by Alistair Hamilton (alistair.hamilton@abs.gov.au)

II. SESAME Framework

3. In the second half of 2008 we developed our proposed "SESAME Framework" (Standards Enabled Shared Active Metadata Environment). This envisages a registry of "concepts and structures" which would allow preferred data and metadata concepts and structures (drawn from an agreed, internationally aligned, Corporate Information Model) to be clearly identified. "Legacy" concepts and structures (required by, and embedded in, existing repositories and processing systems) could be identified in the same registry (with a non preferred status), together with "structure mappings" that identified how the legacy concepts and structures related to the preferred ones.

4. The mere act of having to explicitly identify concepts and structures associated with each existing repository and processing system, and to establish and agree the mapping of these to the preferred reference model, would be a huge step forward for the ABS. At first the framework would function primarily as a "Rosetta Stone", providing a standard reference model for mapping data and metadata content into, and out of, different environments (enclaves) within the ABS and beyond the ABS. In this way it would provide a standard "bridging" format rather than, potentially, each environment needing to define and maintain completely separate direct mappings to each other environment with which it needed to interact.

5. Over time, however, as repositories and processing systems were redeveloped or replaced they would adopt preferred concepts and structures wherever possible. (Previously, despite good intentions, redevelopments often led in practice to yet another different - explicit or implicit - information model.) Due to the use of standard bridging, such changes internal to a particular environment would not have large impacts on other environments within the ABS "ecosystem". In fact, such internal re-engineering would streamline interactions and lead, over time, to the corporate information model more often shaping in practice "the way things are" rather than just "the way things should be".

6. The SESAME Framework proposed use of SDMX to define structures, and to exchange data and metadata based on those structures, using DDI (Data Documentation Initiative) to describe unit record level datasets. The use of well supported and flexible standards, that already interoperate with other relevant standards such as ISO/IEC 11179 and XBRL, should also also facilitate interoperability beyond the ABS - both in terms of drawing content in from outside the ABS and disseminating content. This also reflects a broader trend toward the practical utilisation of standards driven primarily by pragmatic business oriented reasons (greater interoperability in several senses, better "market" support, better "future proofing" etc) rather than simply as a matter of principle (eg based on the conceptual depth and elegance of the standard).

7. For more background on the SESAME Framework see the following links

http://www.oecd.org/dataoecd/8/11/41860256.pdf?contentId=41860260
http://www.oecd.org/dataoecd/38/35/42019191.pdf?contentId=42019192

8. The SESAME Framework is yet to receive top level sign off within the ABS but has attracted strong practical support from the technology and "dissemination futures" divisions that lay beyond the Methodology and Data Management Division as well as interest from one other agency within Australia and some overseas NSOs.

III. Statistical Metadata and Enterprise Information Architecture

9. There is an ever increasing emphasis on practical meshing of metadata management repositories and services with end to end statistical production process, including applying the former to drive the latter. This was a key aspiration in the 2003 Metadata Strategy. Experience since then suggests that if we are to actually achieve it then we need a very strong, practical framework - which we hope that SESAME can evolve into. Experience also suggests that it requires a melding of traditional "metadata management" conceptualisation with Enterprise Information Architecture (EIA), and, through that, melding with other aspects of Enterprise Architecture (EA). In the ABS context we have found that, to some extent, this addresses a gap in from the EA perspective as well because EIA was perhaps the least well defined element of EA previously. Strong EA, and well defined EIA in particular, appears essential if - for example - implementation of SOA (Service Oriented Architecture) is to coherent across the statistical life cycle and across the different subject matter domains.

10. The central players in terms of ABS EA have adopted the METIS GSBPM (Generic Statistical Business Process Model) for a number of purposes in preference to the older ABS internal model. There is no enthusiasm so far, however, within statistical business areas that understand, and have structure thinking and documentation around, the older ABS model. There is ongoing discussion about how far it makes sense to encourage them to change versus being content to adopt the GSBPM as a reference model focused on external purposes that is supported through mappings from the older internal model.

IV. Statistical Metadata and Statistical Information Management

11. Related to the previous point, the closer integration of "traditional" statistical metadata with data and with other information (eg other processing parameters used to drive end to end operations) means that metadata management is also meshing with a broader field that might be termed "Statistical Information Management" (SIM). In some ways SIM and information interoperability appears to provide a less "alien" heading when talking with other domestic organisations about, what amounts to, interoperability of data and metadata.

12. For some of the agencies concerned production and/or use of statistics is a "secondary" function. Other of these agencies have a focus on other types of data -

geospatial, scientific, research, administrative. Discussing SIM - along side more general IM - reminds them that we are all in the IM space although we approach it from slight different angles. The similarities and the differences are both important, including ensuring the former isn't under-estimated and the latter over-estimated.

V. Collaboration

13. In looking to modernise aging infrastructure, and move to a fundamentally new, holistic, Statistical Information Management framework, the ABS is actively seeking collaboration opportunities. While sharing of already developed capabilities remains one aspect, shared and/or rationalised design and development of capabilities is seen as an increasingly practical and attractive mode. (A rationalised approach might see one agency take the lead on one open, collaborative development while another leads development in a different but related area, rather than each agency spreading its own development resources across each area .) There are both push and pull drivers. The ABS - like most agencies - lacks the resources to simply "go it alone" in terms of achieving the overall desired outcomes within a period that is measured in years rather than decades. Even if we had the resources, however, the "go it alone" approach no longer appears to be a cost effective approach. (Which is a shift from the past where there was genuine room for doubt in regard to both the true cost and the true effectiveness of collaborative/shared approaches versus going it alone.) Modern open standards, various IT developments, various existing international initiatives and broader changes in organisational thinking (ie beyond just the ABS) greatly facilitate the proposed direction. While the primary focus will be on collaboration within the international statistical community, related communities (eg geospatial data, scientific research communities, data archive community) are seen as potential partners in regard to some activities. It is envisaged most collaborations would focus on quite specific capabilities where there is clear practical common interest rather aiming to operate on a grand, sweeping scale. A broader vision and framework within the ABS then becomes the means to ensure that each "little piece" in the puzzle to be filled in through a specific collaboration does fit reasonably within a bigger picture rather than ending up with a patchwork of "point" solutions that fail to join up coherently.

14. The overall "puzzle" being solved by a collaboration partner may be different in detail to the ABS one, as long as the particular piece that is the subject of the collaboration fits well within both. Nevertheless the SESAME Framework (eg in terms of a high level reference model and shared data and metadata structure definitions) aims to have as many aspects of the "bigger picture" shared with other NSOs, which increases the likelihood that there will be practical business cases for the ABS and for other NSOs to undertake collaborative development, based on shared interest, in regard to individual "puzzle pieces".

VI. Looming Challenges for Geospatial Presentation of Statistical Data and Metadata

15. As in many countries, the geospatial presentation of information (of which statistical data is a small but growing element) is receiving a lot of attention from the government, business sectors and the public. While the ABS is making (somewhat slow, but steady) progress on supplying data to the geospatial world on an sustainable basis, a big unresolved question (internationally I gather) is the extent to which, and how, the richness of the statistical aspects of that data (eg measures of, and information on, statistical rather than geospatial quality, statistical structural metadata, annotations related to observations) can be made visible once "meshed up" with geospatial data, using geospatial data and metadata standards and presented through a geospatial lens.