Distr.
GENERAL

WP.7
11 April 2011

ENGLISH ONLY

**UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)**
**CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION**
**STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)**

**ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD)**
**STATISTICS DIRECTORATE**

**Meeting on the Management of Statistical Information Systems (MSIS 2011)**
(Luxembourg, 23-25 May 2011)

**Topic (i): Architectures, models and standards**

# Microdata Dissemination Architectures and Systems

**Supporting Paper**

Prepared by Rochelle Thorne, Paul Nicholls and Kelly Boettcher,
Australian Bureau of Statistics (ABS), Australia

## I. Introduction

1. User demand for access to statistical Microdata for analysis and linking is increasing at a rapid rate. Statistical agencies must respond to this demand while maintaining privacy and confidentiality standards that underpin data provider confidence and adhere to legislative requirements.

2. Increased Microdata dissemination is a major objective of the Australian Bureau of Statistics (ABS) strategic change program, the Information Management Transformation Program (IMTP). IMTP takes a new approach to ABS business, applications, information and technology architecture. The automated flow of ABS data internally and externally will be enhanced by adopting the Generic Statistical Business processing Model (GSBPM)[1], the Generic Statistics Information Model (GSIM), metadata standards such as DDI[2] and SDMX[3] and service oriented architecture (SOA)[4].

[1] http://www1.unece.org/stat/platform/display/metis/The+Generic+Statistical+Business+Process+Model
[2] http://www.ddialliance.org/
[3] http://sdmx.org/
[4] http://www.cio.com.au/article/268531/soa_101_an_executive_guide_service-oriented_architecture_soa_/

3. Such a framework will allow the ABS to provide Microdata dissemination services that will satisfy user demand and mitigate business risk. This paper documents the ABS experience in progressing microdata dissemination, including recent learnings from the IMTP "pathfinder" project, Remote Execution Environment for Microdata (REEM).

# II. ABS Microdata Dissemination Strategy

## A. Background

4. ABS has been making Microdata available to approved statistical users since 1985. Access was initially restricted to the release of a limited range of Confidentialised Unit Record Files (CURFs)[5] on Compact Disc. ABS microdata is made available under Clause 7 of the Statistics Determination 1983[6] with users being required to sign an undertaking regarding their use of this data. Access is limited to authorised users for statistical purposes only.

5. Over the last decade, ABS has subsequently invested significant resources into on-site Microdata dissemination enclaves. ABS released its initial microdata dissemination tools, Remote Access Data Library (RADL) and ABS Data Laboratory (ABSDL) in 2003. ABS is currently developing a system called Micro to automate the majority of manual administrative work associated with managing external client access to microdata dissemination systems. The most significant current investment of ABS resources in microdata dissemination is a next generation on-site tool called Remote Execution Environment for Microdata (REEM).

6. In 2009, ABS released TableBuilder[7] on the ABS website to allow users to define tables from the full Census unit record file. TableBuilder is accessed from the ABS website and allows users to produce output tables that have been confidentialised using a perturbation technique that also protects against differencing. The TableBuilder product has been jointly developed by the ABS and Space-Time Research (STR)[8]. TableBuilder is a key component in the ABS REEM architecture.

## B. Current Business Drivers and ABS Microdata Strategy

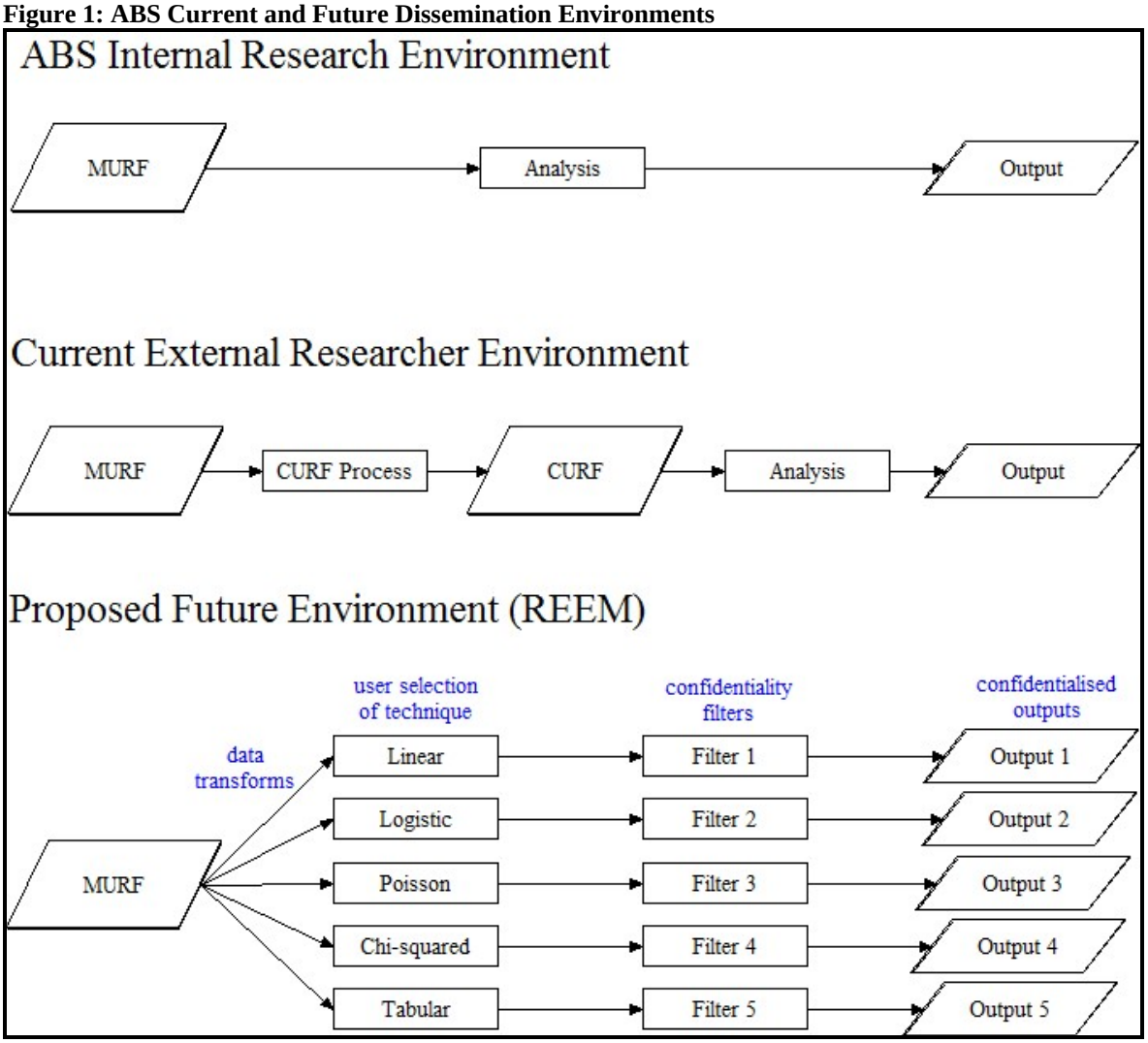
7. User demand continues to grow for access to a wider range of microdata, more detailed unit record data and more flexible analytical tools that provide outputs in real-time. ABS has released a wide variety of household survey microdata but very few business CURFs. Improvements to the CURF creation process will enable the release of more microdata, more quickly, to meet user demand. IMTP will provide metadata infrastructure that will allow for the release of both microdata and searchable metadata.

8. ABS will continue to produce basic CURFs for analysis by users in their local environment. REEM will be progressed as a future replacement for RADL primarily for table generation and basic statistical analysis on more detailed unit record files. ABSDL will be enhanced to allow complex analysis of detailed data including linked and longitudinal datasets. To progress the analysis of detailed unit record files, ABS will need to confidentialise outputs rather than inputs. Menu driven interfaces will support real-time user analysis and improved user outputs such as graphs. Metadata standards, such as DDI and SDMX, will underpin system-to-system interfaces and data management. The diagram below shows current and future microdata dissemination environments (Figure 1).

---

[5] http://www.abs.gov.au/websitedbs/D3310114.nsf/home/CURF:+Microdata+Entry+Page
[6] http://www.comlaw.gov.au/Details/F2004C00203
[7] http://www.abs.gov.au/tablebuilder
[8] http://www.spacetimeresearch.com/

**Figure 1: ABS Current and Future Dissemination Environments**




## C. Information Management Transformation program (IMTP)

9. IMTP is a business transformation program with outputs that will enable the following outcomes:

- Increased quality/reliability of ABS products and services
- Increased granularity of data
- Increased discoverability of ABS data / information
- Increased access to ABS products / data
- Decreased time to market of statistical products / data
- Increased coherence of ABS / other data sources
- Increased levels of service to developing countries within the region
- International collaboration to develop a statistical industry

10. There are a number of IMTP outputs that will enable the ABS to deliver on microdata dissemination objectives. The formal adoption of DDI and SDMX will provide a foundation for ABS metadata infrastructure that is end-to-end, reusable and active. As these standards underpin key strategies, ABS is committed to contributing to the ongoing development of DDI and SDMX in the

international statistical community. The design, selection and development of metadata registries and repositories during the IMTP will support DDI and SDMX services. Such registries and repositories will be the foundation for all ABS business processes including microdata dissemination.

## D. GSIM and GSBPM

11. GSIM and GSBPM are viewed by ABS and other NSOs as key standards to facilitate the "industrialisation" of statistics. They are designed to work together. GSBPM provides a reference model for statistical business processes and GSIM provides a reference model for the information flows between the GSBPM components.

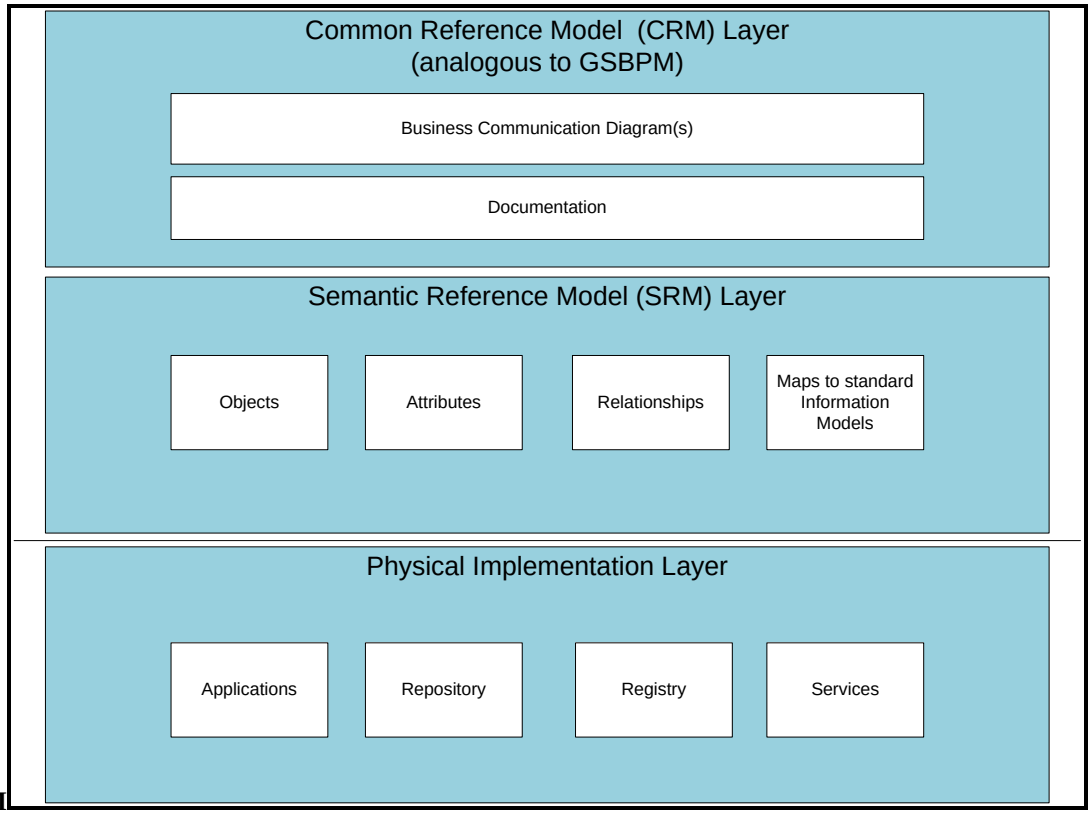
12. The idea of a Generic Statistical Business Information Model was discussed at the MSIS meeting in April 2010. The inaugural meeting of the Informal CSTAT[9] workgroup on stronger collaboration on Statistical Information Management Systems occurred two months later and identified an essential role for GSIM in providing a consistent reference model when defining information required to drive statistical, production processes, and output from these processes. A collaboration team, (OCMIMF)[10], led by the ABS, was tasked to operationalise a common metadata/information management framework, including:

- Develop a generic statistical information model (GSIM)
- Map SDMX and DDI to the GSIM framework
- Influence the evolution of the standards
- Implement SDMX and DDI in a test situation
- Operationalise the use of metadata as a driver for business processes, based on the Generic Statistical Business Process Model (GSBPM)

13. In simple terms GSIM can be envisaged as providing a basis for statistical organisations to agree on common terminology and definitions to aid their discussion on developing metadata systems and information management frameworks. The Business Communication Diagram is intended as a simple (not technical) diagrammatic representation of GSIM (see Figure 2). It is anticipated that during May 2011 early thoughts on the CRM Layer will be made available by the collaboration team and they will encourage review and input from all agencies and initiatives which have an interest in GSIM.

[9] CSTAT is the OECD Committee on Statistics
[10] http://www1.unece.org/stat/platform/display/msis/Statistical+Network

**Figure 2:**




**GSIM**

## E. Services Oriented Architecture (SOA)

14. ABS technical architecture fundamentally takes a services oriented approach. This provides ABS with the ability to cut costs through the reuse of well designed and constructed services. SOA will enable the ABS to provide better client service and improved collaboration with other NSOs due to the loose coupling of systems, both internally and externally. This architecture will enable the ABS to be more agile in meeting client demands for microdata and metadata services, while consistently meeting privacy and confidentiality requirements.

15. SOA will enable ABS to produce a system that supports the National Statistical Service (NSS) and International collaboration objectives by being extensible to allow other organisations to harness and\or supply capabilities, such as confidentialisation methods, and to amalgamate microdata from other NSOs . Improvements to existing services, new methods and new data services will be "plugged in" or "pulled out" of microdata dissemination business solutions, with minimal business risk and minimal change management cost. The management of this constantly changing environment will be transparent to clients using ABS business systems.

# III. ABS Dissemination Architectures

## A. RADL (Remote Job Submission)

16. Background

RADL technical design, like many early Microdata dissemination systems, was based on the Luxembourg Income Study System (LISSY) project. RADL allows users to remotely submit code, (SAS, SPSS, STATA) which is run within the ABS's secure environment, with the output returned to the user. Typical external RADL clients are academics and policy researchers. A significant number of RADL clients are from other Australian government departments. RADL is a multi-tier, multi-server application with a Notes/Domino front-end and a Windows application server at the back-end.

17. Strengths

Microdata is released to approved RADL clients in the form of CURFs. The CURFs available through RADL may be more detailed than those distributed on CD-ROM. This architecture provides a lower risk environment as the ABS is able to intervene when privacy or confidentiality issues are identified.

18. Weaknesses

Submitted code and output are subject to a range of automated or manual checks. Certain input commands are not allowed, including graphical displays of data. Some output is not released back to the analyst if it does not meet ABS confidentiality requirements. Metadata is only available as an external technical guide on the ABS website. RADL is not suitable for advanced analysis. RADL protections are not tight enough to enable analysis of more detailed data. Turnaround times for submitted jobs can be a disincentive for users of the service. The viability of this architecture is under threat from increasing risk of identification due to increased computing power (both hardware and software) and better managed external datasets.

## B. ABSDL (Onsite, Real-time Analysis)

19. Background

ABSDL, is an on-site facility which allows ABS clients with statistical expertise to interactively analyse statistical datasets, primarily CURFs, using a secure environment in ABS offices. ABSDL is primarily an infrastructure system based on remote desktop over a segregated VLAN.

20. Strengths

ABSDL enables richer analysis than RADL for certain types of research. Clients obtain results of analysis in real time. ABS staff are able to intervene when privacy or confidentiality issues are identified.

21. Weaknesses

This system is rarely used, mainly due to the cost which is recovered from users. The costs are high as specialist ABS staff are required to arrange access and equipment and monitor each session.

## C. REEM (Remote, Real-time Analysis)

22. Background

ABS expects REEM to meet user expectations for flexible access to richer microdata datasets such as more detailed household datasets, linked datasets, longitudinal datasets and business survey datasets. REEM was initiated in 2009 as a pathfinder project within the IMTP. A schedule of deliverables,

including integration with other IMTP components such as metadata registries and repositories , will be delivered in a series of Phases. Please see the aspirational architecture in figure 3.

23. Phase 1 – Delivered in 2010

This phase has enhanced TableBuilder functionality:

- Support count based measures (eg events, people, families, households)
- Comply with ABS release policy, including protection against differencing
- Accept DDI format input data from the ABS Information Warehouse (ABSIW)
- Provide a 'Save as' SDMX option to enable data integration with other tools
- SDMX web services to achieve automatic machine to machine queries (limited functionality)
- Preliminary research into confidentialising regression parameters for categorical household
- Relative standard errors and annotation of estimates

24. Phase 2 – To be delivered in 2011\12

Phase 2 includes:

- TableBuilder will be able to manage continuous household data items
- More detailed microdata available for tabulation in TableBuilder
- Enhanced metadata linked to the user interface
- Development of an analysis service
    - Menu-driven queries
    - Modelling including diagnostics remotely executed on detailed data
    - Automated methods and tailored outputs to ensure data providers are not likely to be identified
    - Confidentialising diagnostics including graphical displays
    - Development of demonstrator for regression with categorical variables including some exploratory data analysis and transformations
    - Preliminary research into regression including continuous household survey variables
- Loading of selected historical CURFs into the REEM environment
- Integration of TableBuilder with household feeder systems
- Data store for DDI instances

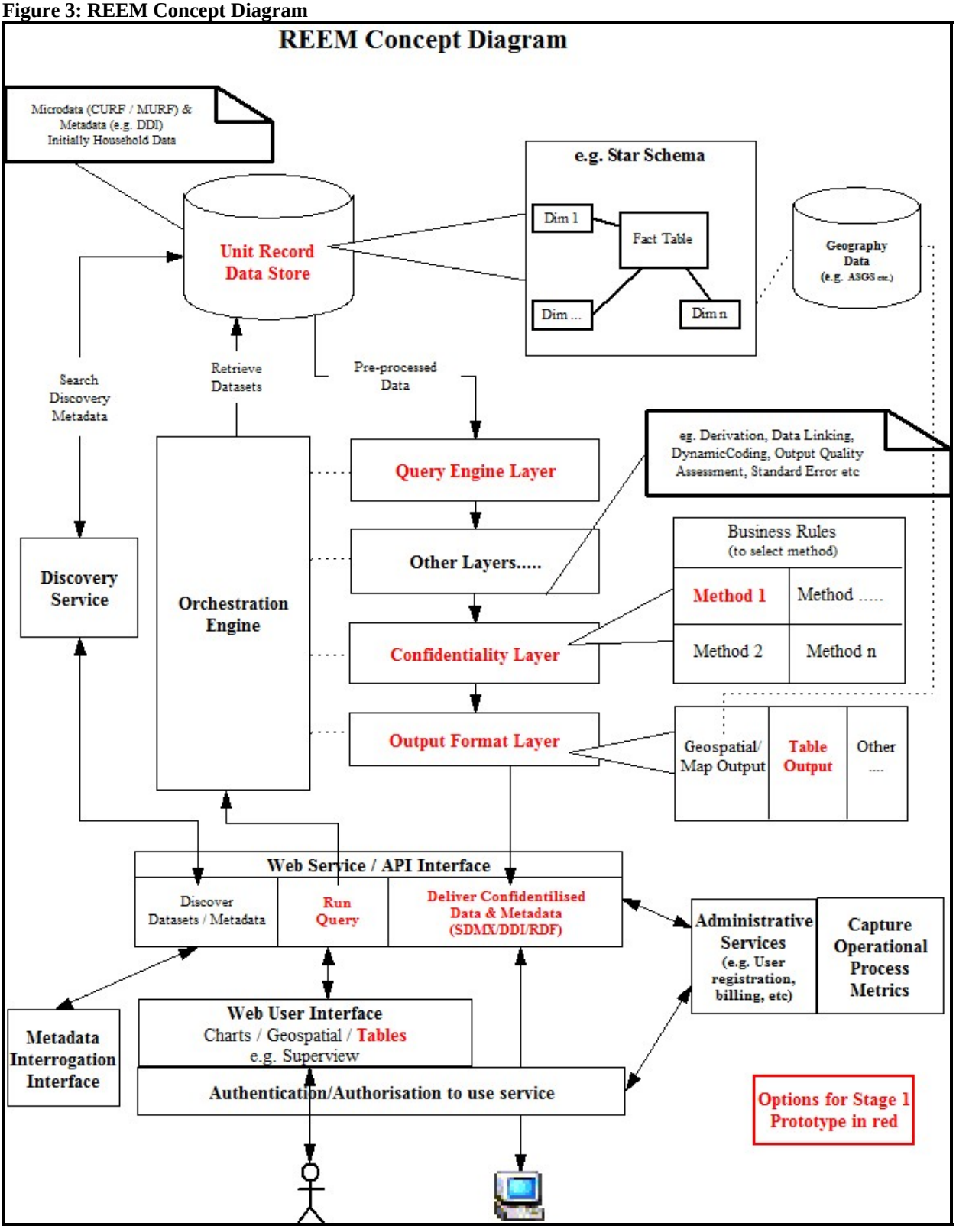
25. Future directions.

Next phases will include:

- Enhanced Survey Table Builder for business datasets including confidentiality methods for tables from business surveys and integration with ABS Business survey feeder systems
- Complex analysis service modules
- Analysis of linked and longitudinal datasets
- Metadata discovery services
- Geospatial mapping display

26. The REEM Concept Diagram (see Figure 3), shows REEM concepts, relationships and interactions. The key features of the conceptual REEM architecture are the clear separation of statistical processes and the structured flow of metadata between components. This design underpins ABS strategy to be responsive and agile in delivery of microdata dissemination services.

**Figure 3: REEM Concept Diagram**



# IV. Learnings and Challenges

27. ABS Legacy Systems and business processes are unlikely to be replaced in the short term. Limited resources will be used to wrap legacy components until funding is available to purchase or redevelop strategic architectural components. Wrapping legacy components will support easier retirement and replacement due to loose coupling between legacy and new components. Legacy components that do not support organisational metadata strategies may need to be remediated depending on their life expectancy.

28. Metadata content required to fill gaps in the metadata between existing components can be manually created to complete DDI artefacts. Automated tools will be purchased or built to assist metadata managers with this process. Underdeveloped features of the implementation of metadata standards, such as versioning and management of unique ids, will need to be addressed internally and then managed by ABS collaboration with NSOs and standards organisations. The lack of existing tools to automate metadata development and management will be a short term constraint, but will give the ABS an opportunity to better understand requirements, identify best practice as well as gaps in the current versions of the standards. ABS will contribute its learnings to the development of such tools in the international statistical metadata domain.

29. Maintaining confidentiality and privacy requirements is a complex challenge for all microdata dissemination architectures. ABS will need to ensure that appropriate resources are applied to the development and comprehensive testing of rules to ensure no breaches of legislation and policy occurs. Confidentiality and output release processes will need to be robust, reliable and easily managed by data administrators, with minimal impact on major outcomes such as performance and usability.

30. The IMTP will invest substantially up front to achieve client education and buy-in from all stakeholders. Engaging business areas to get their support and investment of scarce resources will be critical for success. Clients will naturally focus on tactical business-as-usual delivery, rather than invest substantially in major change programs. It is critical that legacy system owners are aligned with IMTP metadata strategies so that support funding for legacy processes and systems is invested effectively. IMTP will continue to invest in path finder projects, such as REEM, to demonstrate the value of key IMTP strategies early, to all stakeholders.

31. The ABS recognises that in today's financially constrained environment, organisations cannot achieve ambitious change without collaboration with similar organisations and software vendors. Investment in frequent communication at strategic and technical levels is extremely important in establishing clear expectations of outcomes from all parties. Internal challenges, such as reaching agreement on the best way to support standards, will exist on a collaborative level as well. Commercial vendors have an interest in simplifying the implementation and meeting deadlines. This can be counter-productive to our objectives, such as developing best practice in the field.

# V. What Are Others Doing?

## A. International Household Survey Network (IHSN)

32. The IHSN website provides a range of microdata dissemination principles, guidelines and documentation World bank has developed a Microdata Management Toolkit[11] for IHSN including a Metadata Editor that imports data from a variety of formats, allows users to enrich the metadata with an interactive interface, then produces DDI 2.1 compliant output. These tools are easy to use and fit for purpose. To build similar infrastructure to support the more complex DDI-L standard will be a challenge.

## B. High-Level Group for Strategic Directions in Business Architecture in Statistics (HLG-BAS)[12]

33. There are a number of active groups related to the HLG-BAS. An International working group on Microdata Access has met in June 2009 and June 2010. The most recent meeting was focussed on the relationship between DDI and SDMX. The group has agreed that microdata metadata should be compliant to a common international standard and it is likely to select DDI-L to move forward. The next meeting of this group is scheduled for July 2011. The working group has a close relationship with the "Nuremberg Group" which is responsible for a series of workshops focussed on microdata access.

## C. Data without Boundaries (DwB)

34. DwB is a project within the European Community's Community Research and Development Information Service (CORDIS)[13]. Participants in DwB include the NSOs of France, Sweden, the Netherlands, Slovakia, Great Britain and Germany. DwB is an active participant in the CORDIS Seventh Framework Program (FP7)[14], one objective of which is to create a common platform for a lasting cooperation between NSIs and data archives. One WP7 task (T7.7) aims to build and maintain effective collaborations with the DDI Alliance and the SDMX sponsors.

# VI. Conclusions

35. Researchers will continue to demand access to restricted microdata to meet analysis needs that are necessary for their work. Such research is required by official agencies to develop and implement vital policy. ABS will respond with ongoing microdata dissemination policy development, tools and data services that will enable ABS and the NSS community to meet demand for data that is both timely and responsible. Revolutionary business and technical architecture change is required to deliver on this promise. Stove-pipe processing must be replaced with managed business processes, common infrastructure and rich metadata, all based on international standards.

36. Collaboration by ABS with other NSOs and other interested parties (eg OECD, Eurostat, Data Archives) is viewed as the most appropriate strategy to reduce the cost and delivery timeframe, to best harness the expertise and experience which is available globally, and to achieve the best returns on investment for the international community of producers and users of official statistics. Such collaboration will be enhanced by partnering with statisticians and researchers from universities. This

[11] http://www.ihsn.org/home/index.php?q=tools/toolkit
[12] http://www1.unece.org/stat/platform/display/msis/Inventory+of+International+Groups
[13] http://cordis.europa.eu/home_en.html
[14] http://cordis.europa.eu/fp7/home_en.html

will provide conduits for research that will result in innovative solutions to complex problems associated with microdata dissemination. Such a broad body of knowledge will provide stakeholders with an ongoing advantage for the design and construction of future microdata dissemination systems.

37. REEM will provide ABS microdata clients with powerful analytics, better web services and more automated flow of data. The transformation of ABS architecture to a highly structured environment will enable ABS to be more agile in meeting client demands for microdata into the future.