Metadata Modelling for Economic Statistical Classifications

Andrew Hancock[1]

[1] Statistics New Zealand, Christchurch, New Zealand. Andrew.Hancock@stats.govt.nz

Abstract:
As economic statistical data is becoming more accessible, available in bigger and more complex datasets, and needing to be analysed and interpreted in so many different ways, opportunities exist for modernising the development processes for the tools and classifications that support the production of economic data. Metadata modelling along with the use of semantic software tools enables significant advances to be explored in the way that traditional economic statistical classifications and economic data are developed, maintained, updated and implemented.

The system of economic statistics is one where there is significant overlap in concepts, definitions, classifications and metadata which often makes search and discovery by non-expert users challenging. Greater uptake of semantic web technology, such as Simple Knowledge Organisation Systems (SKOS), and Resource Description Frameworks (RDF) provide innovative methodologies for managing and describing economic data and introduce a dynamism and flexibility that did not previously exist.

This paper explores the use of metadata, conceptual and entity modelling to replace the traditional methodology of hierarchically structured, sequentially code based statistical classifications that currently underpin economic statistics.

Keywords:
Economics, Classifications, Metadata Statistics

1. Introduction:

Traditional statistical classifications provide the taxonomical basis for information management, data description, and production of economic statistics, and are a fundamental component of key economic frameworks such as the System of National Accounts (SNA), the Balance of Payments manual (BPM) or System of Environmental-Economic Accounting (SEEA). But they are difficult to use, maintain and update across statistical systems or information management systems because of their complexity of detail and their traditional hardcopy nature. The data becomes increasingly less reflective of the contemporary reality of the global economy.

In a world driven by the use of social media tools such as Instagram™ or Twitter™ human communication and interaction is changing. Economic statistics need to take account of that changing world and move accordingly. Staying with traditional approaches and frameworks provides consistent time-series but doesn't enable contemporary data to be created that will influence economic policy and decision-making.

1.1 Background

Whilst economic statistical classifications are a cornerstone of official statistics, the continued belief that turning text into code for storage, data production and analytical purposes using mutually exclusive and sequentially numbered categories in structured statistical classifications is no longer a viable approach. There is often great expense in time and resourcing to develop, revise and maintain a statistical classification which negates the ability to reflect the contemporary real world of economic data. There is also a need for good supporting metadata, something economic statistical classifications and standards do not always provide

Metadata standards such as the Statistical Data and Metadata Exchange (SDMX), ISO 11179 Information Technology – Metadata Technologies, or the Data Documentation Initiative (DDI) are based upon conceptual or entity models that chunk content down into component parts for easier understanding, usage and consumption

Economic statistical classifications, like metadata models, have concepts, definitions, codelists, entities, categories and other similar attributes. However, those components are not seen in isolation from each other, only as a whole because as human beings we have limited capacity to deal with data growth. This adds to the realisation that traditional approaches to classifying economic information need to change as human intervention is gradually replaced with machine learning and automation.

The real-world changes rapidly and the traditional classification and standards model doesn't keep up and is no longer supportive of economic data. For example, ISIC broadly defines manufacturing as the physical or chemical transformation of materials into new products in plants or factories using power-driven machines and materials-handling equipment, and that it is about the transformation of materials into new products. ISIC either classifies activities into the types of processes used for manufacturing or by the products produced.

Does this mean that all transformation of materials constitutes manufacturing? No – logging, whereby a tree is cut down and turned into logs is not considered manufacturing, but then the logs when transformed into building frames or furniture form part of the manufacturing process. There is a conceptual relationship which may not be easily articulated or visible in a classification structure without a lot of cross-referencing or inclusion/exclusion text.

2. Methodology:

Innovative classification approaches can be introduced to alleviate the pain caused by the traditional methodology and this will need to be undertaken in a considered and transitional way. The variety and volume of economic information from sources that didn't exist 10-20 years ago such as ATMs, Global Positioning Systems (GPS), mobile phones, supermarket scanners, internet activity and social media, highlights new activities, new ways of describing entities and categories for which the current economic statistical information model cannot keep up. The use of relational databases, innovative classification management systems, computer created matrix software, advances in ontological engineering, semantic web and other ICT technologies need to be utilised to improve the development of economic classifications and the search and discovery of associated metadata.

What matters is semantic consistency across the measurements of economic data, something that isn't ideally achieved within the current way of developing and maintaining economic statistical classifications.

Metadata modelling provides a new way of thinking which begins with a clearly defined concept, which may then have relationships to any number of other concepts or sub-concepts. Each concept is unique and forms a scope for all the entities or words that may then be categorised by that concept. This then leads onto the use of entity-relationship models and relational database thinking as a way forward for how economic classifications can be better developed and integrated. It is about better identification and description of an information object, how it behaves, its function and use, how it relates to other information objects and how it is managed over time.

A move to a matrix style approach of relationships formed by linking multiple categories together and away from the traditional parent-child structure will enable more fit-for-purpose views of concepts and provide users with greater flexibility around their data, without comprising consistency.

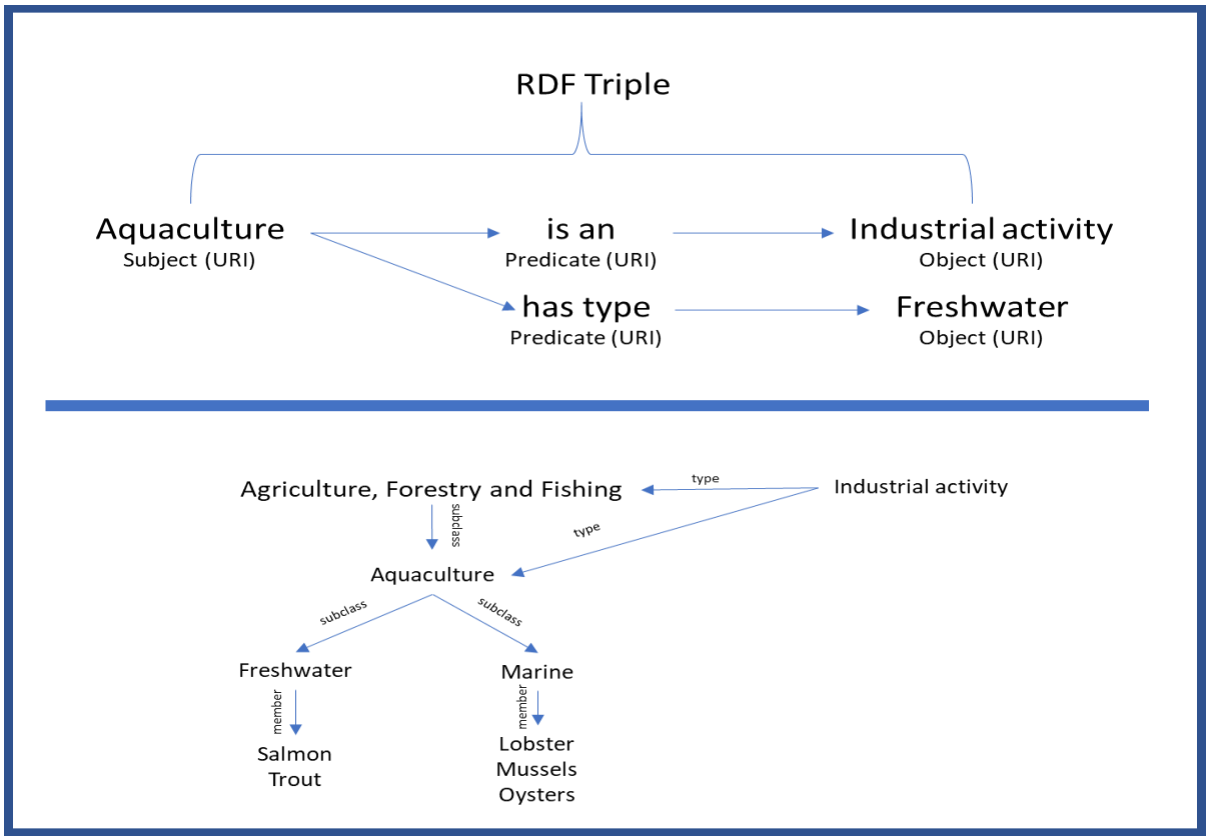
Changing from a traditional approach to the development and maintenance of economic statistical classifications provides numerous opportunities for efficient data management. There is a need to exploit existing data resources and map data, systems and analytics particularly if there is greater need to assimilate big data and/or open data processes and principles.

Using new methodologies will require service-oriented architecture (SOA) which allows for integration with other system components or platforms, especially when utilising a cloud environment. The uptake of the Simple Knowledge Organisation System (SKOS) and/or integration of other systems such as the Statistical Data and Metadata Exchange (SDMX), ISO/IEC 11179 or the Generic Statistical Information Model (GSIM) provide information models which specify concepts, relationships, rules and other elements that are not dissimilar to the content of economic statistical classifications.

Introducing these different methodologies enables better usage of taxonomies, thesauri, ontological engineering and concept management ideas to mix structured and semi-structured data to give new insights for how economic statistics can be produced.

RDF is used for representing resource information on the internet and is a tool for sharing information. It uses unique web identifiers for describing resources or entities which makes it a powerful option for structuring and storing economic classification content and provides a quantum leap forward for sharing an identical concept or category across multiple classifications. For example, the concept of aquaculture could be stored once and reused or shared across an industry, product, trade or sector classification, rather than having single entities in single classifications, each with potentially differing definitions.

The other aspect of RDF which makes it so powerful is the RDF triple. This comprises a subject (which describes the web resource for the information), predicate or relationship (which defines a property that the information is sought about) and an object (which contains the value for that predicate) enabling classification content to be disassembled into component parts to enable easier integration and sharing with other systems or frameworks. An illustrative example relating to aquaculture is shown here:



By connecting the triples, a graph network of relationships is defined within a set of controlled vocabulary terms . A graph network is a set of nodes joined by a set of lines or arrows and can be created using graph knowledge software. An example of the software is SPARQL which is a query language used to retrieve and manipulate data stored in an RDF format.

In SKOS, concepts can have multiple relationships like the notion of an extended human family, electronic thesauri or neural network model. A traditional economic classification nearly always requires a parent-child relationship between each level due to the narrower to broader aggregation approach for refining the groupings.

Using SKOS, concepts can be identified by using unique resource indicators (URIs), labelled with lexical strings that can utilise multiple languages, which can assign notations and link to other concepts and organise this into informal hierarchies and networks using defined concept schemes.

Making use of URIs changes the way in which content can be labelled, used and discovered which removes the constraint of single descriptors or mutually exclusive labels. In addition, the use of synonyms or aliases for economic categories provides additional flexibility and power for describing and presenting information.

A major advantage of this approach using SKOS is that it makes for more granular metadata and easier integration accompanied by greater ability to share concepts and content across different classifications and/or views. This eliminates the time-consuming and costly overhead that comes with then having to create mappings or correspondence tables between classifications.

3. Result:

The move to a concept-based classification model is in some respects a natural evolution of the process of developing statistical classifications. As noted earlier, classifications are comprised of component parts - some of which are given more prominence than others when developing the classification. For example, the general approach is to identify all the things that need classifying and then it becomes a top-down, bottom-up process of determining what parent-child relationships are needed, how many levels are required, what the top or broad levels need to be and then putting the whole structural development on a sequential code pattern. This approach does not fully maximise the potential of the component parts nor allow a true reflection of the concept that is being measured.

As a starting point for the new way of doing things the concept becomes the crucial element or entity around which everything is built. Each concept is given a label, and a definition which needs to be agreed upon by users, such that the definition forms the scope of what the concept measures. As with the use of subjects, objects and predicates within RDF, each concept will also have a relationship to other concepts which enables a conceptual framework to be created and an easier way for merging and transferring data – in some respects removing the need for one-to-one classification correspondences but also enabling a faster way of creating those correspondences. As concepts are related, the categories they contain are then linked to the other concepts very much in the vein of a neural network or electronic thesaurus.

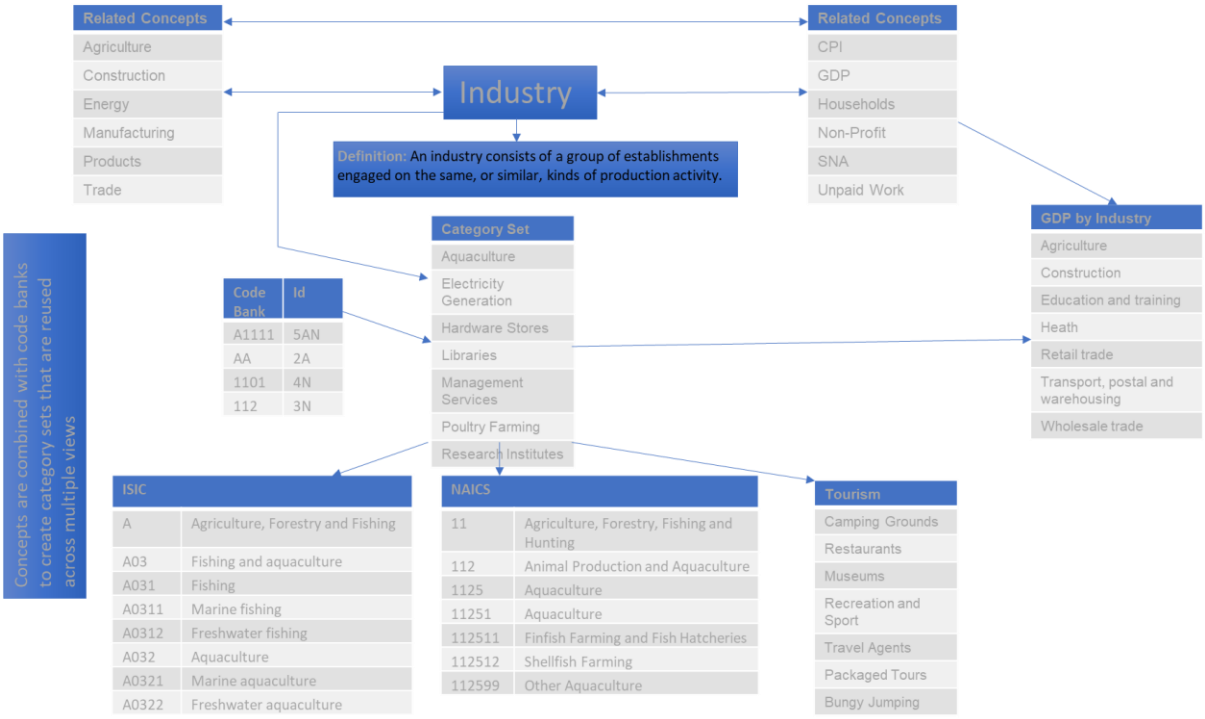
Sitting underneath the concept is a category set which is simply all the words that fit within the scope of the concept definition – very similar to an SDMX codelist (but without the codes). These words can be user defined/suggested labels which can be dynamically added to and updated, and which could include words that would traditionally be confined to an alphabetic index or coding index.

Alongside the concept and category set is a code bank which provides all the codes that could be used with the concept. This code bank holds the approved codes, whether alpha, numeric or alpha-numeric, relevant to the concept and which can be used in the creation of any view or classification from a category set. The bank will potentially be linked to the category set by a universal resource indicator (URI). Again, these can be user defined or standardised but by storing these in a bank, users can determine which codes best suit their system or output needs. It also removes the problem that occurs when implementing economic statistical classifications in that users want to have an alpha-numeric code, but the standard only allows for one specific code type such as numeric.

Finally, the outcome of all this is the introduction of views of the category sets – these can be standardised, or user defined and are reusable and able to be shared. These views of category sets (or what we used to call statistical classifications) are effectively a cut and dice of a master list of content like ‘Lego™ blocks’, all linked via an overarching concept and only including content that is within the definition of that concept.
The category set is a dynamic list of words associated with the concept that can be added to at any time. All instances are time-stamped and approved (either manually or automated) and users are then notified of changes to the master list. Users can choose to adopt change immediately or business rules can be applied to the concept to release updated content at regular points of time, for example quarterly, six-monthly or annually.

In this example below, noting it is only illustrative of content, the primary concept of industry has a relationship to many other concepts which allows for the creation of related views. The concept has a definition, a category set, and a code bank attached to it. The category set can be used to produce standardised views to represent ISIC, or the North American Industrial Classification (NAICS), or an output view for the Tourism sector for example. Additionally, aggregated views such as GDP by industry can be created because of the relationships

between the concepts which allows the category set to have wider application or linkages.



This provides a significantly greater range of outputs and views for users to create and to match with their specific datasets. The interlinkage of the concepts means that other content can be pulled through to provide hierarchical structures, flat lists, cross-cutting views or amalgamated views of concepts.

4. Discussion and Conclusion:

To move the development and maintenance of economic statistical classifications into the ideal of metadata modelling and conceptual classification management will take some time. But the benefits for national statistical offices in terms of cost-reduction, better resource utilisation and greater responsiveness to user demand, outweigh the continuation of the traditional time-consuming process of developing and maintaining statistical classifications that are out of date upon publication. Applying the thinking of metadata modelling and the greater use of conceptual relationships that are fully described, and which utilise the best features of the semantic web is the most practical way forward. Such an approach contains a wealth of information about the concept used in classification and provides rich and flexible information about the relationships and properties within economic statistical classifications.