The Modernisation of Statistical Classifications in Knowledge and Information Management Systems

Andrew Hancock
Stats NZ, Christchurch, New Zealand
Andrew.Hancock@stats.govt.nz

Abstract: As technology transforms knowledge and information management systems, statistical data is becoming more accessible, available in bigger and more complex datasets and is able to be analysed and interpreted in so many different ways. Traditional approaches to the development, maintenance and revision of statistical classifications no longer support or enable description of data in ways that are as useful to users as they could be. The ability to search and discover information in ways that were previously not possible means that new methodologies for managing and describing the data, and its associated metadata, are required. The development of structured lists of categories, often hierarchic in nature, based on a single concept, limited by the constraints of the printed page, statistical survey processing system needs, sequential code structures or narrow user defined scopes results in statistical classifications neither dynamically reflecting the real world of official statistics nor maintaining relevance in a fast changing information society.

Opportunities exist for modernising the developmental processes for statistical classifications by using, for example, semantic web technology, Simple Knowledge Organisation Systems (SKOS), and Resource Description Frameworks (RDF), and for better describing metadata and information within and across multiple, interconnected information and knowledge management systems. These opportunities highlight the difficulties that come with using traditional approaches to statistical classification development and management, and encourage new thinking for different and more flexible options for developers and users. This paper explores the need to dispense with traditional practices for developing statistical classifications as cornerstones of metadata, knowledge and information management, and comments on the need to change the underlying methodology within statistical classification theory, best practice principles and how they can be used in associated information management systems.

Keywords: Classification Theory, Metadata, Statistics, Taxonomy

1. Introduction

Traditional approaches to the development and maintenance of statistical classifications have generally been unchallenged or unchanged for as long as national statistical offices have needed to produce statistics. This is due to the stove-pipe production processes that these agencies use and the need to store information in machine readable, coded form. Turning text into code for storage, data production and analytical purposes, and for producing outputs and aggregations has required structured, usually hierarchic, statistical classifications to enable this, often at great expense in cost of time and personnel to an organisation.

Statistical classifications are the cornerstone of statistics and without them survey information, responses and data is somewhat difficult to interpret and use. They are a fundamental requirement for any statistical system and their role in official statistics is significant. This is increasingly so as national statistical systems explore ways to integrate and identify new data needs for the global indicator framework and 2030 Agenda for Sustainable Development as articulated at the 2017 meeting of the United Nations Statistical Commission and available on the United Nations Statistics Division website (http://unstats.un.org/sdgs/metadata/).

Statistical classifications can also form the basis of taxonomies for document storage and knowledge management in a range of applications. For example the International Classification of Statistical Activities (UNECE, 2014) is an illustration of this, as is the work on taxonomies in XBRL undertaken by Statistics Netherlands and discussed by Roos (2010). They also support keyword strategies for search and discovery of information across relational databases. But they are difficult to maintain and update, particularly when used across statistical systems and when embedded in document management systems or repositories. Often there is disconnect in their use due to the contextual situation of statistical application versus knowledge management. This inconsistency arises through the different mind sets of the designers of statistical classifications compared to the knowledge managers, and the purposes for which the classifications are then used or needed. A strong organisational culture is the most important factor for achieving successful

Reference this paper as: Hancock D, "*The Modernisation of Statistical Classifications in Knowledge and Information Management Systems*" *The Electronic Journal of Knowledge Management Volume 15 Issue 2 2017*, (pp126-144) available online at www.ejkm.com

knowledge management (Lee and Lee, 2007) but this can prove to be the biggest impediment when trying to articulate change, and change the way in which people can reuse codifiable information.

As with flexible and easily maintained knowledge management systems being a driver for organisational improvement, statistical classifications are essential for managing knowledge effectively but whether they are the most practical solution is really the key question.

Yet there is always scope for improvement and for challenging the existing thinking to allow for innovative and new processes to be explored. Technology allows more and more transmission and exchange of information and data than ever before. The use of collaborative tools to enable people to work together and across locations, the use of web technology and portals that can provide access to information and manipulation of that information means that traditional classifications that have underpinned metadata and outputs are actually hampering the description, storage retrieval and reuse of data.

Statistical classifications have often acted as a proxy for metadata registry systems because of their use of definitions and descriptions, scope and characteristics of data, guidance on use of data and the fact that they organise data in a standardised way. Their use as a proxy for metadata standards also stems from the variety of metadata standards in existence and the lack of clarity around their purpose and how those are often only developed for a specific context. Examples of this are the Statistical Data and Metadata Exchange (SDMX), ISO 11179 Information Technology – Metadata Registries, and the Data Documentation Initiative (DDI).

For the purposes of this paper, metadata is defined as data that defines and describes other data (ISO/IEC, 2004). Effectively it is descriptive information about data which explains the measured attributes and structure of that data.

1.1 Background

In an ever-changing data-driven world and in an environment of increasing big data and the sheer availability of information, a new information architecture and associated thinking needs to be developed to enable a modernisation of statistical classifications in knowledge and information management systems. Without taking this approach national statistical agencies will struggle to be more cost-effective in their data production and dissemination processes, and will continue to have difficulty producing real-time insights into the data using relevant classifications and well-articulated concepts. With diminishing resources and demand for automation statistical agencies need to change their technical infrastructure and store information in better structured and conceptually organised systems, and to encourage reuse of existing content rather than proliferation of like. It is that challenge to achieve a balance between reuse of data against the cost in codifying that knowledge that underpins the desire to make inroads into changing the use and understanding of statistical classifications. The world needs to be able to better share ideas, issues, problems and resolutions but to do so requires a form of standardisation of concepts and content to enable those conversations to be about the same thing. In other words the data and knowledge needs to be classified, codified and described in a consistent way such that everyone starts from the same page of understanding.

1.2 Purpose

This paper seeks to raise questions on whether the traditional classification methodologies are still valid, and whether innovation derived from methodologies used in knowledge and information management systems makes life easier for statistical taxonomists, developers and data users now and in the future. Exploring the possibilities that come with the integration of big data and taking a digital first approach for sourcing and producing statistical data, and in particular exploring and utilising metadata information modelling systems will facilitate and encourage a shift in traditional thinking on the way statistical agencies approach the development and maintenance of their existing and future classifications. It should be noted that this paper does not explore data mining as a methodology for improving the developmental techniques for statistical classifications, even though that provides significant options for modernisation of classification development. There are advantages in using statistical analysis, neural networks and data visualisation methods to find information and patterns to build classifications (Shaw et al, 2001) but this was not a consideration per se in undertaking the research described in this paper. Nor does this paper truly explore ontologies for classifying text documents even though there are similarities to approaches discussed in the literature, for example as explored by Lytvyn et al (2017). The paper concludes with a case study of developing a tool to support the change in philosophy that has been undertaken by Statistics New Zealand.

2. Methodology

This paper has been undertaken as a piece of interventionist research (Dumay, 2014) to not only try and resolve an organisational problem but present a way-forward through the discussion on the innovative approach described in the case study.

Much of the literature relating to taxonomical and ontological advances within knowledge management is focussed on search and discovery elements, and enabling better storage of information. Whilst this is pertinent to the future of statistical classifications, there seems little discussion by statistical agencies on new approaches in relation to better data integration and description. Agencies will argue that they have databases with conceptual, semantic, syntactic and structural organisation, but whether the database can be observed or analysed at the appropriate level may well be hindered by the nature of the statistical classification that underpins the data (Thanos, 2017).

2.1 Rationale

The point of engaging in this research has been to identify new insights and manage innovations to improve the quality and timeliness of statistical data and associated information and knowledge. The work has begun from questioning why the current methodology is still being used within national statistical offices. Is there a better way of managing knowledge and statistical classifications that gives the user what they are really after, and which enables a better definition and understanding of the available data? This also provides an opportunity to investigate what knowledge, both institutional and external, requires support to improve the accessibility of information that describes the statistical data.

Primarily the work initiated from the organisational perspective, in terms of where money can be saved and where long-term efficiencies can be gained. An aspect of this initiative was to effectively enhance the ability to capture, retain and exploit knowledge and enable creation, aggregation and use of new information. In particular it was about whether there is a problem to solve or an organisational process to be improved. However, this quickly evolved into an ideal that if we started with a blank page, and had no previous knowledge or expertise in statistical classifications or information/knowledge management, what would we do?

It was important to also treat this piece of work as one component in understanding the approach for an organisations knowledge related strategies, not as a replacement for, or alternative to, centralised information or codified data. It is also an exercise in changing a traditional approach to classifying information as opposed to making any distinctions between tacit and/or explicit knowledge as discussed by Kabir (2013). Whilst tacit knowledge exists, it has to be encoded to enable exchange and distribution and often tends to be confined to specific practices and methods of communication and transfer. Statistical classifications are entirely supportive of this application just as when knowledge is easy to access and transfer. However their usefulness in expanding the pool of knowledge remains limited because of the strict rules of codification that generally apply and the formality in structure they provide. So relaxing the rules that are traditionally used in taxonomy or concept development and management can make the sharing of information more intuitive and flexible, thus reducing the need for traditional statistical classifications.

The challenges in attempting to change the way in which statistical classifications are developed and used to support data, are very similar to approaches for building business ontologies as discussed by Gavrilova, Leschva and Strakhovich (2015). Ontological engineering provides a mechanism for specifying hierarchic conceptual relationships which then provides a more flexible and visual approach for sharing information, knowledge and statistical data. This approach can work very well for the redevelopment of statistical classification structures but is also a consideration in the approach discussed in the case study later in this paper.

2.2 Outcomes

The desired outcome is to enable construction of a conceptual model that structures information through the creation of a relationship matrix for the content. This supports search and discovery of information and allows for a form of diagrammatic representation of metadata through the potential use of broader to narrower concepts, and the production of subsets which can be aggregations or disaggregations of concept based sets. This leads to better categorisation of knowledge and information, and enforces consistent usage of concepts and terms, and a move away from static, hierarchic structures for classifying data. The other overarching challenge in changing methodologies is the statistical need to codify all survey data for ease of storage, manipulation and analytical purposes. Turning text into code is a subjective, labour intensive approach which limits the ability to share meaningful information and knowledge as the focus is on the code rather than the text and its semantic relationships.

The case study presented in this paper may not advance the science of knowledge or information management as such, but it exemplifies advances in concept and classification development and management that enable real-world information gathering, integration and dissemination. The vision and philosophy that sits behind the ideal that statistical classifications must be revitalised in their development and maintenance if they are to remain relevant, will transform national statistical agencies. The case study does expose thinking around viable options for improving and transforming a traditional science and enabling a more dynamic approach that can be supported by a cloud-based tool, and which can be utilised by any person or organisation wanting to manage information and classification related knowledge. This transformation in approach maintains alignment with the steps for a knowledge process in that the knowledge content is created, captured, processed and stored to enable retrieval and access as described by Staab et al (2000). But what does transpire from investigating the impact on the transformation processes is how ICT system architecture influences the decision-making and the case study provides an example of this.

3. What Are Statistical Classifications?

The starting point for exploring the modernisation of statistical classifications in knowledge and information management systems is to begin by stating what they are and why we need them in the statistical production and dissemination process.

3.1 Background

Human beings, through the evolution of cognitive development and science, have a need to classify the world around them, to organise information and knowledge (Salkind, 2004), and to turn complexity into order. This is particularly important in the world of official statistics where there is a myriad of complex and often unrelated concepts, information and data, which is poorly stored, managed and described. Whilst classifications have been used to structure much of the information that is available in the statistical world, those structures often do not adequately support knowledge sharing or effective information management. This is particularly problematic for the electronic transfer or communication of information as classifications are very much point-in-time reflections of data. Transferring the appropriate knowledge to the right person or place often then requires the translation of data from the existing classification structure into a new format or different aggregation, something a traditional classification structure cannot easily allow. Standards, and in particular standard statistical classifications have been seen as important for enabling reuse of data and for transforming raw information into manageable data for wider consumption. However they are hampered by inconsistent use within subject matter areas of national statistical offices and other organisations.

3.2 Definition

For the purpose of this paper, the definition documented in the United Nations Best Practice Guidelines for Developing International Statistical Classifications (UNSD, 2013, p.5) applies when using the term statistical classification.

> *"A statistical classification is a set of categories which may be assigned to one or more variables registered in statistical surveys or administrative files, and used in the production and dissemination of statistics. The categories are defined in terms of one or more characteristics of a particular population of nits of observation. A statistical classification may have a flat, linear structure or may be hierarchically structured, such that all categories at lower levels are sub-categories of a category at the next level up. The categories at each level of the classification structure must be mutually exclusive and jointly exclusive of all objects in the population of interest."*

Fundamentally a statistical classification is a set of discrete, exhaustive and mutually exclusive categories which can be assigned to one or more variables or concepts used in the collection and presentation of data, and which describe the characteristics of a particular population. Those categories describe the data and support knowledge sharing provided there is accompanying robust, clearly identified and well described metadata.

3.3 Why are statistical classifications necessary?

Understanding statistical data requires consistent and robust frameworks that enable interpretation of input from survey responses and/or statistical or administrative datasets to produce relevant and understandable output. That output has to provide data insights that are readily able to be analysed to support policy and decision making, and which informs society. Often classification systems are seen as a cornerstone for knowledge organisation (Geipel, 2015), information management, and for providing the metadata which is so essential to users in interpreting their data. Statistical classifications are an essential foundation block in building a statistical picture of a nation and the world but the current methods for creating them do not enable an accurate and flexible picture of that world to be created. Stand-alone statistical classifications, also do not enable reuse, sharing and integration of content across concepts or systems thereby reducing the ability to increase knowledge and information sharing.

Statistical classifications have traditionally been seen as necessary because they enable the grouping and organising of information, knowledge and data into meaningful, systematic sets according to common criteria or concepts to facilitate the production of official statistics. For example, they:

- identify the similarity between events, ideas, people, information and concepts;
- provide a framework to assist information and knowledge management, policy and decision making;
- establish trends for comparison of information and data over time;
- allow the aggregation and disaggregation of big data and complex datasets in a meaningful way for analysis; and
- facilitate the harmonisation and coordination of statistical information and data compilation and comparability worldwide.

The use of common statistical classifications and standards has always been a significant criteria of the United Nations Fundamental Principles of Official Statistics for informing government and society with statistics and information that are practical, relevant, consistent and impartial. In particular, principle 9 of those principles states: “The use by statistical agencies in each country of international concepts, classifications and methods promotes the consistency and efficiency of statistical systems at all official levels.” (UNSD, 2014, p.2)

Statistical classifications provide a means for facilitating data exchange and for managing information. They do not always enable value to be added to the content or context of the data due to the nuances that exist within the data and metadata that cannot be reflected using rigid and often inflexible classification structures. Further they do not easily enable the reuse of existing content because of their perceived inflexible nature. Statistical classifications are but one component in the technological infrastructure that comprises an information or knowledge management system. Without them however, the data loses a significant proportion of its meaning but the value of that meaning is currently limited through the lack of innovative approaches to classifying and codifying information that current statistical classification approaches promote. The value of the meaning is also hampered by the traditional fixation that text needs to be turned into a number, or code to make it useful, storable and retrievable. Codification therefore becomes a millstone for a classification developer to remedy as the code ranges and patterns constrain the structural build of the classification as exampled by the International Statistical Classification of Diseases and Related Health Problems ICD-10 (WHO, 2010). The focus on codes means that the creation of semantic relationships are difficult and the ability to advance organisational capability through search, discovery and sharing of information becomes constrained.

The current usage of common statistical classifications provides the ability to integrate and re-use data in a limited way which doesn’t always ensure that information is collected once and shared more easily by suppliers and users. It also reinforces the scenario whereby people and systems retain pieces of information or knowledge that could be better described and shared if statistical classifications were more flexible in their approaches and structures.

There is greater need for flexibility and innovation which can be obtained through the investigation and adoption of knowledge and information management strategies to build from classification concepts into relatable views of data and information. This is where the modernisation of statistical classifications can occur, otherwise the traditional approaches really only give a limited understanding of a data driven statistical environment, and only continue to provide an inefficient framework for classifying, storing and sorting information and data in knowledge and information management systems.

4. Why Modernise Approaches For Developing Statistical Classifications?

In the current world where the production of, and the need for statistical information should be faster and more immediate, and where technology now gives more options for collecting and presenting data, statistical classifications are not necessarily keeping up, or taking advantage of those new technologies within the knowledge and information environment. There is a notion that traditional statistical classifications are becoming redundant as knowledge and information capability and associated technology circumvents the traditional methods. Statistical classifications are also often seen as a burden in the statistical processing environment because of their perceived rigidity, dated content and inflexibility. Alternative survey specific versions may be created to circumvent standardisation or comparability because the survey sees itself as special or unique and above the common elements of a standardised statistical classification.

In addition, the early technological advances of the twentieth century that came with the introduction of computers unwittingly constrained classification development due to the limited computer storage capabilities, width of a computer screen to display text, and the need to continue to print to hardcopy coupled with the limitations of the printed page in terms of page width and text string limitations. Hardcopy classifications have become labour intensive to produce, time consuming and resource intense and generally result in the classification being almost redundant by the time it is published. Whilst classifications are now produced electronically, it is this traditional hardcopy mentality that still drives their development. Embodying knowledge into a book does not translate into useful and usable knowledge (Al-hawandeh, 2002). It still has to be codified, processed and communicated in easily understandable formats, something a statistical classification does not necessarily enable.

It is not as though there has never been a desire to develop statistical classifications in a radically new way as they generally perform as designed and enable the collection and dissemination of data that meets the general need of statistical agencies. But the approaches proposed in this paper, along with the case study, are a radical step forward for changing the philosophy of the known, as they have never been really tried before in the statistical environment. There are plenty of examples of taxonomy and ontology development for knowledge management using semantic technologies in the literature (e.g. Kabir, 2013; Fensel, 2011) but this paper is exploring modernisation and the use of semantic methodologies from the perspective of a statistical agency in order to understand the value for data that can come from using a new approach.

4.1 The current situation

Traditional classification structures often do not reflect the fact that the world is an information society where everything needs to be instant and accessible (McNab, 2009). Statistical classifications need to adopt new approaches or they will no longer have a relevant place in enabling the interpretation of data to improve knowledge and understanding of real-world issues.

The failure of statistical classifications in the modern context is generated by the demand for information and data in real-time, and the ability to change and keep up with that dynamic scenario. “Classifications age because reality changes. Classifications need to be revised periodically, but that is a difficult exercise for statisticians.” (Boeda, 2009, p.3).

Relevance is critical to ensuring that data and information is reliable and can meet the needs for policy or decision making. However the constraints imposed by traditional taxonomic approaches and the time-consuming nature that comes with updating and refreshing these classifications, really consigns them to the annals of history as a valuable or even useful tool for managing information and knowledge.

4.2 Industrial Activity Example

For example, in a classification of industrial activities whereby businesses are classified by their predominant activities, the concept of manufacturing is still premised on the idea that you build a factory, employ staff, invest in plant and machinery, source raw goods and produce a product. Then you group and present information using a classification framework based on either manufacturing processes, such as fruit or vegetable processing, or by the products produced, such as bread or stationery.

With the new concepts of outsourcing or even factory-less production of goods, this whole traditional notion of what manufacturing represents no longer applies (Bernard and Foot, 2015). The focus shifts to identifying services and intellectual property (IP) ownership and the ability to understand those concepts and adequately classify them. The information is adequately described or available, it is just that the static nature of a statistical classification doesn't then reflect the reality, thus skewing data as emerging concepts have to be classified to outdated categories. Generally companies that are of a widely diversified nature make it impossible to determine their core business and therefore make any decision for codifying to an industry classification difficult (Hagedoorn and Duyster, 2002).

The most well-known example of this is with Apple Inc. which would traditionally be recognised and classified as a manufacturer of electronics and software, as well as a wholesaler of those products. However the reality is that now Apple Inc. do none of that work. They hold the original intellectual property and may monitor or control the quality aspects of the products that are produced with their brand, but they allow firms in China (or elsewhere) to do the whole process from obtaining raw materials, building factories, hiring staff, making the products and distributing them. This is because modern technology is becoming so complex that R&D partnerships and alliances are required as companies can no longer afford to develop a new product in isolation (Hagedoorn and Duyster, 2002). So are they still a manufacturer or are they a business service, or something completely different? We know Apple Inc. exists, and we know they have Apple products and we can obtain information about them in terms of income and expenditure, number of employees etc. but is the issue in the search and discovery of information, or the fact that any data associated with or about the company doesn't reflect the reality when it gets classified.

What this demonstrates is that the infrequent updating of statistical classifications makes it difficult to try and classify emerging realities into a dated structure and that means that the information is not current and the data is not accurate i.e. it becomes hidden in an arbitrary, best-fit category. This then relies on increased metadata or guidance on how to treat entities such as Apple Inc., but does not necessarily alleviate the search and discovery for users trying to find information about the enterprise, nor enable new collections of data to easily integrate with other data sets. The ability to share and compare information and data is hampered by the division created in having real-world information classified against dated concepts. Reuse of the existing category or concept without changing its scope and coverage means that the ability of researchers to understand what the data really means is hampered. It is important that in making data reusable it is complemented with appropriate metadata information (Thanos, 2017) as often a common vocabulary or terminology is missing.

4.3 Occupation Example

Another example comes from the dynamic nature of the Information and Communication Technology (ICT) sector in terms of measuring and classifying occupational information. A traditional occupational classification may be based upon the concept of skills, and tasks and duties performed, which enables grouping of occupational categories into related groupings such as managers, professionals, technicians, etc. However job titles change so rapidly based on employers wanting to glorify a vacancy to attract a certain style of applicant, or there is the elevation of qualifications and skills to impact on the interview process and enable employers to be picky about what entry level they want, and then there are issues pertaining to skill shortages and immigration which further impact on how occupations are classified.

This constant real-world change coupled with the ever-increasing re-use of traditional job titles to mean other things creates increased frustration for developers and maintainers of classifications, plus for users of the associated data. For example, an architect is no longer exclusively a designer of buildings and structures, they are now more represented in the ICT sector as solutions architects, enterprise architects, software architects and the list goes on. The title of architect needs significantly more information around it to enable accurate classification and representation in the data that informs workforce planning, skill shortages and other policy related initiatives.

5. Challenges for a national statistical office

The difficulties the examples above illustrate for a statistical classification in terms of just the simple ability to update and reflect real-world content is further affected by the ability of the classification developers to acquire the relevant real-time information and knowledge about a concept or category to enable them to consider doing something about it. Is there a definitive source for the information that enables a clear decision to be made about what is needed for the classification, and is there sufficient user demand to actually measure the concept to warrant changing a classification? What invariably happens is that the developer compiles a list of issues for consideration to work through the next time a classification is reviewed but this list may also be dated by the time any work is undertaken.

Further exasperations are encountered if the user is interested in a real-world phenomenon, has data about it and wants to share that information but is limited to a traditional, static classification to describe the information. It is all very well to have the information in XML and/or exchange the information via SDMX processes but the meaning of the category or the scope of the concept being measured is not accurate. Theoretically this may not be seen as an issue because the processes are still doing what they are intended to do, e.g. machine-to-machine. The problem lies with users of the information not being on the same page as their views or perspectives are constrained by what classification or conceptual scope they are using.

The need to change is also underscored by the need for national statistical offices to cut costs and reduce overheads for which traditional classification development or review projects are a significant resource burden. Despite this, there are always numerous road-blocks put forward to maintain a traditional approach to statistical classification developments most often encapsulated in the ideal that 'this is the way it has always been'. Innovation or new ways of doing things invariably bring about aspects of low trust or fear of the unknown coming to the fore (Torben Rick, 2011). This combined with staff needing retraining, changes to outputs systems and processes, and increased publication costs that come with change do not seem to make agencies move from blindly staying with the usual or traditional approach. The overarching reluctance for change is coupled with an acceptance that it is alright to have the outcome of this process as only producing a classification that reflects the point in time it was developed and which often has no relevance to the data when finally implemented into statistical, knowledge or information management systems.

5.1 Organisational management perspective

From a management perspective the challenges faced in terms of meeting that real-time demand for information means a strategic rethink on priorities, and a re-evaluation of resourcing availability which then has to be compared against the cost of finding and using new and better ways to undertake the work. The cost of establishing a project team to do the work that leads to the cost of producing a book is becoming more and more prohibitive. This impacts on the decision-making with regards to whether or not statistical agencies chose to pursue a more dynamic approach to the creating of statistical classifications.

It may, from a management perspective, be easier to do nothing, with the resulting impact that the classifications do not meet the need, or reflect the reality as it is financially better to take that approach and risk the relevancy of the information and stay with the known regardless of how dated the information is. This type of thinking tends to suggest that maintaining a statistical time-series trend is more useful and important for national statistical agencies rather than reflecting the current real-world. But then why can both approaches not be utilised? In addition the number and variety of internal systems coupled with information being stored in one form and not able to be shared or reused across systems is seen as a driver for not doing anything.

As an example to build a classification from scratch, as opposed to revising an existing classification can take a number of years, particularly if the information is deemed, or thought of, in the context of a hardcopy publication. Establishing the need for the classification, in terms of the concepts to be measured or the knowledge or information need, followed by clarifying the scope of the work and allocating and planning sufficient resource requires a lot of time and consultation. The simplicity of identifying an initial structure for a classification, for example identifying the broad groupings that might be used to describe occupations or jobs, let alone the number of levels and the detail information that needs to be classified at the lowest level, can be quite time-consuming. This is all before the actual content and detailed information is gathered, interpreted, structured and a classification built. On top of this is the need to consider the information needs and the outputs required, how the information is to be collected such that the classification can be built, and determining the relevant systems and tools to support the implementation and ongoing maintenance of such a classification. Time series mitigation combined with alignment and integration of administrative with statistical data, whilst able to improve statistical quality still imposes large administrative costs especially when adding or improving definitions and associated statistical measures (Penneck, 2007).

Aligned with this cost is the need to determine the appropriate custodian of the classifications and the rules and processes for future updating and maintenance. In an environment where traditional classifications and hardcopy are produced, management decisions have to be made about whether the cost of change is warranted, whereas consideration of a different approach to determining and managing the information, may have longer term benefits in terms of financial cost reduction and/or employee reduction.

5.2 Benefits of changing approaches

So why modernise a traditional science, theory or approach for producing statistical classifications that has worked relatively well for a very long time? It is generally accepted that a traditional grouping of like categories in a hierarchic structure or flat listing is the most appropriate way of classifying data. This has generally been the way of creating a taxonomy since the days of Aristotle whereby knowledge was very much part of personal experiences and ideas and communicated and shared verbally. Yet this approach is no longer meeting the needs to produce data that enables the modern reality of the world to be described or understood, nor does it specify the concepts of data and knowledge in ways that support the sharing of information (Theocharis and Tsihrintzis, 2016).

The future for modernising statistical classifications is about adopting innovative approaches or recognising advances in metadata modelling whereby statistical classifications are broken down into their component parts i.e. concepts, categories, codes, definitions and coding index entries, rather than focussing on reviewing and developing one size fits all, comprehensive monoliths.

The arrival of the internet, semantic web technology and other associated web and ICT technologies, particularly in the fields of knowledge and information management, gives so many advantages that can now be used to improve how classifications can be developed and applied to data, and to improve search and discovery of metadata and knowledge information. Technology gives the ability to link information using searchable glossaries, instant definition visualisation and improved metadata, particularly via the use of metadata information modelling approaches. The interoperability of systems and the treatment of information as an asset needs to be incorporated into the thinking that enables society to understand the vast pool of information available to it.

With the immediacy of data availability from a multitude of sources such as smartphones, supermarket scanners, ATMs, Global Positioning Systems and other electronic media, it is important for statistical classifications to be more responsive and use terminology and categories that reflect current world usage and understanding. However the immediacy does not come without risk or cost, and just because there is more data doesn't mean the inferences or implications that can be taken from it are sound (Harford, 2014). There is still a need to establish a clear need for collecting and classifying the data in a way that is meaningful and to undertake this in a way that doesn't over-elaborate the content beyond the practical need of and for the users.

A change in approach can use relational databases, innovative classification management systems or other computer created matrix software to produce dynamic electronic formats, compared to the old hardcopy using traditional editing and publishing processes. For example understanding the use of Resource Description Frameworks (RDFs) to mix structured and semi-structured data across different applications gives a different insight into how statistical classifications may be created. It also opens up the use of ontologies and taxonomies to be better retrieve information using web services and query tools. Familial relationships of content can now supersede the traditional parent-child approach thus producing more options to structure knowledge, information and metadata within a classification and its associated outputs.

The thinking that underpins statistical metadata or information models give more flexibility for considering strategies to modernise the way in which statistical classifications evolve. The common need is to really ensure that everyone is talking about the same concepts, categories and content in the same way, something not always the case when utilising a traditional statistical classification framework.

“Creativity and innovation at work are the process, outcomes, and products of attempts to develop and introduce new and improved ways of doing things. The creativity stage of this process refers to idea generation, and innovation refers to the subsequent stage of implementing ideas toward better procedures, practices or products. Creativity and innovation can occur at the level of the individual, work team, organization, or at more than one of these levels combined but will invariably result in identifiable benefits at one or more of these level of analysis.” (Anderson, Potocnik and Zhou 2014)

6. Introducing A Vision For Change

The overarching benefits for knowledge and information management that come with moving away from a traditional approach provides numerous opportunities for managing the data more efficiently.

Data is nothing without the metadata that supports it. Whilst metadata is usually described as information about information it is fundamentally characterised as either structural, reference, descriptive or administrative metadata (Riley, 2017). This covers general metadata concepts for producing context, terms for describing statistical methodology, terms and data, and terms specifically for data exchange or sharing. But the metadata is primarily intended to create a common understanding of the meaning of data, or the semantics of it to enable the correct and proper use and interpretation of data by all users.

Metadata standards not only chunk down entities and objects, and define relationships between them, but also require definitions and an understanding of concepts and classifications. Taking a metadata based approach means that each attribute within the statistical classification, knowledge or information system are better identified and implemented.

This gives a new insight into how knowledge and information can be managed in a more flexible and relevant environment which allows users to share and contribute content to meet real-time needs. How to chunk classification content to enable a richer outcome is driven by thinking about how structured data and metadata can be better incorporated into a classification framework when investigating the way knowledge and information models are described.

It is no longer just about codification of information into a single knowledge structure, it is also about enabling sharing of content in a consistent way, and taking a conceptual relationship approach to the content. To truly support knowledge management processes a different classification, ontological or taxonomical approach is required.

The new approach proposed for modernising statistical classifications through this document for codifying information is focussed on the ability to create multiple relationships between concepts with the critical decision point being whether to treat everything as a concept, or only treat some elements as concepts and others as categorisations of those concepts. This impacts the thinking on how the mapping process occurs within the system or framework developed, and may also impact the ability to reuse taxonomies in multiple environments, for example in a data management system, document repository, records management system or other environments that are not specifically data or metadata focused.

More time may need to be spent on understanding and creating those relationships. It is not just about a one to one relationship but potentially a one to many mapping and establishing all relationships between concepts may not initially be easy. But there is scope for machine-learning and automation to occur over time to reduce any manual requirements for mapping. All of this leads to better knowledge management in a consistent way and facilitates search and discovery for all users. The information then has greater value and is enriched.

6.1 Use of new technical solutions and methodologies

A new technical solution/approach means better use of service oriented architecture (SOA) which allows integration with other system components or platforms (both internal and external), particularly when moving to a cloud environment. Use of the Simple Knowledge Organisation System (SKOS) and integration of extensions based on ISO/IEC 11179 or even the Generic Statistical Information Model (GSIM) give wider opportunity to do more things with classification components, structures and views than ever before.

A critical driver for moving into a new methodology and thinking around how statistical classifications would work in knowledge and information management systems was the visualisation that comes with utilising the thinking behind metadata conceptual models for international metadata standards such as Statistical Data and Metadata Exchange (SDMX) or Data Documentation Initiative (DDI). These metadata standards are supported by information models which specify concepts, relationships, rules and other elements and which show mappings of either real-world or abstract entities.

The biggest advantage moving forward is the utilisation of RDF/XML technologies to provide flexibility in terms of managing assorted and associated properties that previously could not be addressed in a traditional classification hierarchy approach. In using XML there is scope for strong typing of information and the ability to transform or reconfigure information and content into formats such as SDMX or DDI making sharing and exchange of classification metadata more accessible and easy.

Moving to a concept driven approach gives a clear element separation providing a robust model to ensure information consistency in terms of content and presentation. Combined with support for simple and complex search and mining capabilities there is significant scope to maximise metadata element reusability, particularly when using the best features of existing metadata standards and semantic web technologies.

Making use of methodologies such as SKOS mitigates the need to strive for perfection within a traditional approach as the functionality that comes from a better use of taxonomies, thesauri and concept management gives more options to classifications developers and information managers. For example, understanding the use of Resource Definition Frameworks (RDF) to mix structured and semi-structured data across different applications gives a different insight for how statistical classifications may be created and used.

6.2 The philosophy of the vision for change

The vision is about introducing a philosophy of concept based classification management that can add value to data by increasing the content and metadata that can be created and stored with a classification category, accompanied by greater integration of administrative and statistical concepts and data. The innovation and value added to the data and for the users also comes from a standardisation process which no longer focusses on standardising to a single classification but which encourages reuse of existing content by storing once and sharing across multiple locations.

However the vision for change requires a culture shift in thinking from parent-child hierarchies to relationships and linkages of information and metadata. The focus needs to shift from a need for a single gold-standard to a classification framework that is of a necessary standard for its intended use (ISO 9000). A matrix type approach may become more feasible than a traditional hierarchic structure as it is about the relationships that are formed by linking multiple categories together, as opposed to just creating a relationship between two sequential entities.

Other drivers for the vision to change meant looking at a user friendly browser based system which was easier to search, operate and access, which provided greater communication with users of classifications, real-time creation of content and introducing flexibility. The critical key in the approach is to ensure that there is semantic consistency across measurements of the entities being classified and which enables the delivery of intelligent user interfaces and systems, and in particular which uses industry standard service driven architecture.

Flexibility enables a more relaxed approach to get fit-for-purpose views of concepts rather than trying to accommodate every possibility into a single classification structure. Educating developers and users to think in this space will be a critical challenge to overcome if there is to be a successful modernisation of statistical classifications and changes within the knowledge and information management systems they support.

Familial relationships of content can now supersede the traditional parent-child approach thus providing more options to structure knowledge information and metadata within a classification. This can be coupled with increased metadata and alternatives to labelling, outputs, language options and definition text that are not an option when applying existing best practice theory and processes to classification building.

The vision is underpinned by looking at the whole process of classification best practice, theory and methodology and assimilating it with the logic and clarity that comes from the conceptual models, glossaries and definitions that support the international metadata standards. A greater consistency and standardisation becomes more achievable if the new way of thinking about classification development is employed.

As a web based solution the philosophy can be supported by the use of an application that is accessible via a cloud environment and therefore accessible from anywhere, which can enable a rich client application delivery of a user friendly and flexible interface, so the technological thinking becomes the driver for change but not the physical technology itself. Ultimately the change in solution and approach will enable potential integration of global classification registries. This will enable greater knowledge and information sharing.

7. Why Use Tools Like Skos Or Rdf To Modernise Traditional Approaches?

In analysing the use of tools to modernise approaches to statistical classifications, and in presenting the case study that follows in Chapter 7 of this article, the perspective is from the viewpoint of a producer of classification management systems, and practitioner of statistical classifications. This is particularly in terms of whether semantic web processes, and or conceptual frameworks are the better long term solution for developing and administering statistical classifications.

The ability to provide multiple output views, different labelling options for categories and linking of multiple concepts to build a structure do not exist with traditional statistical classifications. This is due to the constraints and limitations imposed by the use of computer systems and the printed page, and the need to publish hardcopy. These constraints are also affected by traditional classifications using only one concept with one strictly defined definition which prevents classification structures from being flexible, adaptable or truly using the notion of broader concepts and narrower concepts. The broad levels (usually at the top of the structure) are delineated into more detail by the need to further define the groupings that sit below until one gets to the lowest level required. It is not usually about applying concepts at each level of a hierarchy and then refining those to enable the categories to be built. Using tools and methodologies such as SKOS and RDF provide options that hardcopy, strictly hierarchical frameworks cannot for applying concepts at each level.

The biggest advantage for modernising statistical classifications in knowledge and information management systems comes from taking a more SKOS/RDF aligned thinking around concepts by starting with a clearly defined and described concept(s) which enables the building of category sets that simply contain labels or descriptions around which metadata can now be placed. What this means is that hierarchies become layers of flat classifications which can be scoped and defined in terms of related, or broader or narrower concepts. This radically changes the thinking around how a classification can be built and structured. It also highlights that taking a traditional hierarchic thinking is actually inflexible and too narrow limiting the ability to record and store knowledge, information and data.

7.1 Simple Knowledge Organization System (SKOS)

As the Simple Knowledge Organization System (SKOS) is a data-sharing standard that was developed from standards in the Semantic Web that prescribed formal logic and structure. It provides a bridge between knowledge, information and metadata systems that require structured and organized frameworks.

The use of the Simple Knowledge Organization System (SKOS) provides a building block around which a new approach can be developed. As SKOS is a formal language and schema which has been designed to represent structured information such as classifications and taxonomies, it seemed a logical step to utilise it to modernise the way statistical classifications can be developed and then used in more productive and meaningful ways. In addition SKOS enables the usage of narrower and broader concepts in a way that traditional statistical classifications are not readily able to do.

In SKOS, concepts can have many kinds or multiple relationships other than the traditional parent-child that comes with most statistical classifications. As stated earlier the narrower to broader aggregation approach of a traditional classification system is more about refining the groupings and relabelling those groupings accordingly rather than applying a refinement of concepts. SKOS also treat concepts as units of thought which could be abstract or finite ideas or objects which can be independent of the terms used to label them, thus giving significant flexibility to how content is described and presented.

Use of unique resource indicators (URIs) changes the way in which content can be labelled, used and discovered removing the constraint of single descriptors or mutually exclusive labels that are the hallmark of statistical classifications that come from trying to continue a hardcopy publication mentality to how classifications are shaped and formed. In addition the functionality to develop conceptual and other content relationships is introduced in ways that currently do not exist and when combined with other options such as the use of synonyms or aliases for categories, additional flexibility and power for describing and presenting information is obtained.

Adopting a SKOS approach makes for more granular metadata and easier integration, and enables the definition of relationships between concepts and the sharing of those concepts across different classifications or views. This removes the time-consuming and costly overheads that come with then having to create mappings or concordances between classifications to understand how data is shared across a concept.

7.2 Resource Description Framework (RDF)

The Resource Description Framework (RDF) is generally used as a language for representing resource information on the Internet, or rather content or information that can be identified on the internet. RDF is primarily a tool or format for systems or applications to share information, and not really a human discoverable or readable identifier.

As RDF uses unique web identifiers, usually noted as Uniform Resource Identifiers (URIs) for describing resources or entities, which changes the way in which statistical classifications are structured and stored, a quantum leap forward occurs in the sharing of an identical concept or category across multiple classifications. This replaces the traditional approach of single stand-alone classifications, linked by concordances and which are not able to share the same metadata and content.

Additionally the use of the RDF triple, which 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) facilitates breaking up classification content into component parts to enable easier integration and sharing or processing with other systems or frameworks.

8. Case Study: Statistics New Zealand's Classification Management System

In 2010 Statistics New Zealand commenced a ten year programme of work to transform the way in which it produces statistics, and to create a statistical system for the future. This provided the opportunity to really rethink the way in which statistical classifications could be developed and maintained. The traditional approaches of producing hardcopy, sequentially numbered structures stored in computer repositories, did not really support the future vision for the organisation, nor enable an approach to address the increasing requirements for metadata. As a practitioner of developing statistical classifications for use in the collection and dissemination of data and official statistics, the time was appropriate to investigate the viability of alternative solutions and provide leadership to government agencies, business and other organisations to ensure that information, and data, was described and managed in a consistent way.

Some of the requirements for moving to a new classification management system were seen as:

- The need to mitigate a legacy system;
- The need to move from a classification repository system to a full classification management system;
- The need to reduce proliferation of like classifications and versions; and

- A desire to introduce a new approach to the management, storage and dissemination of classification related attributes and entities;
- The need for improved mechanisms for searching and discovering statistical concepts, definitions, and categories to assist in knowledge improvement and management.

Metadata models focus on the relationships between entities and ensuring that there is good description of those entities. For statistical classifications, this meant applying a similar approach by breaking the classifications into their component parts to give a greater emphasis on the central concepts and their definitions supported by categories with their definitions and associated metadata. This philosophy underpins the vision for the new classification management strategy.

It is intended that the system sits in a cloud environment with access to all users of the statistical data ecosystem in New Zealand. This will allow more standardisation and reduces proliferation of like classifications as users will be required to use existing content unless there is good reason not to. In addition, the approach uses the best features of existing metadata models and standards to build a flexible and extensible system which can easily accommodate future system redesign. The ability to take a modular approach to development and maintenance of a system has provided opportunities to realise cost savings and efficiencies for all producers of official statistics.

8.1 Advantages for a new system

Advantages include greater relationships between entities, leading to more efficiency and automation in the development, authorisation and dissemination process, combined with improved search and discovery options. The key advantage is that information is stored separately in its component parts, but only stored once, and then used in multiple locations.

The core breakdown of the underlying model is based around the following components:

Core: This identifies, versions and describes contexts within which classification or views of categories are used.
Conceptual: Where concepts and their uses are modelled
Classification: A general model which provides a generic approach to classifications with options for more specific extensions to deal with formal structures or derived classifications.
Coding: Enables interaction and integration coding systems by storing items such as synonyms (often survey responses or classification categories) used in coding statistical surveys.
Concordances: Focuses on all the relationships between concepts and entities, enabling system generated as well as user generated concordances

The outcome is a true classification management system and not just a repository for classifications, concordances and associated classification metadata. As a concept based system it has increased information and metadata about the relationships and properties of classifications and their component parts. System flexibility allows it to be extensible to allow for future system developments or changes without the need for large-scale system review and modification. Classifications become views of category sets rather than single entity structures thus giving greater option for output and design.

The system uses semantic web functionality which allows for common formats for the integration and combination of data drawn from multiple and varied sources. It also utilises the Simple Knowledge Organisation System (SKOS) as a means of supporting the classification schemes and taxonomies that can be used in a semantic web environment. Using the Resource Description Framework (RDF) information is passed between computer applications in an interoperable way. This opens up the system to integration with administrative systems often using an access layer or dashboard rather than system to system or machine to machine access.

The biggest advantage for classification development is the significantly richer quality of information that can be stored and held within a classification due to the relationships between concepts and categories. Beginning with agreed concepts and definitions enables the creation of many related and derived views of content that previously would not have been possible. Aspects of traditional classification theory still remain such as mutual exclusivity, agreed concepts, but other aspects of rigid code patterns, hierarchical structures and traditional constraints for publication are removed.

Objects contained within the new model still retain much of the traditional elements such as levels, versions, codes and categories. What changes is the way they are related to each other, and the information that is stored with them. Greater emphasis is now placed on defining objects and ensuring there is a clear statement on the scope and content of an object.

From a research perspective point in time research is more possible due to the maintenance strategies around content and the dynamic ability to reflect real world requirements in a timelier manner. Time-stamping every entity will show when it was in use or valid. This allows for research on a concept and/or categories to be undertaken on any time pattern, as opposed to the existing practice of the life cycle of a classification. From a user perspective there is more flexibility in reusing existing approved content and utilising agreed concepts thus removing the need to reinvent or duplicate content which is often required by existing technology and classification management systems.

8.2 Impact on users

A key feature of the new methodology and philosophy is the ability for users to make greater decisions around when to adopt and use content. Users will no longer have a cyclical review process based on a five or ten year cycle inhibiting the adoption of real-world issues. Information is updated dynamically and time-stamped to indicate when it was added to the system. Users are notified of any changes and can decide if and when they wish to adopt the change thus reducing implementation planning and associated costs. Adopting changes rather than being forced into having to plan for a block change or replacement of a classification is the new norm. The system with its auditing functionality provides the ability to better monitor uptake and usage of concepts and their associated views. This enables better maintenance and standardisation information to be available to system administrators or content custodians.

Moving from a repository system into a management system gives greater search and discovery options, and facilitates interlinking of concepts and content. Previously this was not possible. Users can 'plug and play' with content to suit their own needs whilst ensuring they are using consistent, and approved content. Any content created in the system is available for other users to adopt and use.

Figure 1 shows the general classification model that underpins the system and the core relationships for entities.

The model enables better versioning of classification structures, allows the creation of category sets, which are effectively master lists of all categories applicable to a concept from which views can be created (either standard or customised) and that the ability to link and relate attributes is much easier to achieve. Classifications are no longer hierarchic, static monoliths limited by codes, levels and structures, they become flexible views which can be reused, integrated and related across multiple concepts.

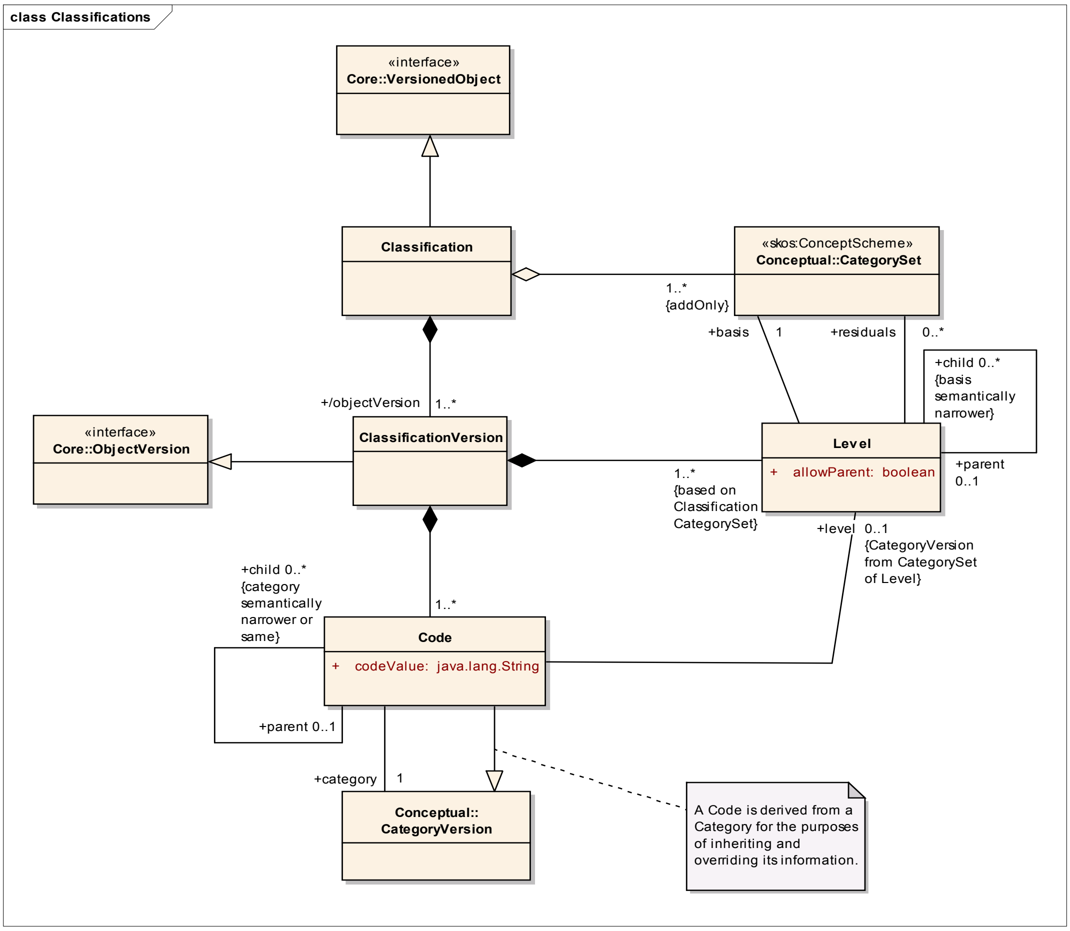

Figure 1: General Classification Model

The concept model (as illustrated in Figure 2) provides the functionality to store concepts with agreed definitions and sources for use in building and relating classifications. It can still allow for the traditional parent-child relationships for classification building if needed but the focus completely moves to defining concepts that can be configured in a flexible manner to increase the richness of information and metadata for users. Relationships move to matrices as opposed to hierarchies i.e. the ability to go sideways and across compared to the traditional up-down approach.

Concepts can be treated as base concepts or narrower versions depending on context therefore allowing a greater understanding of the exact relationships between different data sets, structures and classifications. The flexibility this brings to the development of statistical classifications and to the integration and analysis of statistical data is a significant benefit from making this shift away for hierarchic, sequential classification structures.

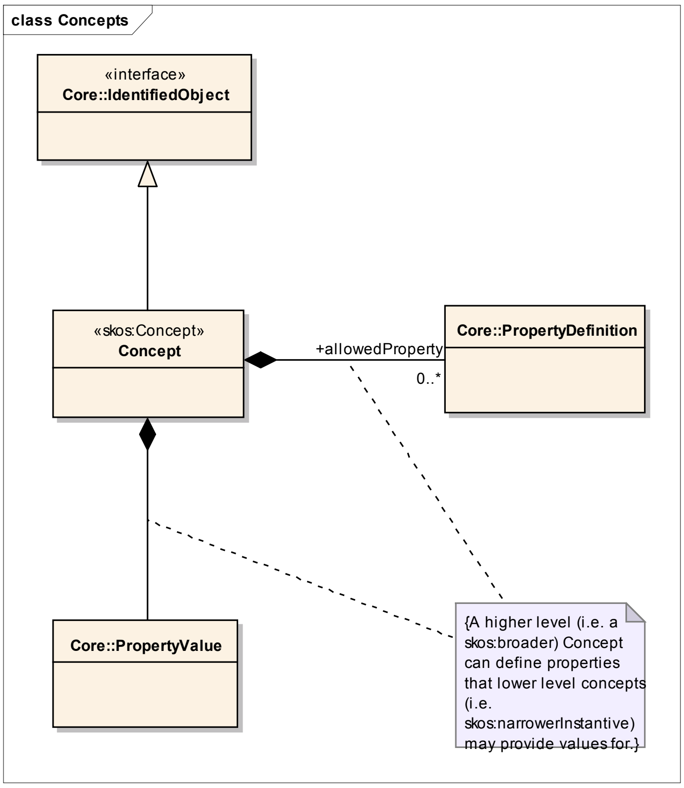

Figure 2: Concept Model

9. Conclusion

Traditional statistical classifications have not yet reached a point of extinction but this paper has tried to demonstrate that demand for instant data and information is better enabled through innovative and newer ways of classifying information. Traditional hierarchies and parent-child relationships limit flexibility, and reduce the ability to better interpret and understand the data. Concepts and categories, relationship matrices and more user defined views of content need to be the norm. Thinking using the principles of SKOS or RDF frameworks is the most viable way to modernise statistical classifications and how they are used in knowledge or information management systems.

It will take time to replace the traditional approaches to statistical classification development and theory with innovative, technological solutions. From a system perspective, much of the development process can now be automated, content can be contributed and approved online, and computer programs are advanced enough to take account of more human thinking methodologies. The challenge is in shifting the human thought processes that are ingrained through our desire to group items together in traditional formats. Educating our brains to accept the chunking down of component parts will take time but the long term benefits for statistical agencies are significant - cost reduction, reduction in staff numbers and more stakeholder engagement supported by more timely and consistent official statistics.

The case study sought to demonstrate what a concept based classification management framework or system might look like but is development is intended to provide a simple, and easily usable tool for search and discovery for most data users. The content is easily updated, mapped and provides a consistent way of managing and describing concepts that can enrich the metadata for official statistics, and which simplifies knowledge management.

Familial conceptual relationships, or visual ontologies providing a relationship between concepts and categories that enhances the richness of the available information, increases understanding of data and society, and will provide flexibility in how statistical information and knowledge can be managed.

9.1 Future direction

The challenges for the future in adopting this approach is changing the mind set of data custodians and knowledge managers to think differently about how information should be structured and how data should be described such that hierarchic classifications are not seen as the only solution for classifying data.

Future work will involve moving from the establishment of relationships between concepts, to then investigating the scope of those concepts to enable the categories or inclusions that define the concept to be linked and related through the use of reusable category sets. This will enable a situation of master category sets from which user defined views can be created thus allowing the flexibility the users want but retaining the standardisation through consistent use of concepts and definitions.

Allowing user-driven content authoring and management, controlled by agreed and strict business rules and processes will in the long term reduce the need for statistical classifications as traditionally understood, and enable statistical agencies to cut costs in resources and capability without compromising their knowledge or information management practices.

References

Al-Hawamdeh, S., 2002. Knowledge management; re-thinking information management and facing the challenge of managing tacit knowledge. Information research, 8(1), Paper No.143.

Anderson, N., Potocnik, K., and Zhou, J., 2014. Innovation and Creativity in Organizations: A State of Science Review, Prospective Commentary, and Guiding Framework. Journal of Management 2014: 40

Bernard, A.B., and Fort, T.C., 2015. Measuring the Multinational Economy: Factoryless Goods Producing Firms. American Economic Review: Papers and Proceedings, 105(5): pp. 518-523

Boeda, M., 2009. The how and why of statistical classifications. Courier des statistiques, English series no.15

Dumay, J.C., 2014. A critical reflective discourse of an interventionist research project, Qualitative Research in Accounting & Management, Vol 7 Iss: 1.

Fensel, D. ed., 2011. Foundation for the Web of information and Services, Hamburg, Springer

Gavrilova, T., Lescheva, I., Strakhovich, E., 2015. Gestalt principles of creating learning business ontologies for knowledge codification. Knowledge Management Research and Practice 13, pp. 418-428

Geipel, M., 2015. Statistical Translation of Hierarchical Classifications from Dewey Decimal Classification to the Regensburger Verbundklassifikation. D-Lib Magazine Vol 21 Number 5/6

Hagedoorn, J., and Duysters, G., 2002. External sources of innovative capabilities: the preferences for strategic alliances or mergers or acquisitions. Journal of Management Studies 39, 2, pp. 167-188

Hancock, A., and Gregory, A., 2013. Classification Management System Development for Statistics New Zealand, Working Paper 18. Work Session on Statistical Metadata. Geneva, Switzerland 6-8 May 2013. Geneva: United Nations

Harford, T., 2014. Big data: Are we making a big mistake. Significance Vol 11, Issue 5

International Organisation for Standardisation, 2004. ISO/IEC FDIS 11179-1 Information technology - Metadata registries - Part 1: Framework. Geneva: ISO.

Kabir. N., 2013. Tacit knowledge, its Codification and technological advancement. The Electronic Journal of Knowledge Management Vol 11, (3), pp. 235-243

Lee, Y-C., and Lee, S-K., 2007. Capabilities, Processes and Performance of Knowledge Management: A Structural Approach. Human Factors and Ergonomics in Manufacturing, Vol 17 pp. 21-41

Lytvyn, V., Vysitska, V., Veres, O., Rishnyak, I., and Rishnyak, H., 2017. Classification Methods of Text Documents using Ontology Based Approach. In: N. Shakhovska, ed. Advances in Intelligent Systems and Computing 512. Ukraine: Springer 2017 pp. 229-240

McNab, C., 2009). What social media offers to health professionals and citizens? Bulletin of the World Health Organisation 2009:87. Geneva: WHO

OECD, Glossary of Statistical Terms. [online] Available at: <http://stats.oecd.org/glossary/detail.asp?ID=6878>

Penneck, S., 2007. Using Administrative Data for Statistical Purposes. Paper presented at the ICES-III, June 18-21, 2007. Montreal: Canada

Rick, T., 2011. 12 reasons why people resist change. [online] Available at: https://www.torbenrick.eu/t/r/hwj

Riley, J., 2017. Understanding Metadata: What is Metadata, and What is it for?. Baltimore: NISO

Roos, M., 2010. Using XBRL in a Statistical Context. The Case of the Dutch Taxonomy Project. Journal of Official Statistics, Vol 26, No.3, 2010 pp. 559-575

Salkind, N. J., 2004. An Introduction to Theories of Human Development. Thousand Oaks, CA: Sage.

Shaw, M.J., Subramaniam, C., Gek Woo Tan., Welge, M.E., 2001, Knowledge management and data mining for marketing. Decision Support Systems 31 pp. 127-137

Staab, S., Schnurr, H-P., Studer, R., Sure, Y., 2001. Knowledge Processes and Ontologies. IEEE Intelligent Systems, Vol 16, No 1, January 2001 pp. 26-34

Thanos, C., 2017. Research Data Reusability: Conceptual Foundations, Barriers and Enabling Technologies. Publications 5 (1), 2

Theocharis, S., and Tsihrintzis, G., 2016. Knowledge Management Systems in the Public Sector: Critical Issues. Lecture Notes on Software Engineering, Vol 4, No 1

United Nations Economic Commission for Europe, 2015. Generic Statistical Information Model (GSIM) retrieved from http://www1.unece.org/stat/platform/display/gsim/Generic+Statistical+Information+Model

United Nations Economic Commission for Europe (2014), Database of International Statistical Activities, retrieved from http://www1.unece.org/stat/platform/display/disa2014/Database+of+International+Statistical+Activities+%28DISA%29+2014

United Nations Statistics Division, 2013. Best Practice Guidelines for Developing International Statistical Classifications, retrieved from http://unstats.un.org/unsd/class/intercop/expertgroup/2013/AC267-5.PDF

United Nations Statistics Division, 2015. Statistics at a Glance retrieved from http://unstats.un.org/unsd/statcom/brochures/unsd_brochure.pdf

World Health Organisation, 2010. International Statistical Classification of Diseases and Related Health Problems 10th Revision, retrieved from http://apps.who.int/classifications/icd10/browse/2010/en