Modernisation of Statistical Classifications

Andrew Hancock

Principal Analyst, Statistics NZ

Chair, UN Committee of Experts on International Statistical Classifications

Background

- International statistics have overlaps in concepts, definitions, classifications and metadata
- Limited integration of the many standards, manuals and frameworks hampers responsiveness to emerging user demands
- It is difficult to easily cross-reference content, and/or search and discover content
- There is rapidly occurring real-world change which is not easily incorporated

Stats NZ
Tatauranga Aotearoa

Purpose

- Advance the use of innovative technologies and approaches for cross-referencing and navigating between the various international statistical standards, manuals and classifications

- Implement new methodologies for managing and describing data, and the categories to which they are classified through greater uptake of semantic web technology

- Allow digital integration with well-established library and other vocabularies, taxonomies and ontologies, to improve cross-disciplinary search capabilities of digitized documents

- Reduce cost, resource and time for undertaking revisions of international classifications and standards

Why have international classifications

- There is a need for standard concepts, definitions and classifications to ensure a consistent approach to classifying statistical data to support global policy initiatives such as the Sustainable Development Goals (SDGs), climate change or the digital economy etc

- They provide a simplification of the real world and a framework for collecting, organising and analysing data, both statistical and administrative, and are the cornerstone of official statistics

- They provide a framework for international comparability and a basis for national development

- They can be used for:
  - Collecting and organising statistical information in a standard way
  - Aggregating and disaggregating datasets in meaningful way for complex analysis
  - Supporting policy and decision making
  - Assisting developing countries and their Official Statistics programs

Issues with using International Classifications

- Understanding the need for an international classification and a lack of encouragement or support by international agencies and NSOs to adopt them in a timely manner
- Obtaining international consensus and input into their development and maintenance, and the length of time taken to develop, revise, maintain and implement them
- Lack of a central agreed global repository for them, and the different formats they are stored in (e.g. html, .pdf, hardcopy, MS Excel) thus limiting integration, sharing, search and discovery, and dissemination
- Traditional approaches of using sequential codes, parent-child category relationships, single category labels, constrained by A4 page or computer screen width, and output table publishing needs
- Inability to identify what is the ‘official’ international standard to be used when there may be a proliferation of like standards e.g. SDMX, DDI, ISO 11179, Dublin Core
- Cyclical review processes based for human consumption, not machine consumption

Stats NZ
Tatauranga Aotearoa

Traditional Approaches

- Hide conceptual relationships, create structural silos, and do not easily address cross-cutting issues
- There is no scope for multiple contexts, or flexibility in approach, no ability to use multiple concepts or entities, and no ability to create aggregated or derived linked views
- Revisions, whether incremental adjustment or larger, are costly and time-consuming for national statistical offices
- Difficulties in timing change for IT systems and platforms to facilitate implementation
- Everything is stand-alone and needs to be mapped using correspondences/concordances
- Limited use of current ontological/taxonomical thinking as still driven by statistical processes and needs

Practice Change Considerations

- Still heavily in a Eurocentric/Western model – mutual exclusivity (a response can only be classified to one category), statistical balance (categories are of similar/equal counts, population is evenly spread) etc
- Entrenched IT systems with hard-coded content, reluctance to change and the fear of innovation
- Impact of digitalisation and introducing digital metadata in machine-readable format e.g. an SDMX api v stand-alone classifications not currently being adopted
- Maintaining time-series and consistency is important but users are struggling to understand that the data means in the contemporary context
- Much of the revised content is already out of date on publication
- Educating users to the idea of change and the benefits that come with it but introducing innovative change in a transitional way

Stats NZ
Tatauranga Aotearoa

Future Thinking/Direction

- Data is now collected from sources that did not exist 10-20 years ago

- There is greater volume and variety in the data and the standards are not keeping pace (the social media world impact)

- It is time to explore the wider use of:
  - relational databases
  - computer created matrix software
  - advances in ontological engineering
  - semantic web technologies
  - more efficient and automated authorisation and dissemination processes

- Concept based classification management systems are the way forward, and can encourage reuse of existing content and reduce duplication or inconsistencies

Concept Based Classification Management

- The vision is about user-driven, dynamic content - doing stuff in real time
- Adds value to data by increasing content and metadata that can be created - expands the data narrative
- Enables greater integration of administrative and statistical concepts
- Introduces semantic consistency across standards
- Encourages greater reuse of existing content by storing once and sharing across multiple locations
- Removes cyclical, labour intensive, time and resource revisions of standards
- Enable more usage of apis and conceptual/metadata modelling

New methodologies

- ❖ Requires better usage of service-oriented architecture (SOA)
- ❖ Enables uptake of the Simple Knowledge Organisation System (SKOS) and XKOS
- ❖ Allows integration with metadata standards - e.g. SDMX or ISO 11179
- ❖ Provides greater usage of taxonomies, thesauri, ontological engineering and concept management to mix structured and semi-structured data
- ❖ Encourages multiple output views, different labelling options, and multi-lingual content linked to an approved concept
- ❖ .xml based and more automation in processes
- ❖ Educating and changing the international thinking – a slow process which is starting to take hold

Metadata Modelling

- Begins with a clearly defined concept which may have relationships to any number of other concepts or sub-concepts
- Each concept is unique and forms a scope for all entities or words that may be categorised by that concept
- Uses intensional and/or extensional approaches to organising knowledge: -
  - Intensional – a concept is listed with properties or categories that the concept must have to be part of set captured by the definition (e.g. concept of country which is defined as independent, a geographic entity, or administrative region)
  - Extensional – a concept is defined by listing or specifying everything that falls within scope of the definition (e.g. concept of country which lists all known countries of the world)
- Enables everyone to talk about the same concept, category and content in the same way
- Makes search and discovery, retrieval and interoperability easier

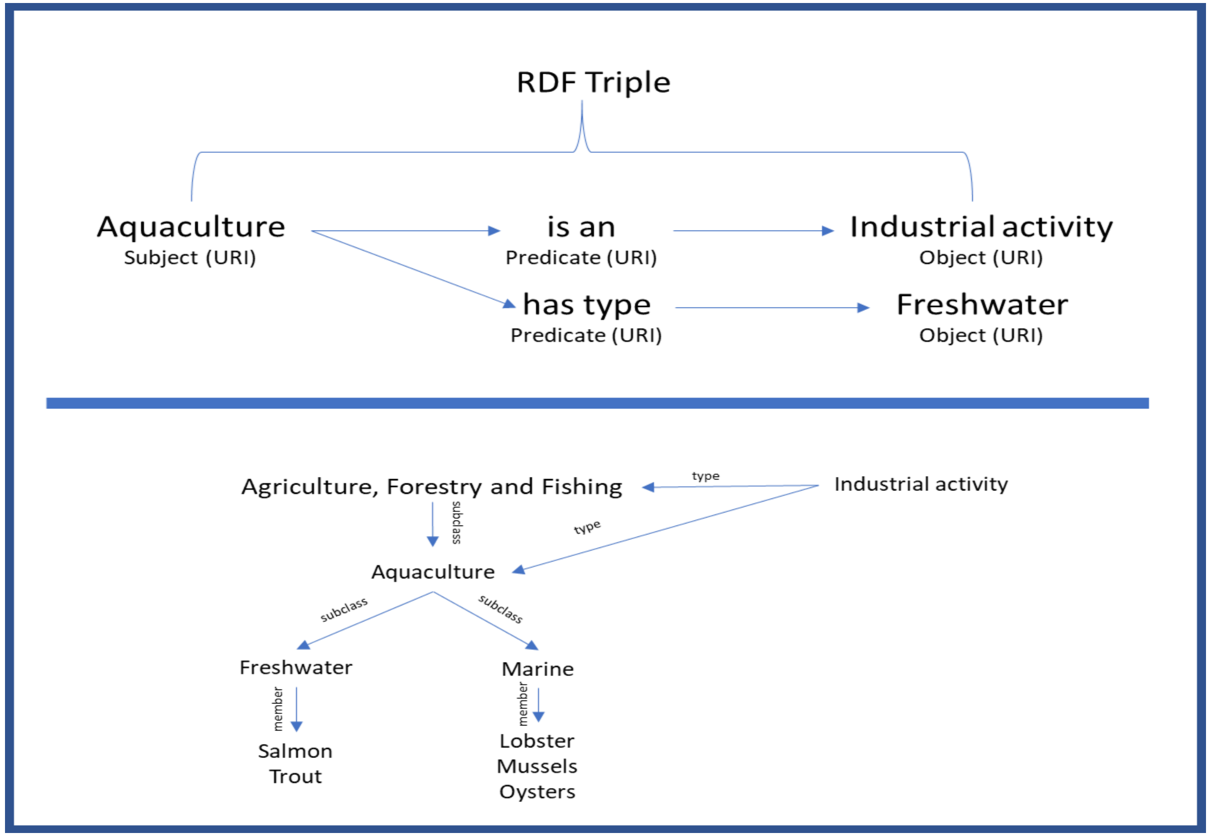
Simple Knowledge Organisation System (SKOS)

- Concepts can have multiple relationships like a neural network model
- Uses unique resource indicators (URIs) linked with lexical strings, assigning notations and links to other concepts
- Can organise content into informal hierarchies and networks using defined concept schemes
- URIs remove constraint of single descriptors or mutually exclusive labels
- Uses synonyms or aliases for categories
- Provides more granular metadata, easier integration and sharing of concepts and content

Resource Description Framework (RDF)

- Uses unique web identifiers for describing resources or entities
- Uses the RDF triple which comprises a subject (web resource), predicate (property) and an object (value)
- Allows classification content to be disassembled into component parts for easier integration and sharing
- Enables reconfiguration or repackaging into traditional content or user defined views
- Utilises graph networks or query language (such as SPARQL) to retrieve and manipulate data
- Enables faster and more dynamic updating

RDF Example - Aquaculture



Concept Based Classification Model

- Concepts are the building blocks for everything
- Each concept has
  - a label and agreed definition
  - relationships to other concepts allowing for conceptual frameworks to be created and making for an easier way to merge and transfer data
  - categories within the concept that are linked to other concepts like in an electronic thesauri or neural network
- There is no scope for multiple contexts, or flexibility in approach, no ability to use multiple concepts or entities, and no ability to create aggregated or derived linked views
- Categories are stored in a category set (similar to SDMX codelists)
- Content can be dynamically updated or added by approved users

Stats NZ
Tatauranga Aotearoa

Concept View - Industry

Related Concepts
Agriculture
Construction
Energy
Manufacturing
Products
Trade

Industry

Related Concepts
CPI
GDP
Households
Non-Profit
SNA
Unpaid Work

Definition: An industry consists of a group of establishments engaged on the same, or similar, kinds of production activity.

Category Set
Aquaculture
Electricity Generation
Hardware Stores
Libraries
Management Services
Poultry Farming
Research Institutes

GDP by Industry
Agriculture
Construction
Education and training
Heath
Retail trade
Transport, postal and warehousing
Wholesale trade

Concepts are combined with code banks to create category sets that are reused across multiple views

Code Bank	Id
A1111	5AN
AA	2A
1101	4N
112	3N

ISIC	
A	Agriculture, Forestry and Fishing
A03	Fishing and aquaculture
A031	Fishing
A0311	Marine fishing
A0312	Freshwater fishing
A032	Aquaculture
A0321	Marine aquaculture
A0322	Freshwater aquaculture

NAICS	
11	Agriculture, Forestry, Fishing and Hunting
112	Animal Production and Aquaculture
1125	Aquaculture
11251	Aquaculture
112511	Finfish Farming and Fish Hatcheries
112512	Shellfish Farming
112599	Other Aquaculture

Tourism
Camping Grounds
Restaurants
Museums
Recreation and Sport
Travel Agents
Packaged Tours
Bungy Jumping

Benefits

- All content is time-stamped and each entity has an unique uri
- APIs are in place to enable integration of systems and to link content
- Uses a customisable lifecycle and approval process
- Content can be disseminated in multiple formats e.g. Word, Excel, SDMX, DDI, .pdf
- Standards, classifications and correspondences are all linked together
- AI/Machine learning will be added to reduce human interaction in the future

Conclusion

- Traditional methods of management no longer work
- Semantic web/metadata modelling is the best way forward
- Governance models can be changed and modernised
- Dynamic and real-time change can be implemented
- Over-arching cost reduction introduced
- Greater consistency in content achieved
- More automation and use of apis/AI/machine learning used