

ANZ Metadata Working Group

Meeting 3 Report

21-22 February 2019, Melbourne

**Background:**

The third MDWG meeting was held in Melbourne on 21-22 February 2019.

The outcomes for meeting were for all members to be better informed about activities and practices; confirm current priorities and activities; to endorse recommended workshop metadata elements; actions since the last meeting and to progress the best practice blueprint.

**Meeting summary:**

Craig Sandy, Surveyor-General of Victoria and ICSM Executive Member, welcomed the MDWG to Melbourne, and noted that the work being undertaken by this group is ‘invisible’. If we are doing our job well, no one notices, it is only when there is no metadata that datasets won’t work or integrate. The more we perform, the less everyone knows about our work. Craig said this group is essential to many other aspects of ICSM, and should also be the focus of all the ICSM groups. Machine to machine will benefit the broader community.

As we move to ATRF, there will be an even stronger driver for working groups to connect with each other, we are unable to create a 3D/4D cadastre unless the data fits together, as well as linking topo, addressing and place name datasets.

Irina said that the group had continued to grow and now had 65 individuals representing 29 agencies. PSMA, TERN and AODN were now members. However since Adam Rice left DTA, we have not been able to form a connection with that agency.

MDWG Terms of Reference

There was discussion on whether the Terms had previously been endorsed, Evert Bleys thought they had been conditionally accepted and wished to have further discussion about them. The Terms of Reference are a living document and can be updated and reviewed as required.

***Action 19/2/1:*** *Evert Bleys to provide his feedback on Terms of Reference to MDWG.*

The roadmap and Profile Sub-Group recommendations were accepted and endorsed at the previous meeting.

Evert said that a minimum set of elements had been identified and accepted, he agreed the concepts had been accepted, but the elements were not endorsed. There would be further discussion and revision during the workshop on day 2 of the meeting regarding this aspect.

Actions from previous meetings.

* Most actions had been completed, a full table is attached.
* A high-level survey was circulated to gain understanding of what is the baseline metadata catalogue in jurisdictions and agencies. There were 17 responses, which gave a good representation across the MDWG. The results will be circulated in full shortly, and a further presentation later in the meeting.
* We are trying to find a contact at Digital Transformation Australia (DTA) and build a relationship with that agency, as it is important to have contacts with *data.gov.au*.
* The impact of GDA2020 needs to further discussed and the implications for the MDWG.
* All final documents are published on the ICSM website.
* Feedback is welcome on the roadmap.
* Investigating best channel to connect and communicate with all MDWG members, possibly through ICSM website, OneDrive and OneNote.
* Further members are most welcome to join the group, and invitations can be extended by any MDWG members to others, especially the technical sub-groups.
* Further elements have been added to the profile recently, to be further discussed by Byron. Further sub-elements to be added.
* Andrew Whiting had a meeting with EMSINA and listened to their needs, challenges and expectations. We need to be clear how metadata can be used and adopted. They have asked when documentation will be available.

Roadmap progress

Andrew Whiting spoke to the progress of the three key tranches and the identified tasks. There is a need to agree conceptually to the core elements, how they map together in order to begin Tranche 2, the Blueprint and supplementary communication documents.

Irina said that the technical sub group had been working to provide tools, documents, applications and user guides. There is continued work mapping implementation of standards between agencies, EMSINA has specifically requested a need to connect metadata services to data, for use in the EMLINK catalogue.

GA (Aaron) has developed validation tool which is being tested by ANDS and AAD, and is available for people to use and validate their XMLs against GA profile (Schematron).

Shanti has been working on mapping versions of the Australia Protective Security Policy Framework, to the current one.

Irina would like to progress the work on the Blueprint for Metadata Best Practice (the Blueprint) by starting a requirement to describe images and adopting the new version of 19115-2 Standard as well as how to implement 19165. Vocabularies are to be defined, common classification, thesauri and tools to be resolved. Members are welcome to join the sub-group, and add to the sub profile activities.

Baseline Survey

The survey sent out was to gain a high level understanding of systems in use by jurisdictions; how often they were updated; how well they were integrated and what was unique for their business requirements.

This provided a knowledge of systems in the community, and to provide a baseline for the Blueprint, as well as determining where the strengths and weaknesses are. The survey will be conducted again in 12-18 months to gauge changes and impacts from MDWG work. The results to be published on the ICSM website, after they are de-identified.

***Action 19/02/02:*** *Publish de-identified baseline survey results on ICSM website in pdf format.*

***Action 19/02/03****: Andrew Whiting will write a short report for circulation to MDWG (and ICSM).*

MDWG Definitions - Byron Cochrane – OpenWork P/L

* Focus of work on definitions make sure that everyone is talking about the same thing where elements or data inside elements
* What is provenance and lineage of data
* Can't reach one standard that fits all but can agree on common elements, customise the experience to the product type
* Bring in DCAT definitions
* ISO definitions and summary definitions provided in table. Tended to avoid cardinalities.
* Poor commonality between Thesauri - good keywords and common definitions is very important need to make these widely available and in an easily accessible form. Important to make definitions public.
  + - * + Victorian Government is now looking at the DAMA method.
        + Having good key words, and common location is really important i.e. common definitions
      * Constraints – **copyright, security and legal** 
        + There is confusion around the implementation of differences licenses
        + Really the copyright applies to distribution not the data
        + End user license agreement
        + Provenance and lineage is testing
        + FSDF LINK is a good example on how to track provenance
* Metadata identifier needed in CSW to identify structure of metadata, mix of identifiers used to metadata and sometimes to the data. Needs further discussion at workshop.
  + - * Interesting discussion on the issues related to **provenance** and constraints
        + Automation of the aggregate record copyright
        + Inconsistent framework approaches I.e. VIC have one field of data applied to another dataset, means the whole dataset becomes locked "One bad egg"
        + **Information Security Classification Framework** where does this fit? SA and Defence use this framework

ISO Met is used for classifying the data

CCBY for the data

This framework needs to ensure there is data custodians and governing roles

* + - * + This also applies to responsibilities
        + Need to determine where in the metadata record we place in these roles and responsibilities
        + When governments MOG –then this places in another dimension of data

Perhaps the best option is versioning the agency, rather than the metadata record

* + - * + SA is beginning to undertake snapshot archives of their data, therefore they need to understand the version of the metadata.

In some examples the attributes change between versions, and this change needs to be somehow articulated

LandXML articulates how to do this.

How to future proof metadata, Shanti has metadata records of metadata.

ISO Standards – Chris Body

* Standards Australia continuing to publish ISO/TC 211 as standards AS/NZ
* Increasing coordination requiring spatial and positioning information
* Smart cities advisory group meeting – seeking Australian input, Australia contributing
* Changes in people Matt Purss new co-convener
* Systematic review ISO 19157 data quality, 19135-2 procedures for item registration,
* Confirm ISO 19129 imagery, gridded and coverage data framework
* Chinese taking more of a role – good resources but focused on their issues and solutions
* New work on discrete global grid
* Revising linear referencing
* 19160-6 addressing – digital exchange models, looking for more GA involvement
* Close collaboration with UNGGIM, OGC and IHO. UN sustainable development goals (very important within region, less engagement from Australian government)

Land administration domain model up for review - LADM v2

Many groups contributing, very many countries involved

Lot of work for this model, needs project management – needs reinforcement with ICSM

Maritime boundaries and BIM related to this also blockchain

**OGC**

* New management regime
* Testbed 15 – secure data, cloud processing and machine learning
* Study real world data with virtual 3D/4D environments, going to work on groundwater, body wear sensors, MetOcean profile
* Meeting in Singapore 25th February.

Lot of interaction with other standards groups much input needed by other groups to allow linking to spatial data

General Dataset Metadata Standards – Simon Cox

Web standards – DCAT 2014 and Schema.org

* Summary of DCAT – wraps around classic Dublin Core covers data set distributions - makes sense to people working in spatial data world. Allows more than one file making part of data distribution
* Used for data set description allows key word and theme to cover what data set describes
* Simple model relatively speaking
* In RDF
* Scope covers records in data catalogues – in many cases this will cover metadata
* Adopted as part of open government – relevant to spatial data (GeoDCAT-AP) and heavily adopted in Europe
* Mapping from ISO19115 to DCAT – allowing geospatial catalogues to be reflected in more general catalogues. Ties down some controlled vocabularies
* DCAT 2014 undergoing a refresh to DCAT-rev scope has been expanded allows cataloguing of data services and allows cataloguing of data catalogues

Schema.org - scope descriptions of stuff embedded in web pages to allow indexing

* Dominant organization using it is Google to allow indexing of webpages
* Google search can find metadata on web pages is web pages follow schema.org
* Google's data set search also harvests DCAT
* Google will crawl pages and index if information is embedded in schema.org
  + - JSON to help users find the webpage
      * This is specifically to make data discoverable through google. Need to be aware that this is Google method, and if you want to play, you must comply
* DCAT structures dropped into schema.org
* Important to make data discoverable by google and indexed in the way that custodians want their data to be indexed
* Mapping and translation from ISO19115 to DCAT and schema.org will allow discovery by google search crawls and people searching using google.

Profiles Guidance & Vocabulary – Nick Car

* + DXWG Profiling work – currently looking at developing a set of rules for profiling
* Proliferation of profiles in DCAT – need rules for profiles
  + Deliver
* Profiling guide
* Profile Vocabulary – describing how the profile is made
* Profile negotiate – how do systems request csw etc
  + The profile vocabulary – this vocab describes what are the parts to the profile, and what are the core parts
* Examples descriptors, validation tools, guidelines – all machine readable
* This vocab will describe all the profile
  + Profile negotiation
* Describes how internet clients may negotiate content – this will have a significant impact over time.
  + Using the profile vocab and testing it on a number of profiles to see if it works
  + Work to be completed in July. The three documents will be then endorsed by W3C
  + Australia Government Linked Data Working Group (AGLDWG) will have many ontologies descried this way

GA Web service metadata records - Aaron Sedgmen

Metadata for services how GA is implementing this

* ISO19119 defines geographic services
* This can include machine to machine or human interactions
* Services currently limited to web APIs
* In future possible link to SPARQL end points
* It is possible to have link to service within metadata
* Having metadata for services allows users to discover independent of finding data – provides more detail about service
* ISO19115 allows linking to service but hard to link other way
* GA profile has been extended to allow service to be recognized
* Not uncommon to have multiple different services on single data set – can accommodate in a single record but this complicates search and discovery
* Collection level record pointing to various services – 3 classes used - service identification class, operations class, service parameter class
  + SA is looking at metadata for services,
    - Moving towards AWS for web services
    - Challenge is the volume that they are being used
    - The more use, then the bigger cost
      * Now trying to find ways to reduce the load on the servers
      * Perhaps the metadata can assist in handling this
      * Discussion – the metadata could do this, however it would require frequent updating
    - Collecting service metadata – now there is a workflow to collect

AAD Implementation experience of ISO19115-1 - Dave Connell

* Have international treaty that requires scientist to make their data available internationally
* Previously used ANZLIC standard, but now use DIF and convert to ISO19115
  + 2019 a catalogue will be developed with the new 19115-1 fits
* Process is not automated and needs human checking (Dave) to quality check content for record meets requirements
* Traditionally the tools were very difficult to use, therefore a new system was implemented. A tool was developed to help scientists fill in records DIF based and has a wizard interface. Now there is a very easy to use metadata template – most people intuitively know how it works.
* All metadata made available through OAM-PMH service
* Have reasonable understanding of implementation of 19115-1 but not –3
* Used GA validator to identify issues – problems with altitude and depth units
* Created DIF-ISO19115 conversion tool

***Action 19/02/04:*** *Lessons learnt register to be created and circulated.*

Metadata catalogues in QLD and TERN - Nick Car

* + Both catalogues are contributing to a total information graph, and both are working towards DCAT2
  + Using COTS
* CSIRO providing help to model data relationships at whole of organization level – two perspectives used:   
  A: Management perspective -every data set must have a record so agency knows what it is accountable for.   
  B: Realm perspective for what GSQ communicates to its clients
  + Currently developing a few catalogues using CKAN, but not all things can be catalogued in this system
* Use Linked Data to bring things together from different systems
* All systems can export in RDF using overarching model
* Catalogues have code lists for vocabularies Use vocabularies management tool – mapping exercise to GA vocabs
* Available in RDF conformant to DCAT
* Samples data held in graph database
* Similar approach used for TERN Graph databases used for sites, samples and observations
* TERN needs data ingestion from different sources but must use an overall TERN data model. Uses a validator that checks if data can be ingested.

Future is cloud based. Need enterprise data model to move to cloud

the LIST example from Tasmania - Todd Baker

* Live demonstration of the LIST from client side
  + LISTmap provides visual search through data layers – linked to metadata portal
  + TheLIST is the central hub for all land tas information
  + TheLIST includes property information, mapping information and metadata
  + All elements are integrated, so the mapping information links to the metadata
    - Users can go from catalogue to data and the other way
* LISTdata provides search through metadata catalogue
* Custodian Users wanted standards based but easy to enter metadata, secure and public metadata
* GeoNetwork used to build system, however content management system draped over top (Open CMS) to make front end user friendly
* Search by text, categories, by map
* Look and feel like car sales, gum tree to make user friendly. Filters used to allow users to narrow search
* Page display easy to read, nice and clean – hyperlink to services
* Integrated license and linking to creative commons
* GeoNetwork instance exposed if people want to link in to it
* Cloning and bulk editing capability available

Workshop preparations on Day 2

* + Mapping spreadsheet – demo on the mapping and analysis undertaken to identify some issues for facilitation
  + Use Creative Commons 4.0; but an issue between who has such rights - the creator, the aggregator or the publisher. Nick Car said that Creative Commons can be licensed in such a way to cover individual rights. Perhaps also geolock such data.
  + **Identifier** discussion -
* Multiple identifiers, single identifiers, CKAN un resolvable identifiers, where to store the identifier
* URI? Can use URIs but issue with CSW records.
* One identifier for the metadata record, citation.
* Alternative identifiers need to be recognized by harvesting services. 4 logical places
* This is a known problem around identifiers and MDWG can assist in
* **Key words and Thesauri** – looking at how to use in the record
* **Contacts and key parties** – objective to get a consistent agreement on where
  + **Cardinality** - one to many relationship, put most important first. Evert working on this.
  + **Tracking Governance Changes** – to be dealt with at a later time.
  + **Constraints** -
    - Run through the concepts in the spreadsheet – to articulate why they are important for different organizations
    - Identification on cross domain elements
    - Have a matrix articulating what are the core recommended
    - Good practice
* How to we handle mapping to other standards - Cross domain mapping?

**Additional Elements**

* People need to understand what is important to them, and appreciating others will have different needs that prevent them from using all recommendations. Best practice vs. good practice.
* Use Ted Haberman charts (<https://www.tedhabermann.com/blog/2019/2/19/metadata-evolution-crossref-participation-reports> ) to gauge progress on elements and analyse how well being used.
  + - Focus on the wrapper wording for endorsement, everyone has different business drivers that affects their requirements.
* Who's your audience? The business creating metadata may not understand the audience, what their users want or need. Clients may need more (or less) than what is provided.
* Understanding what harvesters are scraping.
* Missing interoperability, missing half the equation and what we should be doing, and what we need to do. Interoperability will be an incentive to drive good practice.
* EMSINA is a good use case, just show us what to do, so we can do it. (Similar experience to Tasmania and AAD).
* Do this at class level
* Importance of element should be decided
  + - Have a list of **recommended elements -** mandatory elements are what are needed at departmental level, never going to get Commonwealth national policy.
  + Workshop – focus on the recommended elements – with the goal of the general user, need to triage the elements.

***Action 19/02/05****: Evert to provide missing elements to spreadsheet****.***

Other issues:

* Unable to engage with data.gov.au (DTA), no representative at MDWG at the moment. They have an infrastructure (MAGDA) that will degrade as they have no particular interest in metadata, let alone standards. This is a risk that should be managed by ICSM and ANZLIC.
* Should GA reinvigorate the FIND catalogue if data.gov.au ceases or fails to deliver.
* Is there anyone in MDWG who would be concerned if data.gov.au fails ?
* When the FSDF Link was being developed, jurisdictions said to just harvest their metadata catalogues, however this was not possible as they were mapped differently and all used different ontologies. We need to start investing back into core capabilities of catalogues, to have consistency in jurisdictions and agencies.

# Day 2 – Workshop by Byron Cochrane

**Review key Elements and definitions**

Byron has created table of elements to share and review

Matrix for making decision on elements General-Specialized vs relative usefulness

Top level items have greatest general use highest usefulness

Evert would like to provide support and advice, his spreadsheet is useful to creators, not for users to access.

Discussion on development of 'profiles' - Best Practice Implementation Guide

Graham Logan said members of this group understand what is needed, but need to communicate this to others users and consumers of data.

Nic: can be multiple levels, there is a need to bring systemisation to it all

Byron would like workshop to produce a high level model, with further specialisation for each business to be added later. Objective is the common elements and the technical details supported by the narrative.

Educate on how to adopt 19115

Principles to articulate how to work within this framework

Can we sell to our managers - what we do is defensible and understood outside this group.

**Metadata identifier** –

* Everyone would be expected to use an identifier, but not recommended as mandatory.
* Everyone likely to do things in different ways due to internal business.
* Not a lot of value in sharing internal identifiers.
* Would need to be a URI if it is to be resolved externally.
* Two fields one text based and other for a URL. Use link as a URI in the URL field to avoid link breakage.
* Recommended that people use URIs

**Date** – should be last update date, some need for date time not just date.

* + Order of preference
    - Revision – last update
    - Publication
    - Creation
  + Precision preferred, don't use mock times if time is unknown.
  + Temporal resolution should be stated – solution to be identified.
  + What to do for deprecated sets, how do you indicate precision: date or date/time.

**Responsible Party –**

* This is a mandatory element
* Contact info (for metadata) should be provisioned for ORC ID to allow tracing of academics moving across agencies. However, position/role identification is preferred for government positions.
* Goal is to aim move away from individual to position
  + VIC is moving to DMBOK where it is now to reference individual
* Contact point responsible for metadata record, not dataset. Possible to have multiple contacts. Don't use citation, this is for author of data set. Can just be an email address. Need enough information for a user to make contact.
  + As there is diversity in options, perhaps we should have options
  + This role needs to ensure there is a contact at a minimum
  + Need to ensure we are future proof – that longer term – liability and risk management
  + Clarification of lineage of responsibility can specify individual.

**Locale, Language** –

* Can go without but should have language (not mandatory) but highly encouraged. Language metadata documented in. Content must be UTF8 for character set, use Australian style guide.
* Needs to contain both sub elements

**Metadata Scope** – should be mandatory, populated from a code list.

**Legal Constraints** –

* for metadata record, very lightly populated (optional) but recommended to populate automatically with creative commons so people can reuse metadata record.
  + Decision: remain optional, however strongly recommend to use
    - This includes legal and security constraints
    - License and security is new to the standard – and this is a link
  + AGRF framework. We need to ensure all government records (including emails) comply to this framework. These legal constraints should align.
  + **There needs to be a vocabulary list of these terms – this also needs to be a technical list**
    - Problem the ISO list is old – therefore we need to create an extra list
      * Who will publish this list – AGLDWG
        + Long term sustainability will be OK through AGLDWG
        + Term governance is more difficult as the terms are used by many different domains

***Action 19/02/06****: Nick to turn list into a Linked Data Vocab****.***

**Identifiers**

* + Evert – different people use it in different ways – depending on business needs – therefore make it optional
  + Several years ago it was adopted by ANZLIC as being optional
  + Nick:
    - UUID is unique, however there is no regime for publishing URI do not resolve
    - URI can do the same, and they have associated governance
  + Issue: GeoNetwork does not support
  + Evert – Add metadata linkage as a new set of elements
    - Keep the traditional identifier
    - Also insert the URL
  + Persistence – this item will stop the "Broken Link"
  + The URI also can show all the end nodes are referencing that URI recommend
  + Recommendation that URI is included in **Metadata Linkage**
  + This is a fundamental outcome that Craig Sandy wanted to achieve.

**Evert's additional elements discussion**

* There is a place in 19115-3 for UUID. This would be recommended if you are going to use UUIDs.
* If your system is doing UUIDs then use an element that declares this. Can be set up to fill by default.
* Parent identifier is worth considering.
  + GDA2020 precision element is important – this will show if the data needs to be modified
  + Addition – this field is very good as it can be used for archiving

Scope – two main components. refine the description of the scope by using both fields.

Classes that describe resource – these sorts of things should be considered:

* If object or data set is spatial declare what it is and what size it is.
* Declare size of vector data set so users can assess if they can cope with it.
* Declare reference system. Problem if there are multiple reference systems
* Data identification highly desirable unless it is a service
* Citation edition – needed if there is archiving

Many additions that are being recommended for particular business reasons. Not all of these reasons apply to all users but many should be considered when implementing a new profile. Evert needs to document the rational behind why someone would need to make things mandatory.

Graham Logan – Two exercises

1. Evert education exercise
2. Broader community – General reference – more in line with what Byron is doing

* By Locate will we be in a position to have Byron work published – the high level document
* Evert can provide a conclusion to the education document in a short term however it will require extra wording

**Messaging at Locate**

* Abstract has been accepted – this will be an introduction and update on the work undertaken
  + The message that we will end up with several profiles
  + Do we need to have some handouts?
  + Working draft on the website – seeking your feedback
* It is important we begin to articulate what we are doing very soon as we are not doing the same as what we have done before
* Working group focus for next 6-8 weeks
  + Need agreement on the focus of the report to put out publicly on the ICSM website
  + Working groups should step through the spreadsheets – Lumeo can assist
  + Evert's work can be brought together soon, then distributed, integrated longer term into the document and updating
    - Provide in spreadsheet
    - Provide this is how you read it
      * the focus will be advising elements which you need to consider
      * First pass will be quickly compiled, second pass will be on the **why**
* **Early April** -
  + Byron report to be published on website
  + Working group approved elements
  + Evert will conclude the spreadsheet and a how to read tab
  + This will support the general guide
  + The complete xml document
  + Next phase, see if Nick can build a tool
* This timeline should meet ANZLIC expectations

* Other Business:
  + Need a discussion forum Lumea is an option or Confluence, Trello is another option
  + Perhaps a lite meeting at Locate or next meeting in June in Canberra
* George can present, SA present
* Need to keep in mind how this will work in reality
* Review key word vocabs which are around
  + Very GeoFocused – but now there are many domains
  + Perhaps the focus should be on common vocabularies