Transform & ingest a vocabulary [Spreadsheet/CSV format]

Transform a vocabulary to spreadsheet format

1. Introduction

A controlled vocabulary reflects agreement on terminology used to label concepts. When research organisations agree to use common language the discovery, interpretation, understanding and reuse of research data is improved. For more information on controlled vocabularies, please see the ARDC guide to vocabularies and research data .

ARDC provides a Vocabulary Service for use by our partners. The service supports research organisations to publish and discover controlled vocabularies.

In order to support our partners in making their vocabularies available and browsable via the ARDC Vocabulary Service, we provide this guide and accompanying ingestion template, which outline a transformation and ingestion process. In addition, section 5 details the discussion of a completed example of this transformation and ingestion process. If you have any additional questions about the transformation or ingestion of your vocabulary, please contact services@ardc.edu.au .

2. Getting started: Questions about your vocabulary

File formats

In what format is the vocab currently being maintained/stored?

- Examples of formats in which a vocabulary may be stored: Spreadsheet, CSV, PDF, text, HTML, SKOS, database tables, etc.
- Has ARDC already developed a transformation and ingestion process for that format? Has your organization developed a process to transform the current format to SKOS? If not, we will work with you to develop a process.
- Note : This is a guide for cases in which there are vocabularies that have a semantic model that can be adequately expressed within the constraints of a spreadsheet or comma separated values (CSV) format. For information about the transformation and ingestion of vocabularies that are maintained/stored in other formats, please consult our other transformation and ingestion guides.

Concept definitions

How are the vocabulary concepts currently being described?

- What are the elements used to describe metadata about the concepts?
- What do these elements mean?
- How do the current elements used to describe metadata about the concepts map to the ARDC Vocabulary Service ingestion template?

In order for your vocabulary to be ingested into the ARDC Vocabulary Service, the information provided in the original format needs to be translated into the ingestion template provided by ARDC. The template allows ARDC partners to indicate what information about the vocabulary should be captured within the following elements:

URI	<uri>	• An identifier which is guaranteed to be unique among all identifiers used within the vocabulary. This identifier will be used to create a unique URI for each vocabulary concept. • If you do not have a predefined URI structure you'd like to use with your vocabulary, ARDC can provide support in this decision. • The URI column is optional for ingestion.
Scheme	<scheme>	• Scheme is an element that allows you to designate a machine-friendly, unique string of characters for your vocabulary. • The Scheme column is optional for ingestion.
Concept	<concept>	• Concept is an element that makes it possible to assign a machine-friendly, unique string of characters for each concept. • Information captured in the concept element may also serve as the preferred label for concepts, if you choose not to make a distinction between concepts and their preferred labels. • At least one concept column is required for ingestion.
Preferred label	<prefLabel>	• Preferred label is an element that makes it possible to assign a human-friendly, unique label for a concept. • If your vocabulary is multilingual, you may use the language tag to provide Preferred labels in multiple languages (see Language Tag below). • The Preferred label column is optional for ingestion.
Alternate label	<altLabel>	• Alternate Label is an element that makes it possible to assign an unauthorized name to a concept. • An example might be a preferred label for the concept “fava bean” and an alternate label of “broad bean.” • The Alternate label column is optional for ingestion.

Hidden label	<hiddenLabel>	• Hidden Label is an element that makes it possible to provide a label for a resource that needs to be accessible to applications performing text-based indexing and search operations, but not visible otherwise. • Hidden labels may be used to include misspelled variants, jargon, or colloquialisms of other labels of the concept. • An example might be a preferred label for the concept “potato” and hidden labels of “potatoe,” “tater” and “spud.” • The Hidden label column is optional for ingestion.
Notation	<notation>	• Notation is an element that captures alphanumeric codes such as "T58.5" or "303.4833" used to uniquely identify a concept within the scope of a given vocabulary, but is not normally recognizable as a word or sequence of words in any natural language. • Classification codes or schemes may be captured using the notation element. • The Notation column is optional for ingestion.
Scope note	<scopeNote>	• Scope note is an element that explains and clarifies what is meant and what is not meant in the definition of the concept and in its use in the vocabulary. • An example might be a preferred label for the concept “vegetable” and a scope note of “ The concept vegetable excludes other main types of plant food, fruits, nuts and cereal grains but includes seeds such as pulses. ” • The Scope note column is optional for ingestion.
Example	<example>	• Example is an element that details an instance serving as an illustration for other instances of the concept. • An example might be a preferred label for the concept “potato” and an example of “ T he prize potato, grown by Peter Glazebrook, tips the scales at a whopping 8lbs 4oz (3.76kg), smashing the previous world record by 9oz. The vegetable, Peter's Kondor variety, was put on show on Friday at the National Gardening Show in Shepton Mallet, Somerset. ” • The Example column is optional for ingestion.

Definition	<definition>	• Definition is an element that supplies a complete explanation of the intended meaning of a concept. • An example might be a preferred label for the concept “potato” and a definition of “ T he potato is a starchy, tuberous crop from the perennial nightshade Solanum tuberosum L .” • The Definition column is optional for ingestion.
Exact match	<exactMatch>	
Close match	<closeMatch>	
Related match	<relatedMatch>	
Broader match	<broaderMatch>	
Broader	<broader>	
Related	<related>	

And the following tag:

Language	@lang	• A language tag allows for the language of text provided in any of the elements to be indicated. Any language code scheme may be used. Guidelines for best practice in language tag usage can be found here . **Multiple languages** If your vocabulary is multilingual, each concept may have more than one prefLabel , as long as each prefLabel is designated with a DIFFERENT language tag. For example, the concept “potato” might have 2 preferred labels: prefLabel "Potato"@en; prefLabel "Kartoffel"@de.

This is not a complete list of all elements which can be captured for your vocabulary in the ARDC Vocabulary Service. If your organization captures extra information that does not fall under the listed elements or tag, we can work with you to create a solution for including that information in your transformation. Please contact services@ardc.edu.au if you have any questions about your transformation process.

Hierarchical structure

- In order for the vocabulary to be ingested properly, the hierarchy (narrower and broader nature of the concepts) must be notated in a machine-readable way. This may require some reorganization of the concepts for insertion into the ingestion template.

Additional preprocessing considerations

What preprocessing needs to be done?

- Are there any additional requirements of the vocab owners or other stakeholders that might impact the transformation or ingestion of the vocabulary into the ARDC Vocabulary Service?
- Have all non-ingestible (non-ASCII) symbols been removed?
- Is the vocabulary multilingual (does it include content in multiple languages)? If so, please provide ARDC with a list of languages used in the vocabulary prior to ingestion.

3. ANZSRC-FOR: An example transformation

The process of vocabulary transformation and ingestion has been performed on the ANZSRC-FOR vocabulary, and the artifacts from that process are provided here as an example for future use.

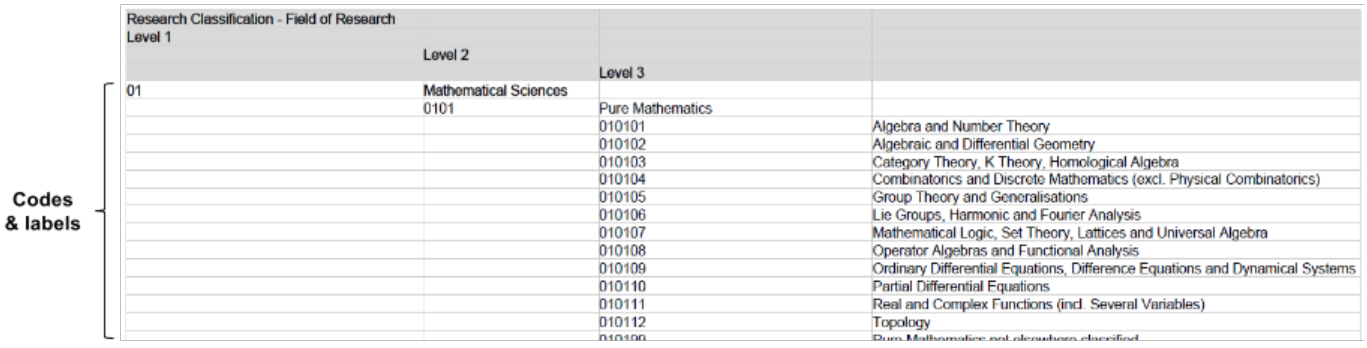
> This is just one example of the transformation of a vocabulary, and is meant to be used as a learning tool. The steps taken in order to transform your vocabulary may vary from those outlined below. Please contact services@ardc.edu.au if you have any questions about your transformation process.

In what format is the vocabulary currently being maintained/stored?

The vocabulary was initially provided as a spreadsheet in Microsoft .xls format, which can be viewed here . An annotated version of this original document (in Google Docs spreadsheet format) can be viewed here . (Completed transformed versions of the ANZSRC-FOR are available in spreadsheet format and CSV format .)

How are the vocabulary concepts currently being described?

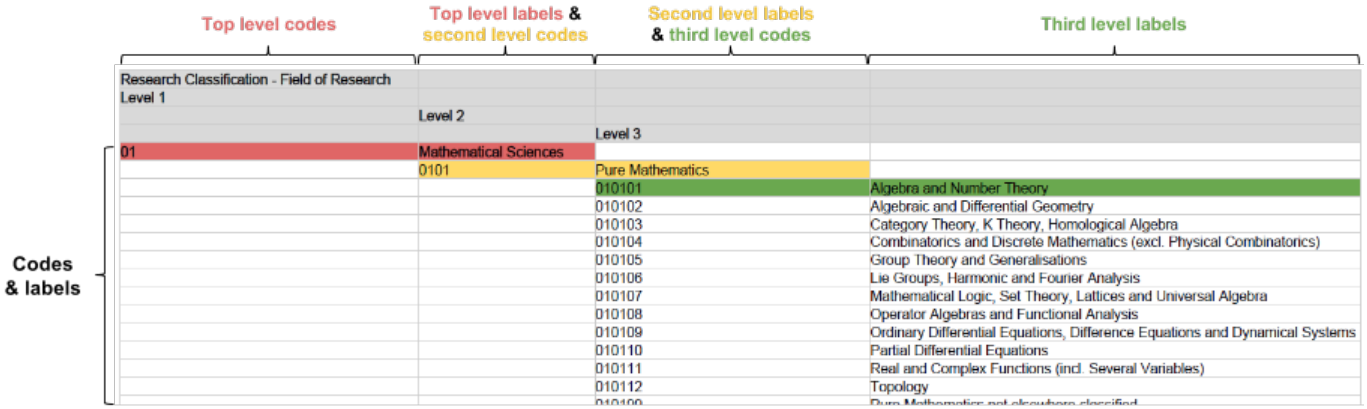
The ANZSRC-FOR vocabulary spreadsheet includes the title of the vocabulary, some column headings to explain how to read the spreadsheet (shown below in grey cells), names of vocabulary concepts, and codes that correspond to the concepts. This structure and code scheme is explained in detail by the Australian Bureau of Statistics here .



Research Classification - Field of Research Level 1	Level 2	Level 3	
01	Mathematical Sciences		
	0101	Pure Mathematics	
		010101	Algebra and Number Theory
		010102	Algebraic and Differential Geometry
		010103	Category Theory, K Theory, Homological Algebra
		010104	Combinatorics and Discrete Mathematics (excl. Physical Combinatorics)
		010105	Group Theory and Generalisations
		010106	Lie Groups, Harmonic and Fourier Analysis
		010107	Mathematical Logic, Set Theory, Lattices and Universal Algebra
		010108	Operator Algebras and Functional Analysis
		010109	Ordinary Differential Equations, Difference Equations and Dynamical Systems
		010110	Partial Differential Equations
		010111	Real and Complex Functions (incl. Several Variables)
		010112	Topology
		010199	Pure Mathematics not elsewhere classified

Preprocessing of the vocabulary

Examination of the vocabulary in its original spreadsheet format reveals that the given column headings (shown below in grey cells) do not provide all of the information we need to transform the spreadsheet. There are multiple types of information recorded in individual columns:



Research Classification - Field of Research Level 1			
	Level 2		
		Level 3	
01	Mathematical Sciences		
	0101	Pure Mathematics	
		010101	Algebra and Number Theory
		010102	Algebraic and Differential Geometry
		010103	Category Theory, K Theory, Homological Algebra
		010104	Combinatorics and Discrete Mathematics (excl. Physical Combinatorics)
		010105	Group Theory and Generalisations
		010106	Lie Groups, Harmonic and Fourier Analysis
		010107	Mathematical Logic, Set Theory, Lattices and Universal Algebra
		010108	Operator Algebras and Functional Analysis
		010109	Ordinary Differential Equations, Difference Equations and Dynamical Systems
		010110	Partial Differential Equations
		010111	Real and Complex Functions (incl. Several Variables)
		010112	Topology
		010199	Pure Mathematics not elsewhere classified

In order for ANZSRC-FOR to be ingested into the ARDC Vocabulary Service, the content provided in the original spreadsheet needs to be entered into the ingestion template provided by ARDC. The template allows us to indicate what original vocabulary content should be captured within the following elements:

URI	<uri>	• Any identifier which is guaranteed to be unique among all identifiers used within the vocabulary. This identifier will be used to create a unique URI for each vocabulary concept. • In the case of the ANZSRC-FOR exercise, unique identifiers were generated based on the labels of the concepts. The URI structure http://abs.org.au/def/anzsrcfor/{conceptidentifier} has been chosen as a standardised format. • ANZSRC-FOR example: library-and-information-studies
Concept	<concept>	• Concept is an element that makes it possible to assign a machine-friendly, unique string of characters for each concept. • In the case of the ANZSRC-FOR exercise, information captured in the concept element also serve as the preferred label for concepts. • In the case of the ANZSRC-FOR exercise, the concept element will map to the concept names (or labels), provided in columns B, C and D of the original spreadsheet. • ANZSRC-FOR example: Library and Information Studies

Notation	<notation>	• Notation is an element that captures a string of characters such as "T58.5" or "303.4833" used to uniquely identify a concept within the scope of a given vocabulary. • In the case of the ANZSRC-FOR exercise, this will map to the codes provided in columns A, B and C in the original spreadsheet. • ANZSRC-FOR example: 0204

> This is not a complete list of all elements which can be captured for your vocabulary in the ARDC Vocabulary Service. If your organization captures extra information that does not fall under the listed elements or tag, we can work with you to create a solution for including that information in your transformation. Please contact services@ardc.edu.au if you have any questions about your transformation process.

In the case of ANZSRC-FOR, the elements used are unique identifier , concept and notation . Because the original ANZSRC-FOR spreadsheet doesn't include content such as concept definitions or alternate labels for concepts (and in fact, these pieces of information don't exist for this particular vocabulary), those columns are left blank in the completed ANZSRC-FOR ingestion template example.

The ingestion template allows for ARDC partners to capture information about the hierarchical structure of their vocabulary and metadata about the concepts in one document.

Concept metadata

A number of steps were performed in order to properly record metadata about the ANZSRC-FOR concepts in the ingestion template.

1. The preferred labels were pulled from columns B, C and D of the original spreadsheet and pasted into the column titled "concept" in the template and the codes corresponding to the Preferred labels (pulled from columns A, B and C of the original spreadsheet) were pasted into the column title "notation" ensuring that codes corresponding with labels were pasted into the same row of the spreadsheet.

For example:

A	B	C	D
Research Classification - Field of Research			
Level 1			
	Level 2		
		Level 3	
02	Physical Sciences		
	0201	Astronomical and Space Sciences	
		020101	Astrobiology
		020102	Astronomical and Space Instrumentation

becomes:

C	D	E	F
concept	concept	concept	notation
Physical Sciences			02
	Astronomical and Space Sciences		0201
		Astrobiology	020101
		Astronomical and Space Instrumentation	020102

1. Because ANZSRC-FOR is a monolingual vocabulary (English language), no language tags are necessary.

1. Unique identifiers for each concept were created based on the Preferred labels using the =lower and =substitute functions and by deleting all punctuation and any text within parentheses. Unique identifiers corresponding with labels were used to create URIs for each concept (using the predefined URI structure) and were inserted into the URI column of the spreadsheet.

For example:

Preferred label of Analytical Chemistry becomes unique identifier analytical-chemistry

Preferred label of Automotive Combustion and Fuel Engineering (incl. Alternative/Renewable Fuels) becomes unique identifier automotive-combustion-and-fuel-engineering

Completion of ANZSRC-FOR example

The completed vocabulary ingestion template for the ANZSRC-FOR is available in spreadsheet format and CSV format .