Workshop Report

Advancing Australian Vocabularies and Mappings

> ***"You can't use an old map to explore a new world."***
>
> ***– Albert Einstein***

Ms Amanda Belton[1], Professor Dougie Boyle[1], Dr Noel Faux[1], Rowan Brownlee[2], Meirian Lovelace-Tozer[2], Roger Ward[2]

07/11/2024

[1] The University of Melbourne

[2] The Australian Research Data Commons

Suggested citation:
Belton, A., et al., (2024) Advancing Australian Vocabularies and Mappings, The University of Melbourne & ARDC

Viewed online at: **Figshare 10.26188/27913398**

TABLE OF CONTENTS

Executive Summary

This report has explored the current limitations of vocabulary mapping within the Australian healthcare system and outlined a path toward improvement. **The primary recommendation is to prioritise a staged approach that focuses on establishing governance mechanisms, leveraging metadata to enhance trust and findability, and strategically integrating AI technologies as they mature.** By doing so, the Australian healthcare community can build a more efficient and effective system for managing vocabularies and mappings. This collaborative effort will not only improve data discoverability and integration but also advance research and clinical systems, enabling more comprehensive and reliable health data analysis. The successful implementation of these recommendations depends on continued community engagement and a commitment to open-source solutions that prioritise interoperability and accessibility. The aspiration is that the developed ecosystem would be applicable beyond the health domain.

Introduction

ARDC strategic drivers

The ARDC is addressing national data challenges through:

- Focus area 1: data strategy and data discovery
- Focus area 2: secure data access
- Focus area 3: data integration
- Focus area 4: advanced analytics

This workshop aimed to understand the requirements and priorities for health and medical vocabulary services, through consultation with key stakeholders from key organisations, see Table 1 for details. The participants representing these organisations also brought their deep knowledge of ISO standards, and key tools and services including MeTEOR, Ontoserver and Snapagogo. Participants also brought their lived experience associated with currently mapping terms between in-house systems to OMOP standards. The workshop and this report aims to support ARDC in focus area 1, Data strategy and data discovery as well as focus area 3, Data integration. The vocabulary mappings could be considered low volume, high value datasets and data relationships.

Workshop Aims

Specifically, the workshop's aims were to understand the expected functionality of a tool(s) used to curate vocabularies and map specific terms, as well as current pain points in the workflow associated with mapping between terms from local vocabulary lists to other vocabularies (i.e. domain specific national and international standards), as well as sharing of these maps to a wider community.

The workshop was divided into three components: Introduction, and background highlighting a number of key tools (Table 2), focus groups based on three use cases. These were:

a. Term lists,
b. Custom Vocabularies, and
c. Mapping between Vocabularies.

At a high level, the feedback from the workshop was that the technology solution needs to be a robust, user-friendly solution that facilitates collaboration, automation, standardisation and integration across the vocabulary and mapping lifecycle (Figure 1).

Vocabulary and mapping lifecycle

The underlying understanding of the vocabulary mapping lifecycle is described in figure 1 below:

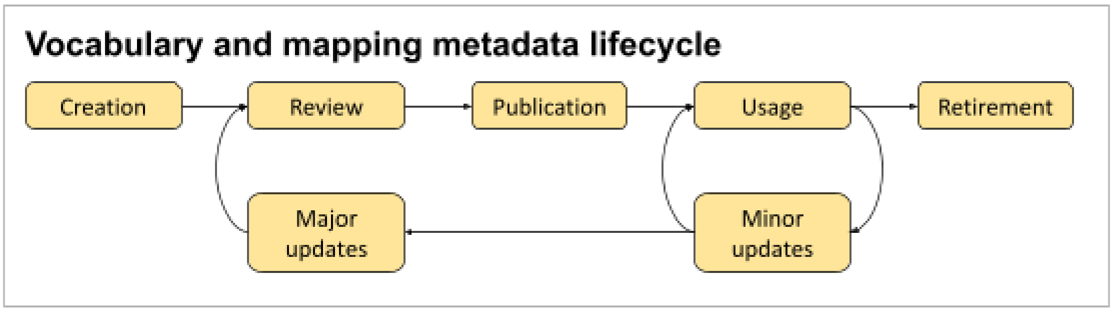


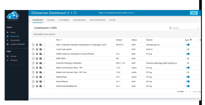
Figure 1: Vocabulary and mapping lifecycle

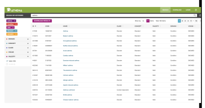
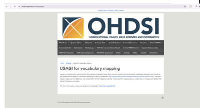
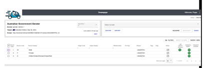
Participants

We are grateful for the expertise and experience shared by participants in the workshop. Organisations that were represented include: MCRI, Murdoch Childrens Research Institute; RCH, Royal Childrens Hospital; CSIRO, Commonwealth Scientific and Industrial Research Organisation; UoM, The University of Melbourne; ANU, Australian National University; KRIA, The Kids Research Institute, AIHW, Australian Institute of Health and Welbeing, EHE, eHealth Education, PHCN and Peninsula Health Care Network.

Ecosystem of tools

The current tools within the ecosystem, we believe, are insufficient. Each addresses specific needs, but none provide a comprehensive solution. The outcomes of this workshop are here to help us understand the attributes of a solution that will address these shortcomings and better meet the requirements.

Tool	Description	Owner/ Author	Strengths and limitations	Screenshot
Ontoserver	OntoServer is a robust, standards-based terminology server that is widely adopted globally for providing programmatic access to major clinical terminologies principally SNOMED CT. It is the means in Australia of managing the availability of SNOMED-CT (the National Clinical Terminology Service - NCTS). Its availability through a free national licence makes it an important piece of terminology infrastructure in Australia. 1. It is designed for runtime use of terminologies, rather than just terminology management. 2. It has deep support for handling complex terminologies like SNOMED CT and RxNorm. 3. It is not specific to the healthcare domain, but implements standard FHIR APIs for terminology access. 4. OntoServer is widely used internationally, with CSIRO having national licensing agreements in 10 countries. 5. It is deeply embedded in the HL7 community, and is also increasingly being used in the openEHR community. 6. There is a free licence available for anyone to use OntoServer in Australia, provided by the Australian Digital Health Agency.	CSIRO		

Athena	Athena is an online tool used for looking up and mapping terms, particularly to SNOMED CT codes. Athena allows users to search for a term and it will provide the possible SNOMED CT codes and concepts that match that term. It has the ability to filter the results, for example, to only show SNOMED CT code.	Observational Health Data Sciences and Informatics (OHDSI) group	Limitations: - It can only search for one term at a time, so it is not well-suited for bulk mapping of large term lists. - It does not have the ability to store or manage the mappings that are created.	
Usagi	Usagi is a widely-used terminology mapping tool that is designed for mapping terms to vocabularies. 1. Maps terms not just for OMOP but for various medical vocabularies. 2. It has access to over 5.5 million terms across these different vocabularies, providing broad coverage. 3. Usagi allows for bulk mapping of terms, rather than just looking up individual terms. 4. However, Usagi does not have the ability to store or manage the mappings that are created. The mappings have to be exported as a CSV file. 5. Usagi provides a confidence score for its mapping suggestions, which some find useful but others find controversial. 6. Usagi is an older tool, around 10 years old, but is continually updated by the OHDSI community.	Observational Health Data Sciences and Informatics (OHDSI) group	Excels at bulk processing of terms, but lacks the workflow and storage capabilities of more modern vocabulary management platforms	
Snapagogo	Snapagogo is a collaborative tool for mapping vocabularies. This tool works within a wider system of technologies to store and curate the vocabularies with a focus on authentication, workflow, and automated mapping capabilities. For people using the tool they may also be using local spreadsheets to manage informal vocabularies. 1. Authentication and Access Control: - Snapper GoGo uses the Australian Access Federation (AAF) for authentication, allowing		Limitations - Currently, Snapagogo's coverage is limited to conditions/diagnoses and medications/drugs. The API for Snapagogo is	

	users to log in with their institutional credentials. - This enables a collaborative workflow where users can be assigned different access levels (read, write) to shared mapping projects. 2. Mapping Capabilities: Snapagogo supports mapping for both conditions/diagnoses and medications/drugs. - It has an "auto map" feature that can automatically map terms to standard terminologies like SNOMED CT and RxNorm. - The tool stores the mapping projects, allowing them to be shared and collaborated on by multiple users.		not documented, but the underlying architecture uses FHIR APIs and an Ontoserver instance. It does not yet support regular expressions and handling narrative text.	
RVA	Research Vocabularies Australia allows usesr to find, access, and reuse vocabularies for research. Some vocabularies are hosted by the Australian Research Data Commons (ARDC) and can be accessed directly through Research Vocabularies Australia.		A limited service that can be enhanced with more detailed and comprehensive information.	
POC Visual Tool (Roma's tool)	This is a prototype tool that is designed to help users with a medical background find matching SNOMED terms for a list of plain text entries. It shows the term being mapped and allows the user to modify the search text to find a better mapping. It displays the suggested SNOMED concept matches in a graph, allowing the user to navigate the hierarchy and select the best match. The tool is focused on handling the challenges of mapping free-text medical terms to the appropriate SNOMED codes, including dealing with synonyms, abbreviations, and hierarchical relationships.		Limitation - this not a complete solution, but rather a prototype that demonstrates some core features for assisting with this terminology mapping task.	

Table 2: Ecosystem of tools. This summary table of the ecosystem of tools was generated from the meeting transcript using Claude Anthropic and reviewed and edited by the presenters and report authors.

Ecosystem Limitations

The current ecosystem of tools, including Snapagogo, do not yet enable creating, curating, analysing and searching across diverse data sources. The complexity and diversity of user requirements in the health space and beyond is challenging.

Approaching this at a practical level, the ARDC medical research data integration framework provides one potential strategy to advance work based on real-world use cases based on vocabulary definitions, data mapping and dataset standardisation to the OHDSI OMOP Common Data Model - a mature, international interoperability standard for medical research. Such an approach could provide foundations for toolsets that may operate across multiple domains (beyond health).

We do not list all limitations here - these vary according to tool. One example of a mature capability we can learn from in the 'search space' is the OMOP CDM EHDEN Portal (https://www.ehden.eu/) - a portal detailing searchable data from 199 European OMOP-converted datasets on over 380M patient records (Nov 2024). The standard data structures, terminologies and governance in appropriate data sharing has built trust through its work providing a single discovery portal for Europe.

Key Insight:
There is currently no accessible method for researchers to search for available mappings. Establishing a system to publish and share information about existing mappings would significantly enhance research efficiency and collaboration.

Workshop Outcomes

Key Functionality Requirements.

The workshop outcomes are organised into ten top level categories: mapping process, governance and licensing, discoverability and integration, automation, collaboration, versioning, documentation, search, API integration and interoperability. These outcomes from the discussion covered more than the requirements for a technology solution.

There were notes (Miro stickies that documented the discussion) that were considered to cross categories. Notes that covered the functionality, Integration and API categories were split into each of the individual categories.

Based on the collection of notes from the miro board and using University of Melbourne's Spark AI instance of the Large Language Model, Anthropic Claude 3 Sonnet (200k), it was asked to identify the overarching needs of the community: (Dr. Noel Faux reviewed and edited these, and are in agreement with his readings of the notes)

1. Robust functionality for managing the mapping process - defining closeness criteria, comparing vocabularies, handling hierarchies, bi-directional mapping, reviewing/curating mappings, visualising deltas, and supporting different mapping types.

2. Strong governance, licensing, and permissions controls - ability to share/collaborate on mappings, control access, validate mappings, and manage mapping maturity levels.

3. Discoverability and integration - See how others have mapped datasets, integrate with standards like SKOS/FHIR, provide a dictionary of available mappings, allocate DOIs to vocabularies.

4. Automation capabilities - Propose mappings based on closeness, learn from accepted mappings, handle patient-reported data, use conventions to suggest mappings.

5. Collaboration features - Centralised updatable source, consolidate local copies, make individual mappings shareable.

6. Versioning support - Map between vocabulary versions, rollback versions, check for updates.

7. Documentation - Capture mapping rules/guidelines used.

8. Search - Fuzzy search for finding terms to map, handle abbreviations/misspellings.

9. API integration - Import/export mappings, check versions, convert formats.

10. Interoperability - Work with existing mapping formats, support conventions beyond just mappings.

Focus requirements for each high level category

The model was then asked to identify the specific requirements for a technology solution for the top level categories: (Dr. Noel Faux reviewed and edited these, and they are in agreement with his readings of the notes)

These have been grouped visually in figure 2 below.

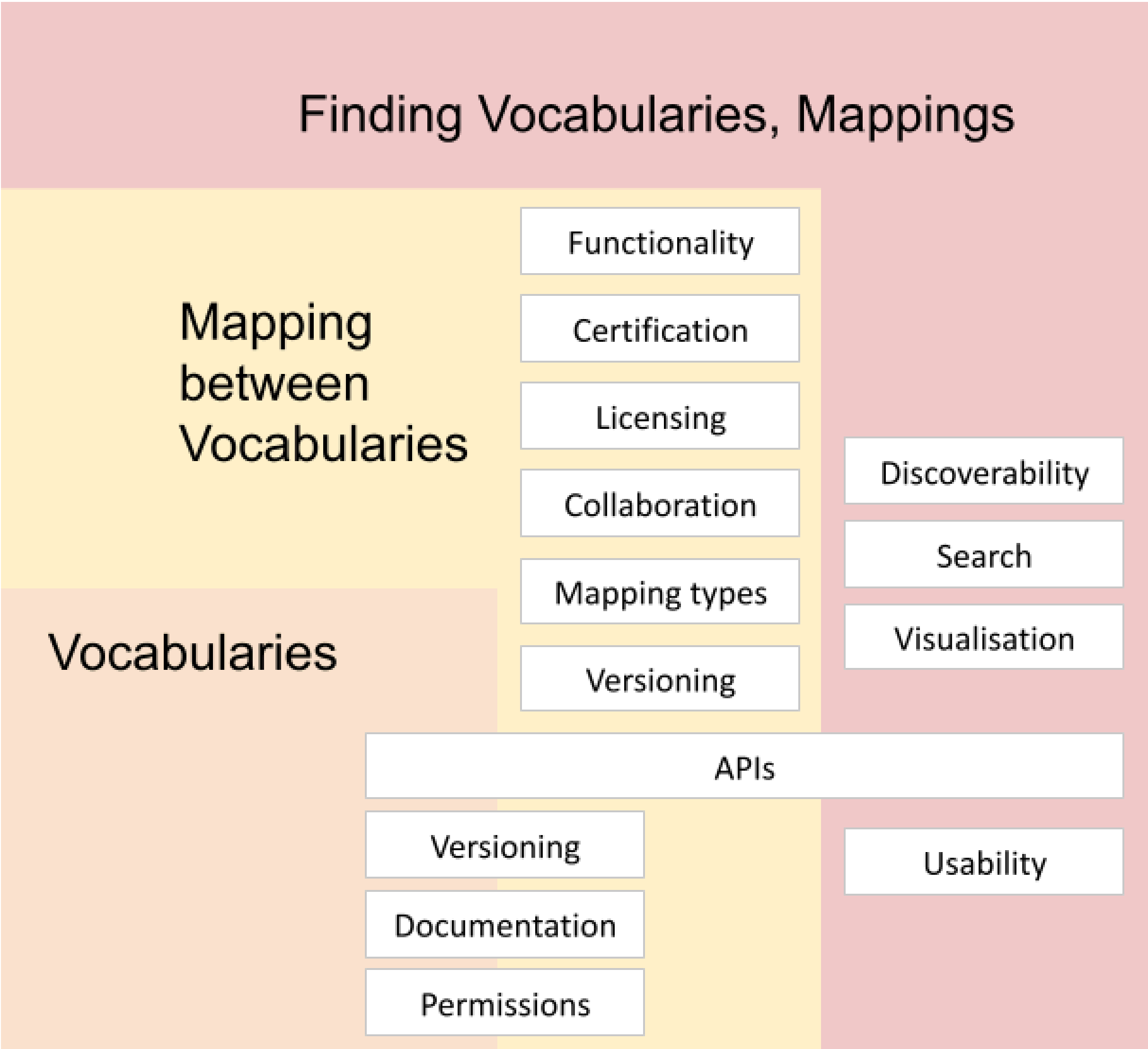


Figure 2: Focus requirements for a technology system

The order of these is not related to community priorities or needs.

1. Functionality

- Allow users to define criteria for determining mapping "closeness" between terms, including context
- Enable comparing source and target vocabularies to identify similarities/differences for mapping
- Support selecting preferred hierarchy levels in the target vocabulary
- Provide bi-directional mapping capabilities between vocabularies
- Offer a simple interface for rapid review, acceptance, or rejection of new mappings
- Allow viewing mapping deltas between users, including hierarchy alignment
- Extract subsets of vocabularies and mappings for specific use cases
- Configurable tool with scripting capabilities to extend functionality

2. Governance & Licensing

- Mechanisms for sharing and collaborating on mappings
- Controls for sharing term lists with selected groups

3. Discoverability & Integration

- Search capabilities to find how others have mapped common datasets
- Allocate DOIs to finalised vocabulary versions and mappings
- Centralised dictionary of available mappings and their purpose
- Support standards like SKOS, FHIR, OWL for integration

4. APIs

- API for checking current vocabulary/mapping versions
- Check for updates to vocabularies/mappings via API
- Import/export mappings via API
- Update mappings and vocabularies
- User authentication and access control
- Convert between different vocabulary formats

5. Automation

- Automated mapping suggestions based on closeness algorithms

- Machine learning to improve auto-mapping from accepted mappings

- Handle mappings for patient-reported outcomes data

- Configure output formats like JSON, OWL

- Leverage mapping conventions to provide automated suggestions

6. Collaboration

- Centralised, live-updated source for multiple curators

- Make individual user mappings available to the community

- Consolidate local mapping copies into centralised version

7. Versions & History

- Version control system (e.g. Git) for managing mapping changes

- Map between different versions of the same vocabulary

- Roll back and forth through vocabulary/mapping versions

- Maintain full mapping history and provenance

8. Documentation

- Capture mapping rules, guidelines and context used

- Annotate mappings with relevant documentation

9. Permissions

- Manage user roles and access controls for mappings

- Privacy controls for individual/institutional mappings

10. Search

- Fuzzy term search for mapping

- Handle abbreviations, misspellings, mixed capitalization

- Use of natural language for search

11. Certification & Verification

- Workflows for expert verification of mappings

- Workflows for curating, validating, and approving new terms/definitions

- Maturity model tracking mapping quality/verification status

12. Mapping Types

- Support mapping between different databases/vocabularies

13. Interoperability

- Integrate with existing mapping formats and conventions

- Handle mapping conventions beyond 1:1 term mappings

14. Usability

- User-friendly interface for mapping activities

- Optimised for different user roles (e.g. curators, reviewers)

15. Visualisation

- Data visualisations to aid mapping understanding

- Visualise similarities, deltas, hierarchies between mappings

Metadata as an Enabler

The workshop participants feedback emphasised the importance of trusted mappings of vocabularies. The technical and community features they raised could be addressed through metadata that encodes:

- Reviews and feedback from a trusted reviewer and/or from mapping users
- Authenticated individual authors and curators
- Links to DOIs to see number of views, downloads, citations
- Terms and categorisations that improve discoverability

Drawing on the vocabulary mapping lifecycle illustrated earlier in this report in figure 1, it can be further explored in relation to the lifecycle of metadata as shown in figure 3 below.

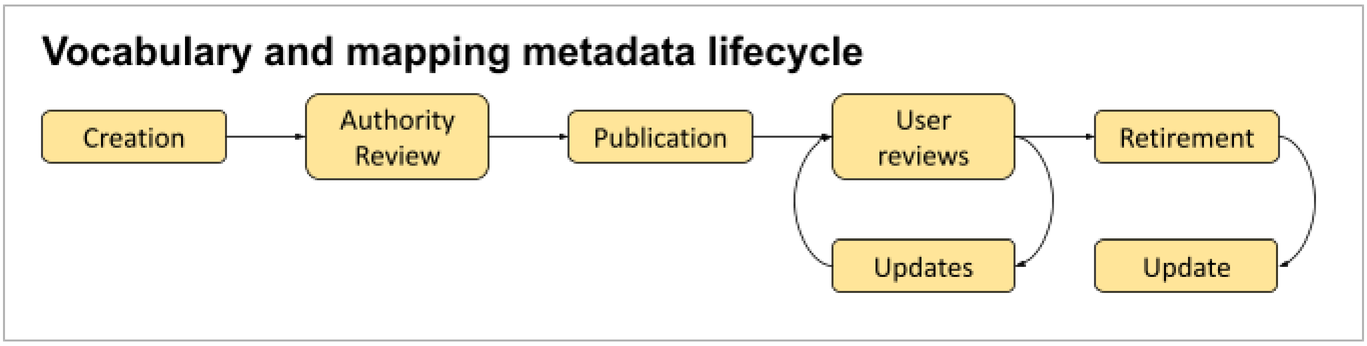


Figure 3: Lifecycle of vocabulary and mapping metadata

Metadata is invaluable as it enables the tracking of provenance and provides detailed information about mapping resources. See the example below:

An Example Metadata Record for General Practice Free Text to SNOMED Mapping

Title: Mapping of General Practice Free Text Medical Terms to SNOMED CT
Description: A comprehensive mapping of free-text medical terms commonly used in general practice to their corresponding SNOMED CT concepts, designed to enhance data standardisation and interoperability in healthcare research and clinical systems.
Keywords: General practice, free-text terms, SNOMED CT, medical terminology, mapping, interoperability, healthcare data, standardisation
Author(s): John Smith
Organisation: University of Technology
Date Created: [DD-MM-YYYY]
Version: [Version Number]
Format: CSV, JSON, XML
Language: English
License: [License Type, e.g., CC BY 4.0, GNU GPL, etc.]
Funding Source: [If applicable]
DOI: EG00252789
Intended Use: To facilitate the standardisation of free-text medical data for research, analysis, and clinical interoperability.
Coverage: General practice terms in Australia
Methods: Manual and automated term mapping processes, verified by domain experts.
Limitations: Limited to terms commonly used in general practice
Contact Information: John.Smith@UOT.edu.au

Future trends

It was noted throughout the workshop that AI, in particular Large Language Models (LLMs), are having an impact in this space, and the participants highlighted some areas it may have an impact on. While there are some aspects that will have a role in the vocabulary lifecycle, and will be discussed later, there was one aspect that was raised frequently and is highlighted below.

AI transcription

AI-powered clinical transcription is a maturing area that is becoming more deeply integrated into healthcare workflows. While there are challenges around ensuring consistency, accuracy, and appropriate human oversight that need to be addressed, the workshop brought focus to the ways that a mapping tool could work with clinical transcription technologies, while recognising that AI transcription tool integration is a longer term goal that will rely on governance structures for quality control and interoperability standards for greater integration of data sources.

Automated transcription tools are likely to become increasingly integrated into clinical workflows. As these tools mature they could use authoritative, certified mappings to improve the accuracy of AI generated medical records.

Recommendations

This report weighs the expected time and effort against the value added when considering how to address the outcomes of the workshop. The intent is to prioritise 'low-hanging fruit' that can be quickly executed over things that may be achieved through AI in the future. To assist structuring future work derived from the workshop outcomes, we propose a staged approach with the timeline of individual activities to be set by priorities in consultation with community stakeholders and aligned with the ARDC's medical research data integration framework strategy and initiative. The ten key functional requirements outlined in this report should be used for consultation on priorities on an ongoing basis.

This staged approach starts with manual mechanisms using existing tools to allow strategies such as the ARDC Implementation Framework to proceed without technology development acting as a bottleneck so work can proceed in parallel.

A staged approach

A staged, incremental approach would institute lightweight, strategic changes that could be incrementally refined and improved through early feedback from key stakeholders. A possible staged plan is outlined below:

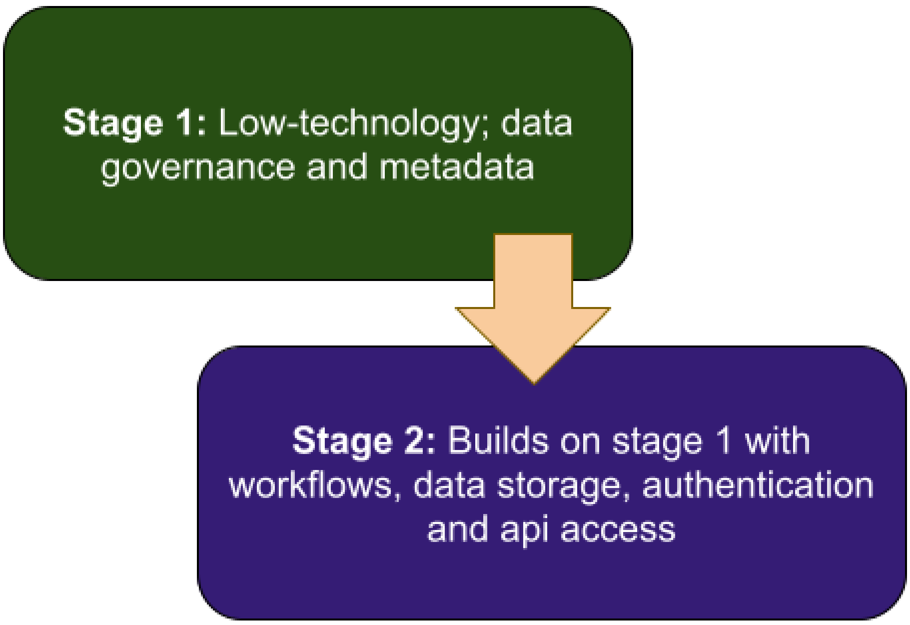


Figure 3: A Staged Approach

Stage 1: Prioritise lightweight, low-technology governance and metadata enablers

1. Establish a lightweight, fit-for-purpose governance mechanisms for collaboration, curation, validation and access controls that could work with commonly used data formats such as excel files Build an initial registry of mappings to uplift the FAIR principles in the community
2. Address metadata as an enabler for findability and conveying trustworthiness of available mappings. Seek consensus around standards for adoption of metadata with incremental development and consultation.
3. Create sample mappings using the metadata standards to improve adoption of vocabulary mappings.

Stage 2: Builds on stage 1 for data storage design and to extend technology enablers

4. Design a modular, configurable architecture with a constrained API featureset that allows extending functionality through scripting and integration of new mapping automation approaches like large language models. Develop minimum viable product for technology solution to uplift discoverability and findability of existing mappings.
5. Extend and refine the governance structures and metadata standards for:
    a. sharing, collaborating on, and discovering mapping
    b. licensing, copyright/usage, credit/acknowledgement
6. Extend API featureset with the refined metadata standards

Key Considerations for a Staged Approach

The proposed staged approach executes an action plan to address the challenges, pain points and opportunities for vocabulary mapping through community driven governance with an initial lightweight technology architecture and open-source solution.

Role of AI in the solution

Based on the discussion and an examination of future trends, AI- specifically LLMs, have a role in the future of the solution. AI is well placed for doing specific tasks, such as search, creating a draft mapping, and initial ETL processes. Given the pace of development in these technologies, there is a questionable need to specifically build such capability. As an example, the future solution may allow the user to choose the best LLM for a specific mapping functionality. It is expected that different LLMs will work for specific applications and domains. For example, the mapping of medications may need a different LLM compared to mapping of procedures or the reason for visit.

The promise of AL/LLMs for search and initial mappings is compelling. During the workshop discussion it was acknowledged AI/LLMs are a long way from being able to do more complex tasks such as controlling or managing governance and licensing. However, AI/LMM do have limitations which need to be taken into account when incorporating AI/LLM into the technology solution, such as training bias and production of incorrect information, or more commonly phrased as hallucinations.

Technology and Governance

The stages for governance would build on existing practices and connect with key communities to establish oversight of shared mappings and vocabularies. This includes mechanisms for collaboration, curation, validation and access controls.

The architecture would allow the governance and technology solution to incrementally deliver and integrate with strong automation capabilities leveraging techniques like machine learning, closeness algorithms and mapping conventions to provide high-quality automated mapping suggestions that learn and improve over time. The technology solution would incrementally develop to implement integration with existing standards (SKOS, FHIR, OWL etc.) and support different mapping formats to ensure interoperability across systems. Consultation and review of related ISO and other international standards to inform the ecosystem development is necessary.

The form of the governance and technology will be shaped by priorities raised through the workshop but also by budget considerations.

Data Design

A standardised location for data storage that is FAIR, compliant with interoperability and vocabulary standards, needs to be addressed. Through a staged approach, data storage may shift through the stages from an initial form such as using Ontoserver's hierarchical vocabulary and JSON FHIR files. Through stages 1 into 2 the data storage would be transitioned to a more suitable platform noting ISO and international standards considerations above. A modular, configurable architecture could support this staged approach where the data layer and presentation layer can develop at different speeds because the interface between them is abstracted and clearly defined. An architecture of maturing technologies could include blockchain controls for the use of vocabularies and vocabulary mappings through smart contracts.

Appendix A: Glossary of Terms

API access	An application programming interface (API) is a set of protocols, routines, and tools for building software applications. It specifies how software components should interact with each other.
Athena	An online terminology lookup tool provided by OHDSI
Australian Access Federation (AAF)	An authentication federation that enables single sign-on access to online resources for the Australian research and education community.
CSIRO	Commonwealth Scientific and Industrial Research Organisation
EHDEN	The European Health Data and Evidence Network partnership that provides a portal for accessing OMOP databases.
ETL	Extract, transform and load process for transforming electronic medical records into analysis ready datasets
FIHR	Fast Healthcare Interoperability Resources a standard for exchanging medical information (not just vocabularies)
ISO/TS 21564:2019	International standard for maps between vocabulary systems with established quality and usefulness requirements for mappings.
JSON	Javascript Object Notation is a text based format for hierarchically organised data that is both human-readable and machine-parsable
Logical Observation Identifiers Names and Codes (LOINC)	A universal code system for identifying medical laboratory observations.
Metadata	Data that provides information about other data that improves findability. Metadata includes the definition, structure, and provenance of data elements.
Natural Language Processing (NLP)	Techniques to analyse, understand, and generate human language, including the processing of clinical narratives and free-text data.
Observational Health Data Sciences and Informatics (OHDSI)	Observational Health Data Sciences and Informatics (or OHDSI, pronounced "Odyssey") program is a multi-stakeholder, interdisciplinary collaborative to bring out the value of health data through large-scale analytics
OMOP CDM	Observable Medical Outcomes Partnership (OMOP) Common Data Model (CDM) international standard for harmonising health data

OntoServer	A terminology server developed by CSIRO, providing programmatic access to clinical terminologies.
SKOS	Standard used frequently in non-medical vocabularies
Usagi	A terminology mapping tool developed by the Observational Health Data Sciences and Informatics (OHDSI) group

Appendix B: Notes on Workshop and Analysis

Workshop Structure

The workshop was held on Thursday 7th November 2024 with participants attending online over zoom.

First was a presentation to the whole group of the ARDC strategic drivers and a walkthrough of the current ecosystem of tools.

Next participants were asked to consider:

1. What are the pain points?
2. What should the tool include?
3. What are the opportunities to enable progress?

To tackle this the workshop split into breakout rooms with each room considering one of the three focus areas:

1. Term lists
2. Custom vocabularies
3. Mapping between vocabularies

Followed by a final large group discussion and feedback from each of the breakout rooms.

Note, the whole workshop was recorded and stored on the ARDC google drive, for future reference.

Methodology - Notes on analysis

Participants and scribes generated ‘stickies’ of textual comments through the workshop discussion, collectively and in small breakout groups, onto a collaborative Miro board. The original arrangement by the breakout groups is shown in figure 3 below.

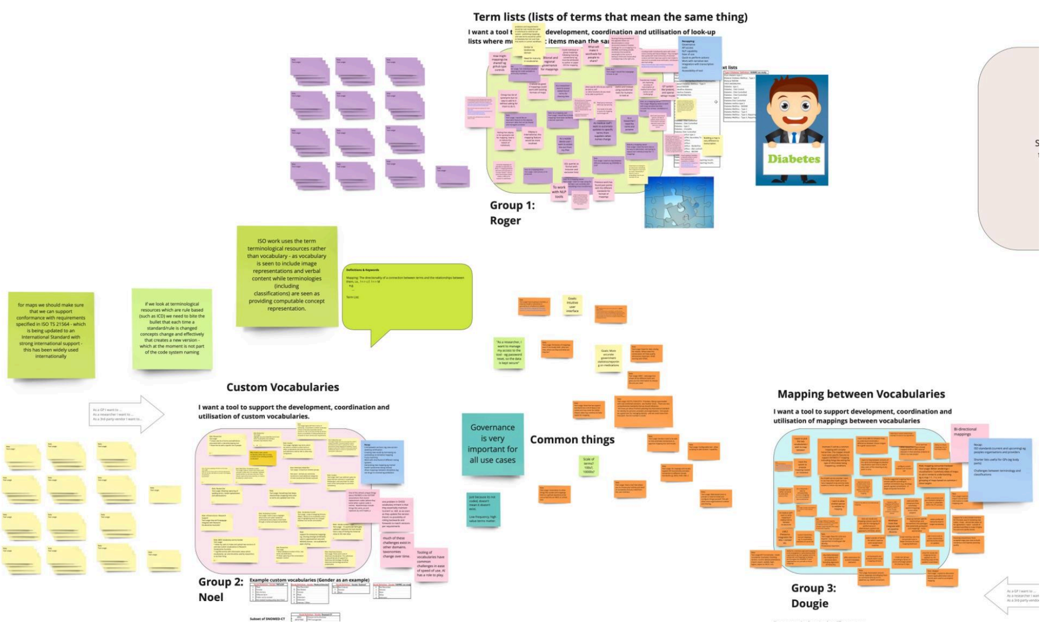


Figure 3: Snapshot of miro board at the end of the workshop

The stickies were organised into clusters (frames) and downloaded as a csv and converted to a well formed json file, and this text was queried using SparkAI’s implementation of Claude anthropic.