A Methodology for Progressively Developing and Harmonising FAIR Vocabularies for Global Interoperability of Geochemical Data

Lesley Wyborn, Australian National University, ORCID; Marthe Klöcking, Göttingen University, Germany, ORCID; Alexander Prent, AuScope, ORCID; Lucia Profeta, Columbia University, USA, ORCID; Angus Nixon The University of Adelaide, ORCID; Kirsten Elger, GFZ German Research Centre for Geosciences, ORCID; Manja Luzi-Helbing, GFZ German Research Centre for Geosciences, ORCID; Rowan Brownlee, ARDC, ORCID; Steve Richard, US Geoscience Information Network Foundation, ORCID; Rebecca Farrington, AuScope, ORCID; Kerstin Lehnert, Columbia University, USA, ORCID; and Dominik Hezel, Frankfurt University, Germany, ORCID.

Co-funded by the European Union

COMMITTEE ON DATA CODATA INTERNATIONAL SCIENCE COUNCIL

AuScope

We acknowledge and celebrate the First Australians on whose traditional lands we meet and pay our respect to the Elders past and present.

Abstract

OneGeochemistry is an international initiative formed to enable global sharing of geochemical data. Unfortunately geochemical datasets are notoriously heterogeneous and are collected by thousands of researchers/research groups on a diversity of samples (rocks, minerals, meteorites, fluids, gasses, etc) using hundreds of analytical techniques across multiple geoscience disciplines. Hence, achieving international consensus on key concepts and definitions requires considerable time and effort.

To ensure compliance with FAIR Interoperability Principle I2 (viz. (meta)data to use controlled vocabularies that also follow FAIR principles), multiple local vocabularies are emerging online that often replicate similar concepts.

To achieve a balance between meeting current demands for FAIR-compliant vocabularies versus time required to reach international agreement, OneGeochemistry has developed a three-tiered approach (local, community, international) towards making semantic resources FAIR and machine actionable:

1) We encourage data providers with locally defined vocabularies or other resources to make them FAIR, available online from a reputable vocabulary service and ensure each term has a persistent identifier (SKOS/RDF);
2) We encourage groups with similar topics to begin harmonising on concepts/definitions and publish these as community resources;
3) We raise awareness of groups harmonising and making semantic resources FAIR-compliant at an international level, particularly those with endorsement from International authoritative groups (e.g., Scientific Unions/Associations/Societies/Commissions).

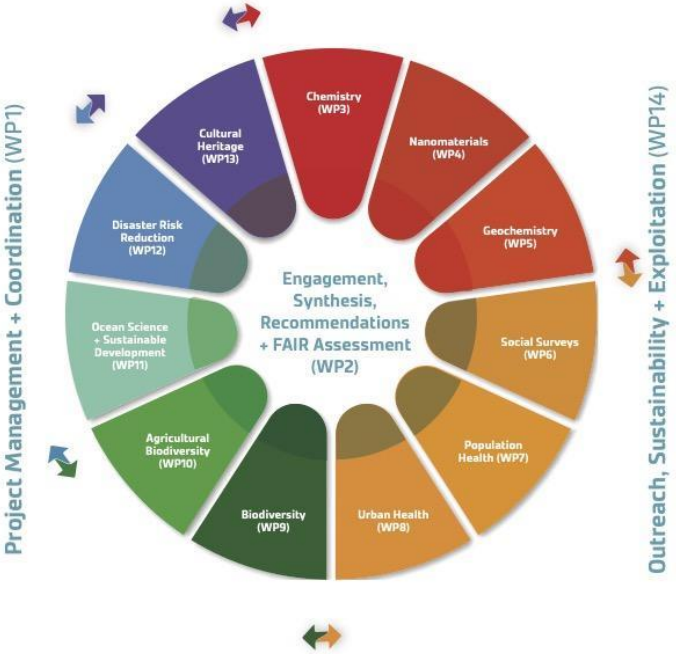
As convergence takes place towards internationally-agreed terms, the size of the community able to share (meta)data in machine-to-machine environments grows. Existing URIs initially used at either a local or community level could be redirected towards internationally endorsed definitions and concepts as these become available.

What is OneGeochemistry?

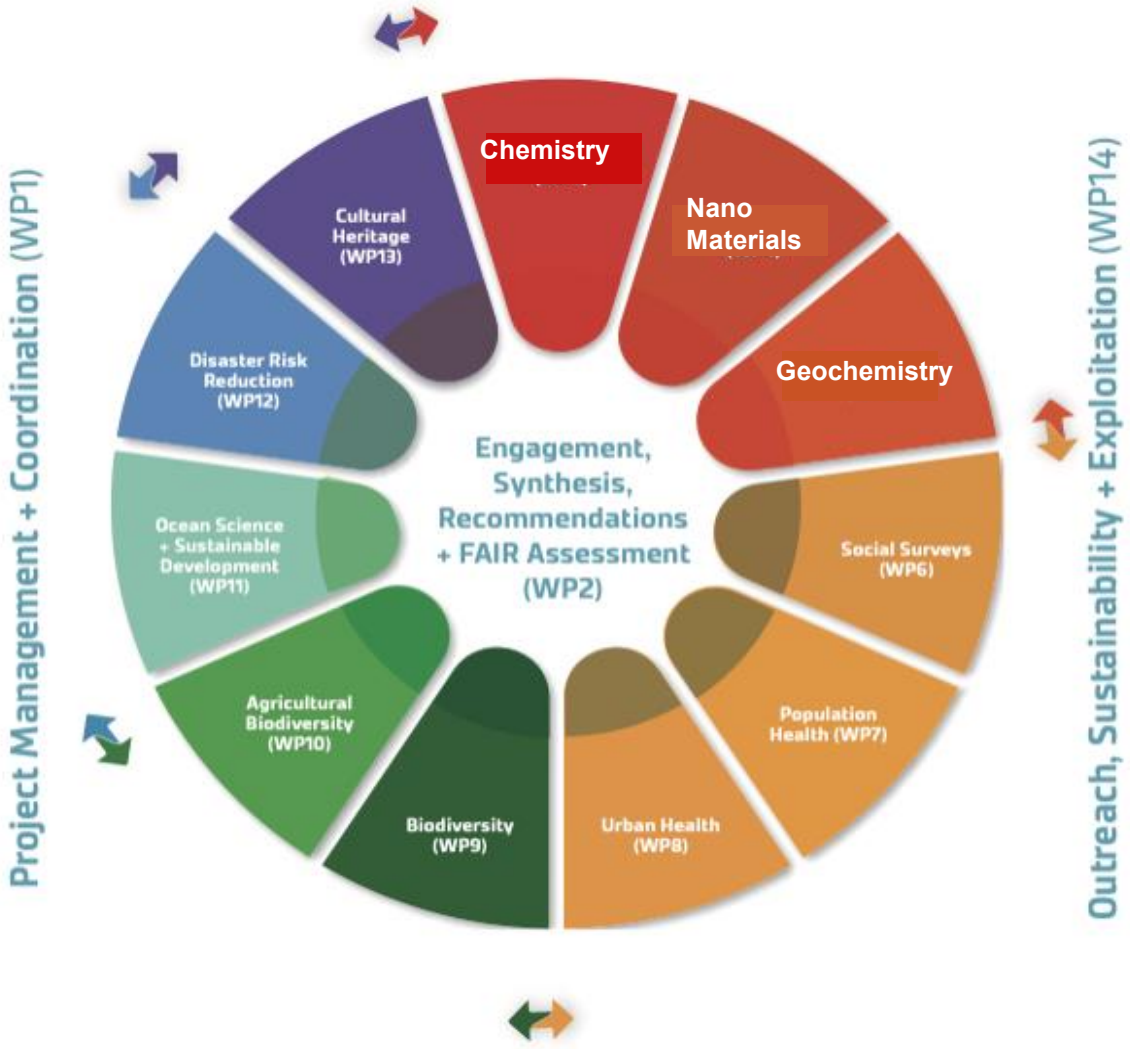
- Collaboration of organizations that support the acquisition, management, and access of geochemical data
- To empower the reuse of geochemical data for the advancement of scientific knowledge and discovery by building and maintaining consensus-driven data standards.

One Geochemistry is part of the EU funded Project WorldFAIR: Global cooperation on FAIR data policy and practice

- Funded by the European Union, HORIZON-WIDERA-2021-ERA-0 — Project: 101058393.
- Two year project from 1 June 2022.
- Nineteen partners from France, Belgium, Cyprus, Denmark, Germany, UK, Ireland, Norway (Europe); Kenya (Africa); **Australia**, New Zealand (Oceania); Brazil (Sth America); USA (Nth America).
- Project contributes to:
  - UNESCO Recommendation on Open Science
  - CODATA-ISC Decadal Programme
  - ISC Action Plan Project 2.1: 'Making Data Work for Cross-Domain Grand Challenges:
- Is based around 14 Work Packages, including 11 case study WPs
- Is doing pioneering work in FAIR implementation profiles to assist in machine-to-machine interoperability



The WorldFAIR project



The 11 WorldFAIR case studies are:

1. Chemistry
2. Nanomaterials — Theme 1
3. **OneGeochemistry**
4. Social Surveys Data
5. Population Health — Theme 2
6. Urban Health
7. Biodiversity
8. Agricultural Biodiversity — Theme 3
9. Ocean Science
10. Disaster Risk Reduction — Theme 4
11. Cultural Heritage → Theme 5

- Exploring features of a Core Interoperability Framework with 11 case studies from a range of research areas Working at extracting the common definitions (Units, vocabularies, data description, data structure, provenance…) across 11 case studies

Geochemical Data Desperately Needs Standardisation

Unclear units stymie science

Nature Comments (2022),
doi.org/10.1038/d41586-022-01233-w

Robert Hanisch, Stuart Chalk, Romain Coulon, Simon Cox, Steven Emmerson, Francisco Javier Flamenco Sandoval, Alistair Forbes, Jeremy Frey, Blair Hall, Richard Hartshorn, Pascal Heus, Simon Hodson, Kazumoto Hosaka, Daniel Hutzschenreuter, Chu-Shik Kang, Susanne Picard & Ryan White

Here's how to make measurements clear and machine-readable.

In 1999, when NASA's Mars Climate Orbiter missed its intended orbit and burned up in the Martian atmosphere, the media had a heyday over the reason: one team had used metric units in its thrust calculations, another, imperial. The navigation software that exchanged this information lacked a built-in process to check units. So when one team's software produced data in imperial units rather than the expected metric ones, the spacecraft was set on the wrong trajectory. The result was the loss of five years of effort and hundreds of millions of taxpayers' dollars.

see www.go-fair.org/fair-principles), and to ensure that open data abide by the 5-star deployment scheme suggested by World Wide Web inventor Tim Berners-Lee, which aims to make them findable, free and structured (see https://5stardata.info/en). Many researchers are now committed to depositing data in free and open repositories with appropriate metadata.

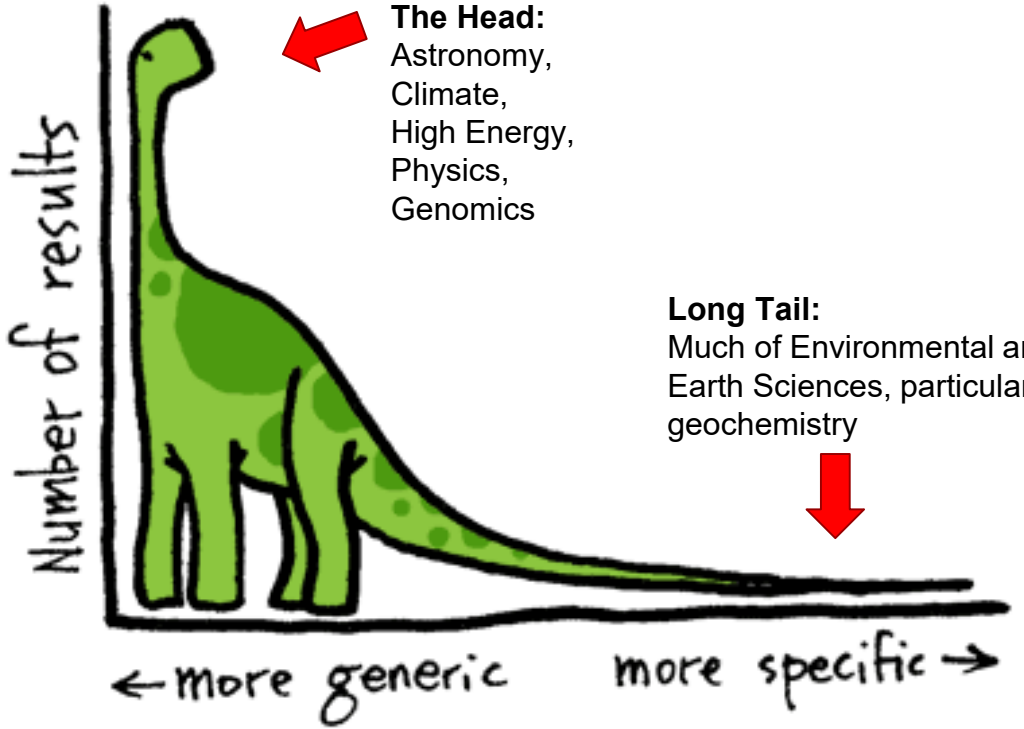
Chaos around units undermines these efforts. Already, many scientists invest more time in wrangling data than doing research. When data are not interoperable or machine readable, researchers' individual informatics approaches are thwarted. The benefits of data sharing shrink.

Unless we take steps to ensure that measurement units are routinely documented for easy, unambiguous exchange of data, information will be unusable or, worse, be misinterpreted. All global chal-

Fewer Mistakes, Failures, Catastrophes

No single authority: there are 5 Unions and ~40 Societies

Geochemistry Data is LongTail and Hence Difficult



The Head:
Astronomy,
Climate,
High Energy,
Physics,
Genomics

Long Tail:
Much of Environmental and Earth Sciences, particularly geochemistry

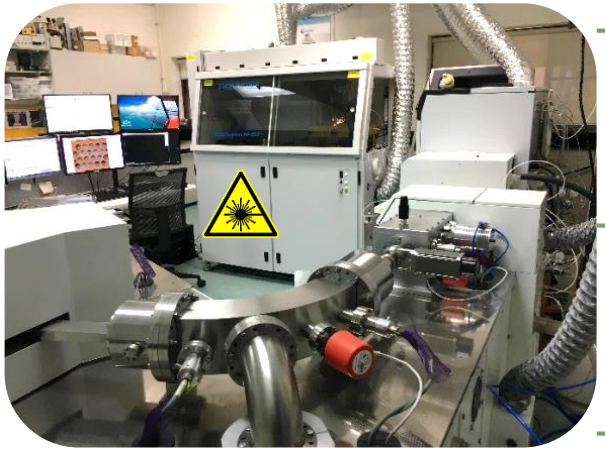
http://juliegood.wordpress.com/tag/long-tail

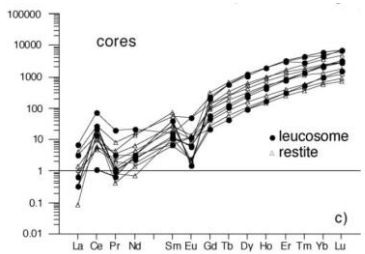
Long Tail Characteristics

- ❑ More specialised
- ❑ Low volume
- ❑ On C drives
- ❑ Hard to find
- ❑ Heterogeneous
- ❑ Collected by many people
- ❑ Citizen science
- ❑ Etc…

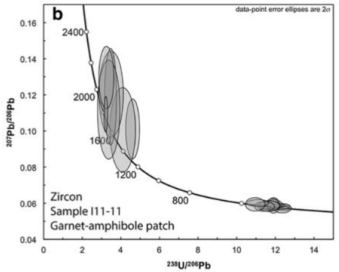
It has many Sub-disciplines

A ‘Long Tail’ community with many subdisciplines, highly specific and small size datasets

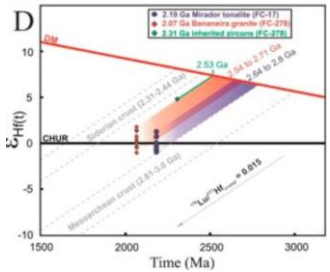




Trace elements



Uranium and lead isotopes



Lutetium and hafnium isotopes

https://upload.wikimedia.org/wikipedia/commons/7/7b/OSIRIS-REx_spacecraft_model.png

Wide Variety of Analytical Methods

https://www.axios.com/2023/10/11/osiris-rex-asteroid-sample-analysis

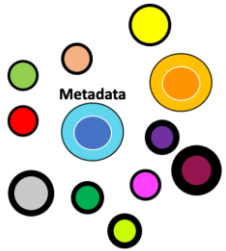
64 analytical methods will be used to study the OSIRIS-Rex returned samples

EMPA	Raman	XCT	VLM	QRIS	GC-MS	LC-MS	VNMIR	NanoSIMS	SLS
μL^2MS	FTICR-MS	SS-NMR	GC-C-IRMS	NMR	MC-ICP-MS	EA-IRMS	SIMS	XRD	SEM/FIB-SEM
TEM	EBSD/TKD	XANES	XPS	HR-ICP-MS	SHRIMP	LAF	APT	TIMS	NI-MI
Q-ICP-MS	FINESSE	NG-NS-MS	ICP-OES	GPYC	SCBTCA	DSC	HR-CL	EDS	EELS
NanoIR	S-XRF	TGA	LA-ICP-MS	NI-NGMS	RI-TOF-NGMS	DESI-Orbitrap	SThM	PCD-AFM	PSFD
XRF	ARGT	SNMS	AMS	COMPT	SV-RUEC	CAPD	DSSM	LIT	ARM
IC	ToF-SIMS	CE-MS	S-IR						

At a cost of $US 1.16 Billion – but all data is required to be FAIR

First steps – How?

Community



Communication

?

! controlled vocabularies !

Enabled

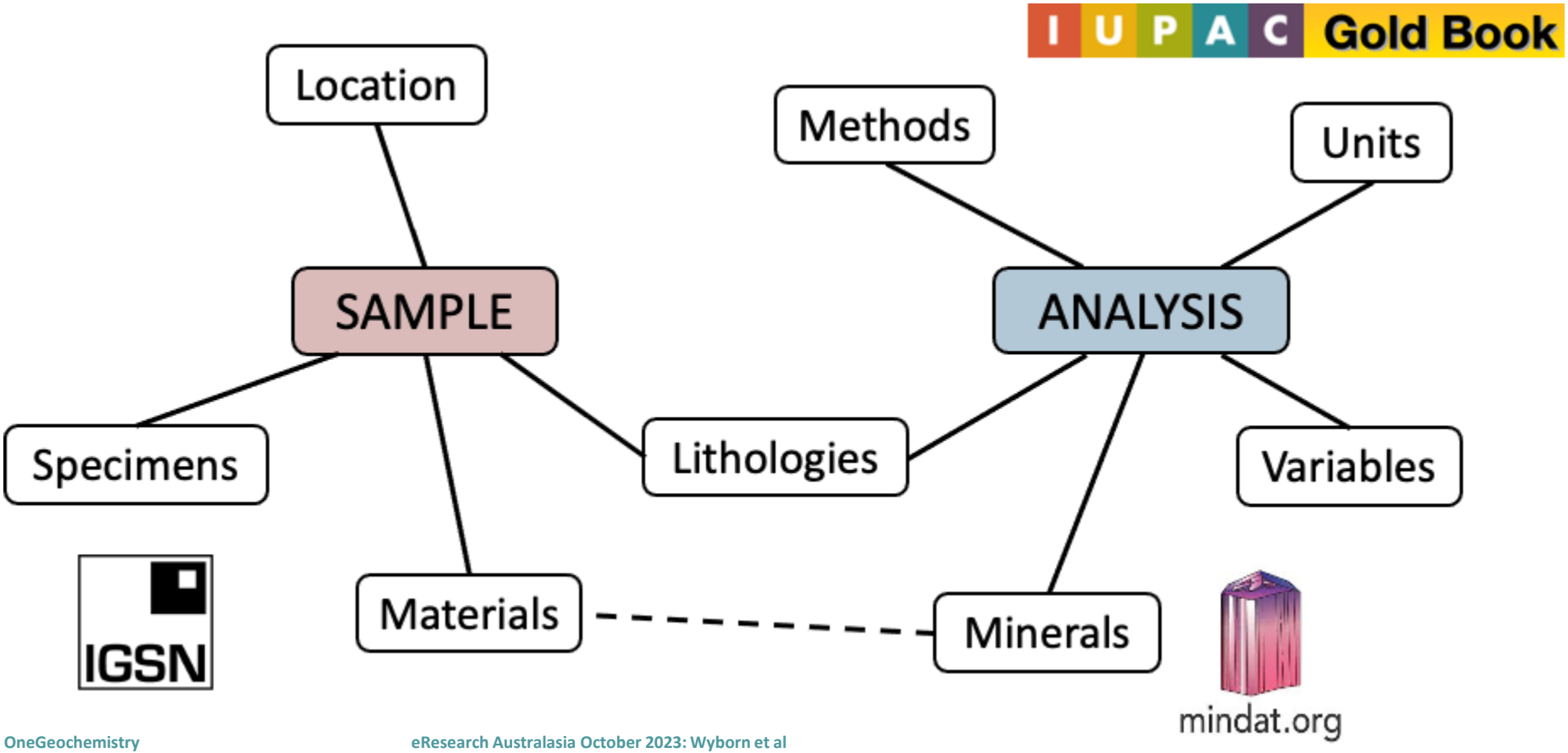
Remember: FAIR means Machine actionable

- FAIR means both human and machine readable
- To ensure compliance with FAIR Interoperability Principle I2 (viz. (**meta)data to use controlled vocabularies that also follow FAIR principles**), multiple local vocabularies are emerging online that often replicate similar concepts.
- People are creating many local vocabularies

In Summary, the Challenge is Diversity and unFAIR

- Variety of data types
  - Variables
  - Methods
  - Formats
- Variety of science
  - Samples
  - Context
  - Communities

But although separate, Vocabularies can be connected



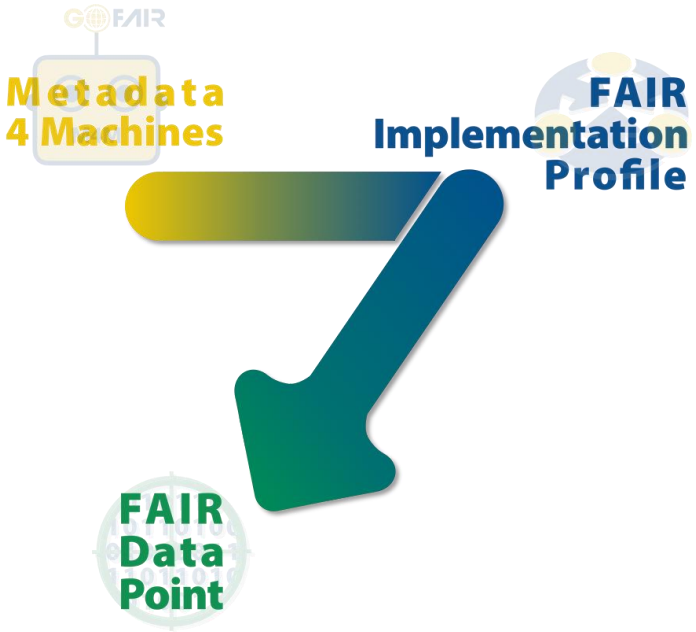
Methods Toward FAIR – FAIR Implementation Profiles

FAIR principle	Question	FAIR enabling resource types
F1	What globally unique, persistent, resolvable identifiers do you use for metadata records?	Identifier type
F1	What globally unique, persistent, resolvable identifiers do you use for datasets?	Identifier type
F2	Which metadata schemas do you use for findability?	Metadata schema
F3	What is the technology that links the persistent identifiers of your data to the metadata description?	Metadata-Data linking mechanism
F4	In which search engines are your metadata records indexed?	Search engines
F4	In which search engines are your datasets indexed?	Search engines
A1.1	Which standardized communication protocol do you use for metadata records?	Communication protocol
A1.1	Which standardized communication protocol do you use for datasets?	Communication protocol
A1.2	Which authentication & authorisation technique do you use for metadata records?	Authentication & authorisation technique
A1.2	Which authentication & authorisation technique do you use for datasets?	Authentication & authorisation technique
A2	Which metadata longevity plan do you use?	Metadata longevity
I1	Which knowledge representation languages (allowing machine interoperation) do you use for metadata records?	Knowledge representation language
[illegible]	Which knowledge representation languages (allowing machine interoperation) do you use for datasets?	Knowledge representation language
[illegible]	Which structured vocabularies do you use to annotate your metadata records?	Structured vocabularies
[illegible]2	Which structured vocabularies do you use to encode your datasets?	Structured vocabularies
I3	Which models, schema(s) do you use for your metadata records?	Metadata schema
I3	Which models, schema(s) do you use for your datasets?	Data schema
R1.1	Which usage license do you use for your metadata records?	Data usage license
R1.1	Which usage license do you use for your datasets?	Data usage license
R1.2	Which metadata schemas do you use for describing the provenance of your metadata records?	Provenance model
R1.2	Which metadata schemas do you use for describing the provenance of your datasets?	Provenance model

- FAIR means both human and machine readable
- To ensure compliance with FAIR Interoperability Principle I2 (viz. (meta)data to use controlled vocabularies that also follow FAIR principles), multiple local vocabularies are emerging online that often replicate similar concepts.

What are FAIR Implementation Profiles?

- Enable documentation that a given community makes about the implementation of each of the FAIR principles.
- For each FAIR Principle, the data provider ‘declares’ any resources used
- A community can be 1 person, it can be an international authoritative source (eg, Science Union)
- FIPs be created at the level of a repository, a data collection and an individual dataset
- Designed to be machine-actionable
- FIPs promote interoperability and standardisation



Schultes, E., Magagna, B., Hettne, K.M., Pergl, R., Suchánek, M., Kuhn, T. (2020). Reusable FAIR Implementation Profiles as Accelerators of FAIR Convergence. In: Grossmann, G., Ram, S. (eds) Advances in Conceptual Modeling. ER 2020. Lecture Notes in Computer Science, vol 12584. Springer, Cham. https://doi.org/10.1007/978-3-030-65847-2_13

FAIR enabling Resources

EARTHCHEM

DIGIS GEOROC

AusGeochem

FAIR principle		
F1	What globally unique, persistent, resolvable iden	
F1	What globally unique, persistent, resolvable iden	
F2	Which metadata schemas do you use for findabi	
F3	What is the technology that links the persistent i	
F4	In which search engines are your metadata reco	
F4	In which search engines are your datasets index	
A1.1	Which standardized communication protocol do	
A1.1	Which standardized communication protocol do	
A1.2	Which authentication & authorisation technique	
A1.2	Which authentication & authorisation technique do you use for datasets?	Authentication & authorisation technique
A2	Which metadata longevity plan do you use?	Metadata longevity
I1	Which knowledge representation languages (allowing machine interoperation) do you use for metadata records?	Knowledge representation language
I1	Which knowledge representation languages (allowing machine interoperation) do you use for datasets?	Knowledge representation language
I2	Which structured vocabularies do you use to annotate your metadata records?	Structured vocabularies
I2	Which structured vocabularies do you use to encode your datasets?	Structured vocabularies
I2	Which models, schema(s) do you use for your metadata records?	Metadata schema
I3	Which models, schema(s) do you use for your datasets?	Data schema
R1.1	Which usage license do you use for your metadata records?	Data usage license
R1.1	Which usage license do you use for your datasets?	Data usage license
R1.2	Which metadata schemas do you use for describing the provenance of your metadata records?	Provenance model
R1.2	Which metadata schemas do you use for describing the provenance of your datasets?	Provenance model

Which structured vocabularies do you use to annotate your metadata records?

Which structured vocabularies do you use to encode your datasets?

Which models, schema(s) do you use for your metadata records?

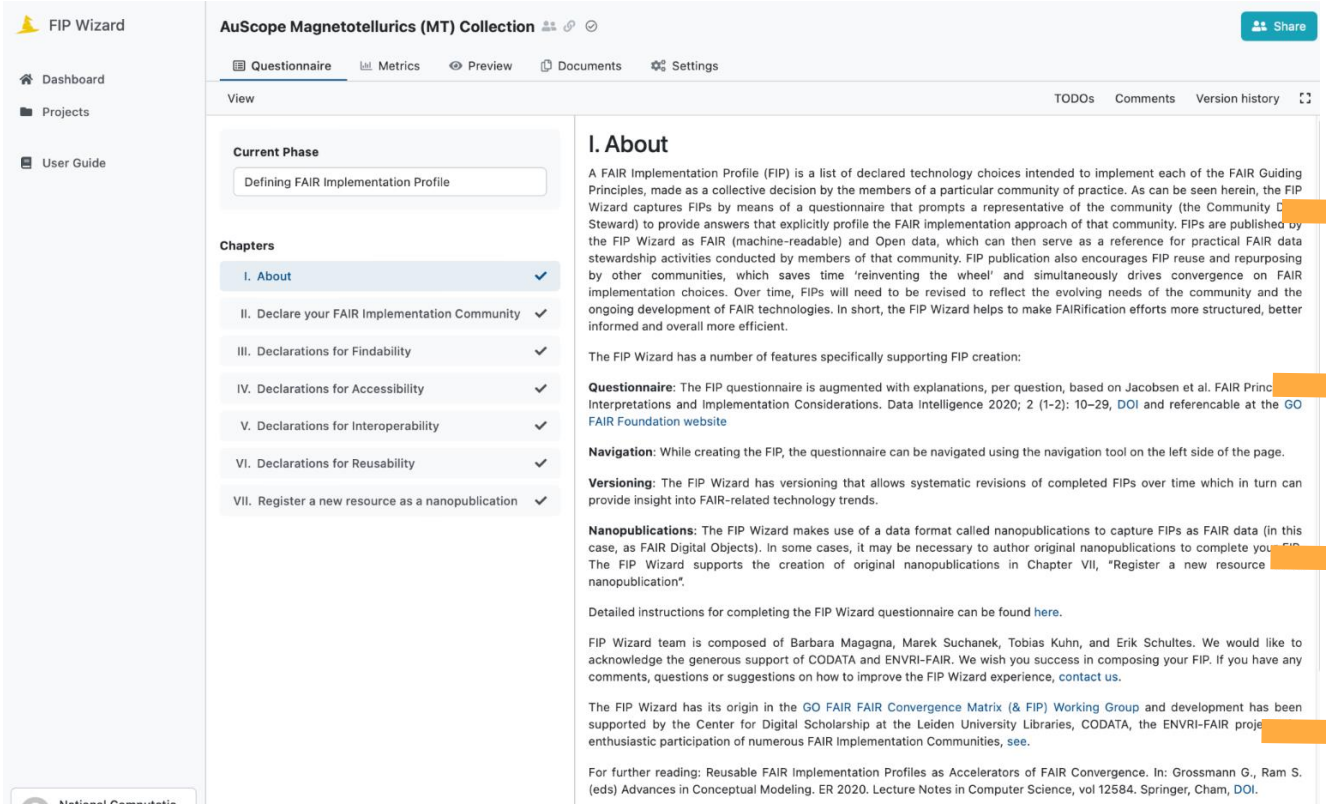
Which models, schema(s) do you use for your datasets?

FER = Structured Vocabulary (published)

Tools to make it easier: FIP Wizard

Slide thanks to Jo Croucher NCI!!!

See Jo's talk on Vertical Dataset Integration, Friday 1400 – 1420 Boulevard Auditorium



PDF

Excel

JSON

Other

FIP Wizard (2023): https://fip-wizard.ds-wizard.org/

FIPs 4 Repository / Platform / Database

EARTHCHEM

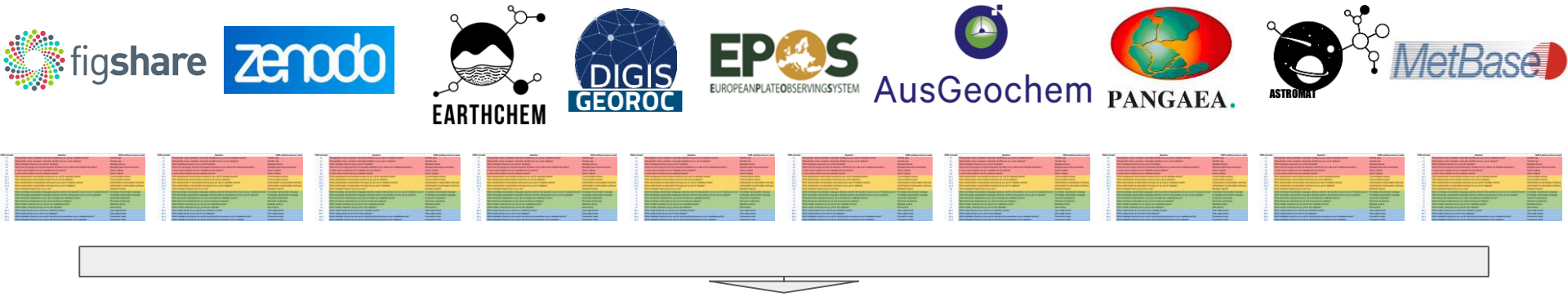
FAIR principle	Question	FAIR enabling resource types
F1	What globally unique, persistent, resolvable identifiers do you use for metadata records?	Identifier type
F1	What globally unique, persistent, resolvable identifiers do you use for datasets?	Identifier type
F2	Which metadata schemas do you use for findability?	Metadata schema
F3	What is the technology that links the persistent identifiers of your data to the metadata description?	Metadata Data linking mechanism
F4	In which search engines are your metadata records indexed?	Search engines
F4	In which search engines are your datasets indexed?	Search engines
A1.1	Which standardised communication protocol do you use for metadata records?	Communication protocol
A1.1	Which standardised communication protocol do you use for datasets?	Communication protocol
A1.2	Which authentication & authorisation technique do you use for metadata records?	Authentication & authorisation technique
A1.2	Which authentication & authorisation technique do you use for datasets?	Authentication & authorisation technique
A2	Which metadata longevity plan do you use?	Metadata longevity
I1	Which knowledge representation languages (allowing machine interoperation) do you use for metadata records?	Knowledge representation language
I1	Which knowledge representation languages (allowing machine interoperation) do you use for datasets?	Knowledge representation language
I2	Which structured vocabularies do you use to annotate your metadata records?	Structured vocabularies
I2	Which structured vocabularies do you use to encode your datasets?	Structured vocabularies
I3	Which models, schema(s) do you use for your metadata records?	Metadata schema
I3	Which models, schema(s) do you use for your datasets?	Data schema
R1.1	Which usage license do you use for your metadata records?	Data usage license
R1.1	Which usage license do you use for your datasets?	Data usage license
R1.2	Which metadata schemas do you use for describing the provenance of your metadata records?	Provenance model
R1.2	Which metadata schemas do you use for describing the provenance of your datasets?	Provenance model

FAIR principle	Question	FAIR enabling resource types
F1	What globally unique, persistent, resolvable identifiers do you use for metadata records?	Identifier type
F1	What globally unique, persistent, resolvable identifiers do you use for datasets?	Identifier type
F2	Which metadata schemas do you use for findability?	Metadata schema
F3	What is the technology that links the persistent identifiers of your data to the metadata description?	Metadata Data linking mechanism
F4	In which search engines are your metadata records indexed?	Search engines
F4	In which search engines are your datasets indexed?	Search engines
A1.1	Which standardised communication protocol do you use for metadata records?	Communication protocol
A1.1	Which standardised communication protocol do you use for datasets?	Communication protocol
A1.2	Which authentication & authorisation technique do you use for metadata records?	Authentication & authorisation technique
A1.2	Which authentication & authorisation technique do you use for datasets?	Authentication & authorisation technique
A2	Which metadata longevity plan do you use?	Metadata longevity
I1	Which knowledge representation languages (allowing machine interoperation) do you use for metadata records?	Knowledge representation language
I1	Which knowledge representation languages (allowing machine interoperation) do you use for datasets?	Knowledge representation language
I2	Which structured vocabularies do you use to annotate your metadata records?	Structured vocabularies
I2	Which structured vocabularies do you use to encode your datasets?	Structured vocabularies
I3	Which models, schema(s) do you use for your metadata records?	Metadata schema
I3	Which models, schema(s) do you use for your datasets?	Data schema
R1.1	Which usage license do you use for your metadata records?	Data usage license
R1.1	Which usage license do you use for your datasets?	Data usage license
R1.2	Which metadata schemas do you use for describing the provenance of your metadata records?	Provenance model
R1.2	Which metadata schemas do you use for describing the provenance of your datasets?	Provenance model

DIGIS GEOROC

https://img.freepik.com/premium-vector/gear-wheels…g?size=626&ext=jpg&ga=GA1.2.1254696273.1668346327

FIPs for each will indicate needed FERs



Comparing FAIR Implementation Profiles will clarify where:

- FAIR Enabling Resources are missing and need to be developed
- Crosswalks should be developed between existing FERs

Publishing FAIR Implementation Profiles will:

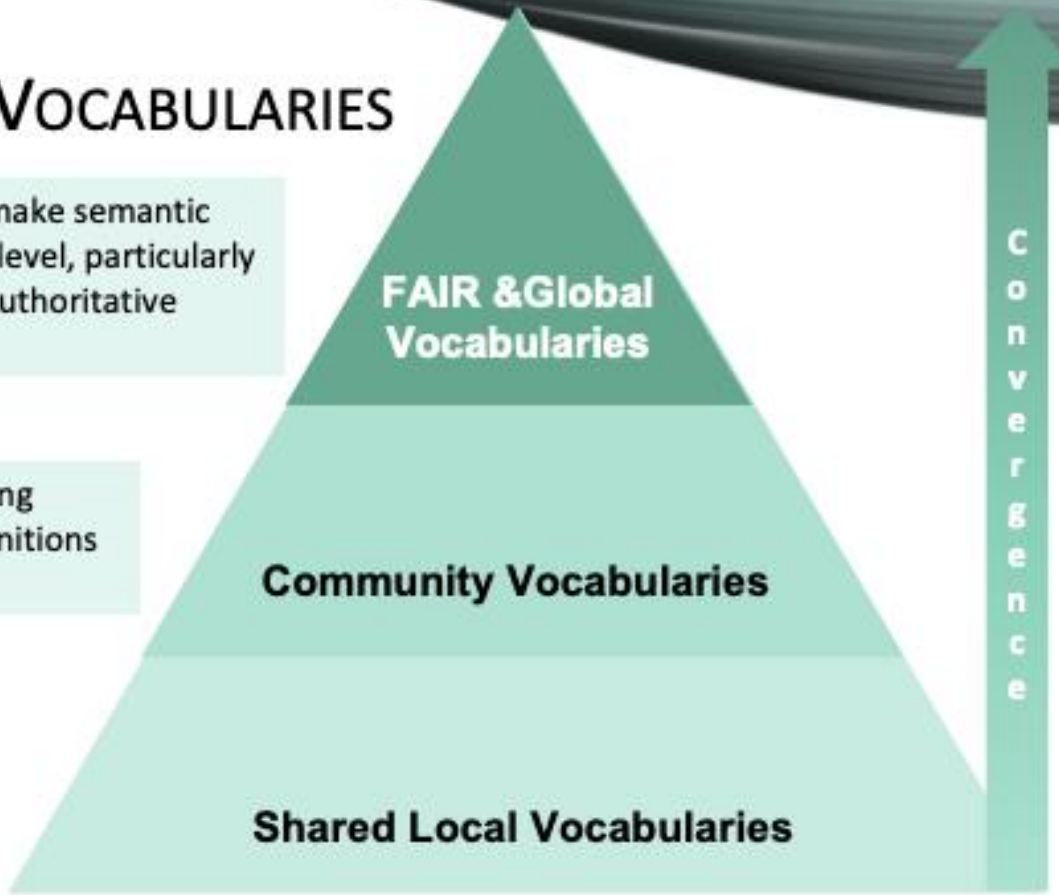
- Enable other (sub)disciplines to use and tailor to the FAIR Enabling Resource used by that discipline furthering cross domain interoperability.

3-Tiered Approach to Vocabularies

Raise awareness of groups harmonizing and make semantic resources FAIR-compliant at an international level, particularly those with endorsement from International authoritative groups (eg Scientific Unions/Societies).

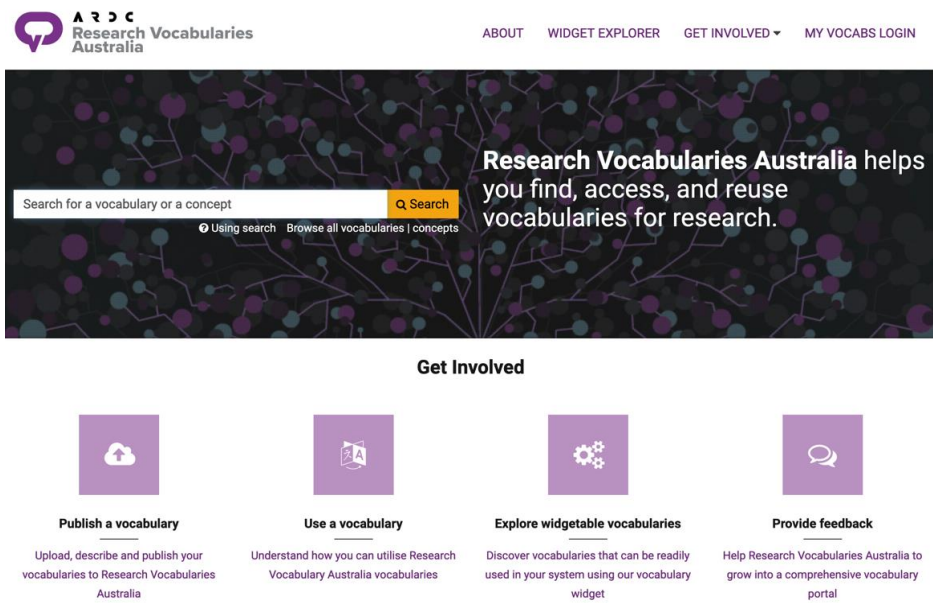
Groups with similar topics to begin harmonizing across multiple locally-derived concepts/ definitions and publish these as community resources;

Data providers with locally defined vocabularies to make them available online and ensure each term has a persistent ID;



Credit: Kerstin Lehnert

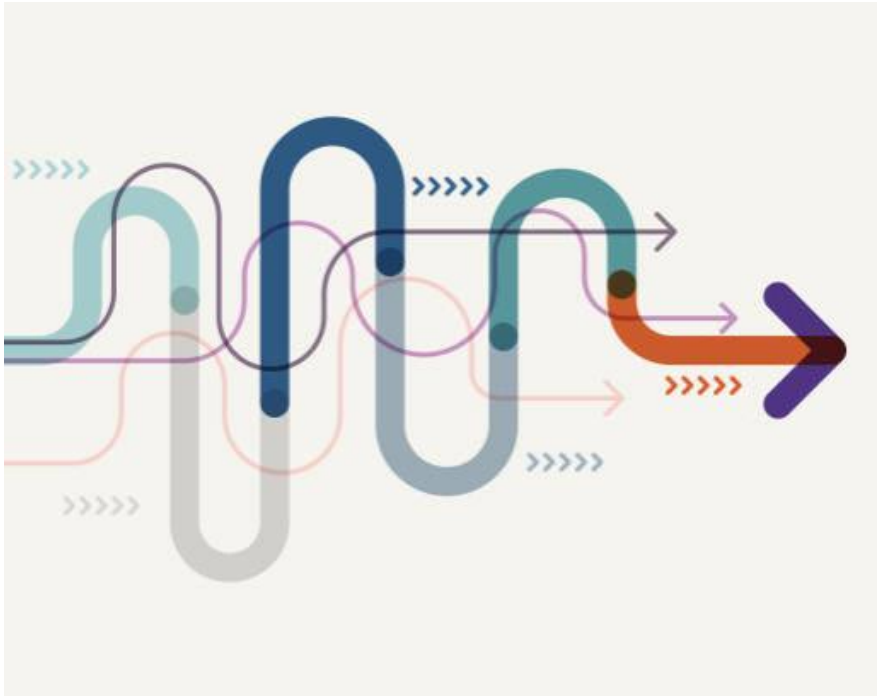
Specifying Locally: Harmonising Globally



- EarthChem, GEOROC, AGN, Astromat are starting to publish their vocabularies in Research Vocabularies Australia
- All terms will have URIs - for a specific vocabulary within a dataset, individual terms could come from multiple global sources.
- Terms can be redirected to new URIs as international vocabularies come online
- Each vocabulary profile can be registered as a FAIR enabling resource (FER)

In Conclusion: one Geochemistry is a work in progress

- Priorities are to expand availability of FAIR vocabularies, by leveraging ongoing projects and publishing what is available now.
- Continue to foster ‘community’ collaborations such as Astromat – DIGIS – EarthChem - AusGeochem to advance convergence.
- Gradually define & refine FAIR Implementation Profiles (FIPs) and FAIR Enabling Resources (FERs) for geochemical data at an international level.
- Work within the WorldFAIR community to enable the international geochemistry community to be part of global interdisciplinary science



Thank you FAIRy much

Webpage

alexander@auscope.org.au

Slack Channel