The Long Tale of Making Long Tail Geochemical Data FAIR, CARE and TRUST Compliant – Are We There Yet?

Lesley Wyborn[1], Marthe Klöcking[2], Kerstin Lehnert[3]

[1]Australian National University, Canberra, Australia,
[2]University of Münster, Münster, Germany,
[3]Lamont-Doherty Earth Observatory (Columbia University), Palisades, USA

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

Artist: Lynnice Letty Church – Tribes: Ngunnawal, Wiradjuri & Kamilaroi (ACT and NSW)
Gadi - "to search for" in Ngunnawal language - January 2020 for NCI Gadi Supercomputer

Abstract

Geochemical data is typical of Long Tail communities and characterised globally by small-sized, highly variable datasets mainly collected by individuals or small research teams. Geochemistry emerged as a discipline in 1838 and has evolved from low throughput, manual analytical techniques to the highly computerised laboratories of today that rapidly produce highly diverse geochemical and isotopic datasets on samples down to the atomic scale. Exponential increases in data volumes are challenging long-established practices and capabilities for organising, analysing, preserving, and sharing data. Increasing applications of machine learning techniques to large geochemical data compilations highlight the enormous value of the curation and harmonisation efforts undertaken by domain-curated data systems, which provide easy access to large volumes of high-quality, well-organised and standardised data.

Unfortunately, geochemistry as a discipline has been slow to change its methods of storing, publishing and sharing geochemical data and only transitioned to electronic publication methods around 2000. Most researchers managed their data locally on C-drives or on departmental servers. Modern data management is now a necessity for the discipline to thrive in the age of digital data and artificial intelligence, particularly as journals and funders now require the formal publication of datasets in repositories.

It has been a long tale to transform the long tail geochemistry community into modern ways of storing and curating data and making them compliant with the FAIR, CARE and TRUST principles. This paper will describe how this transition is taking place. Although focused on geochemistry, it is relevant to many other long tail communities.

Setting the scene: Geochemical Data

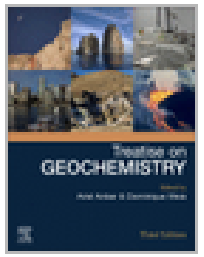
- Geochemical data is typical of Long Tail communities and characterised globally by small-sized, highly variable datasets mainly collected by individuals or small research teams.
- Geochemistry emerged as a discipline in 1838 and has evolved from low throughput, manual analytical techniques to the highly computerised laboratories of today that rapidly produce highly diverse geochemical and isotopic datasets on samples down to the atomic scale.
- Exponential increases in data volumes are challenging long-established practices and capabilities for organising, analysing, preserving, and sharing data.
- Increasing applications of machine learning techniques to large geochemical data compilations highlight the enormous value of the curation and harmonisation efforts undertaken by domain-curated data systems, which provide easy access to large volumes of high-quality, well-organised and standardised data.

Further Reading

ELSEVIER

Treatise on Geochemistry (Third Edition)

Volume 8, 2025, Pages 97-135



Geochemical databases

Marthe Klöcking [a], Kerstin A. Lehnert [b], Lesley Wyborn [c]

Show more

https://doi.org/10.1016/B978-0-323-99762-1.00123-6

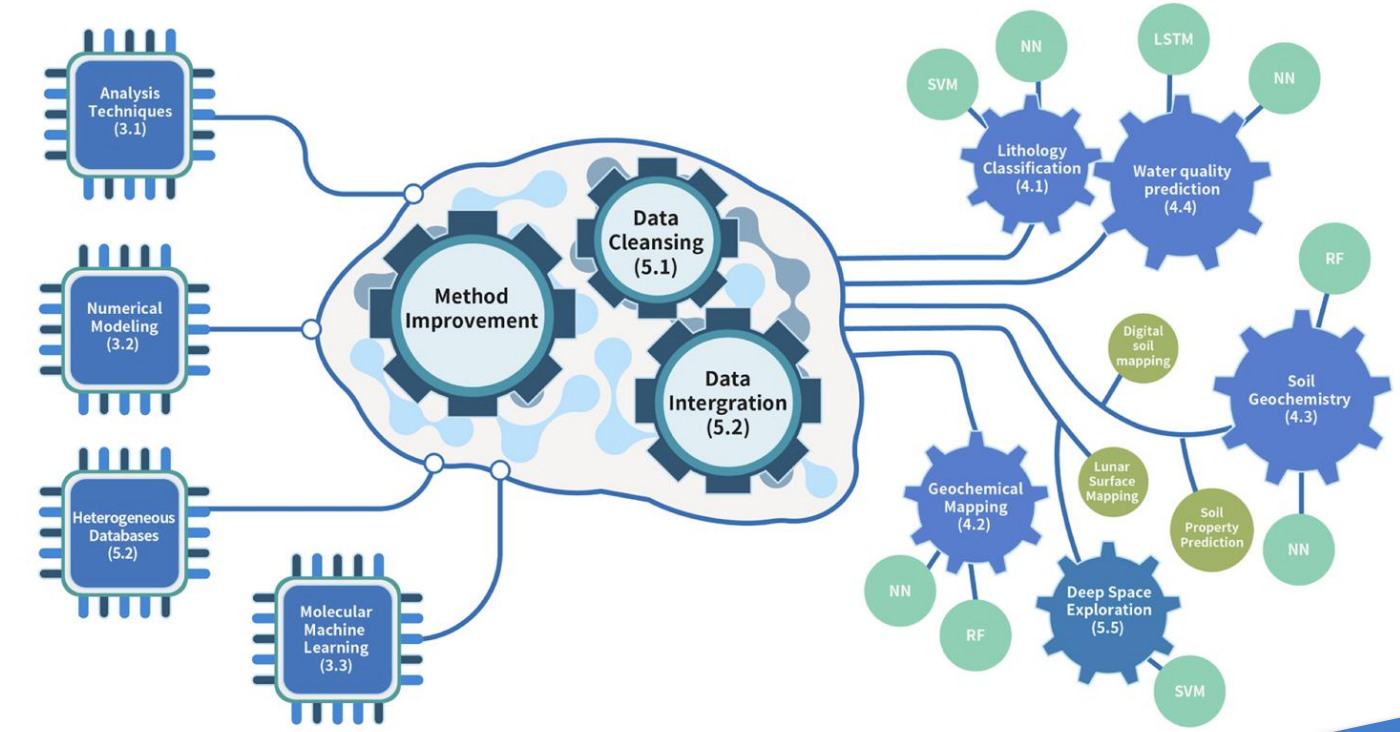
+ Add to Mendeley Share Cite

https://doi.org/10.1016/B978-0-323-99762-1.00123-6

Get rights and content

Warning – it is 39 pages long

The rise of Data-driven Geochemistry



He, Y, et al. "A review of machine learning in geochemistry and cosmochemistry: Method improvements and applications." *Applied Geochemistry* 140 (2022): 105273.

70 abstracts with ‘Machine Learning’ at Goldschmidt 2024

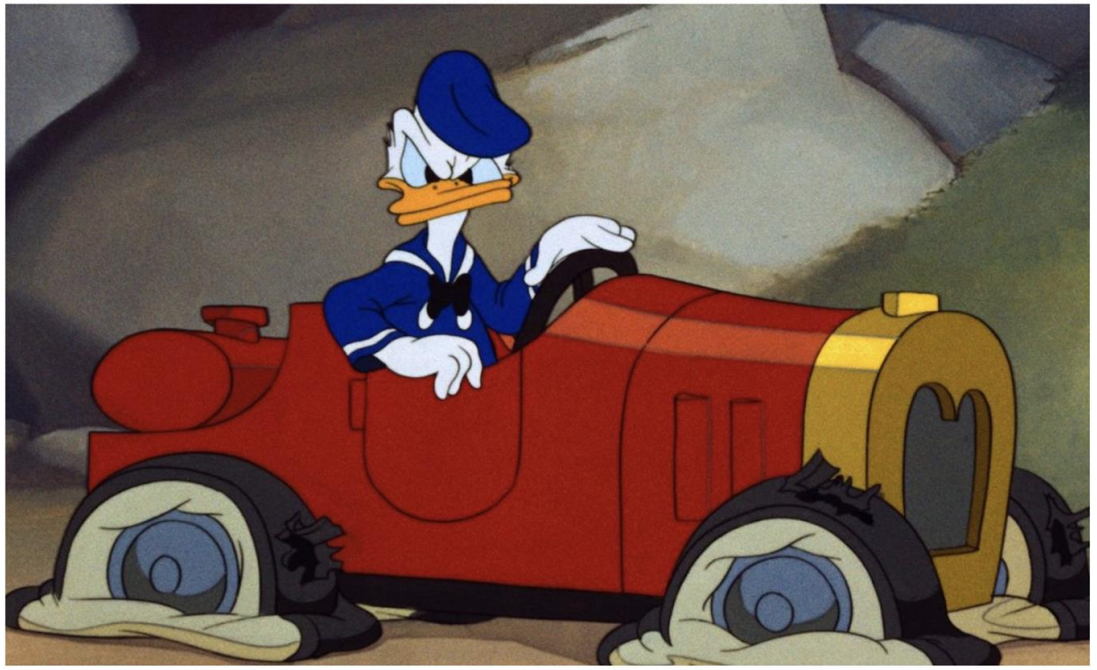
We need “Science-Ready’ Data

- Data that is OPEN
- Data that is FAIR
- Data that is scientifically **validated,** harmonized, integrated, ready to be used for statistical analysis, modelling, ML/AI

The Data Scientist with no data!

Agron Fazliu Follow

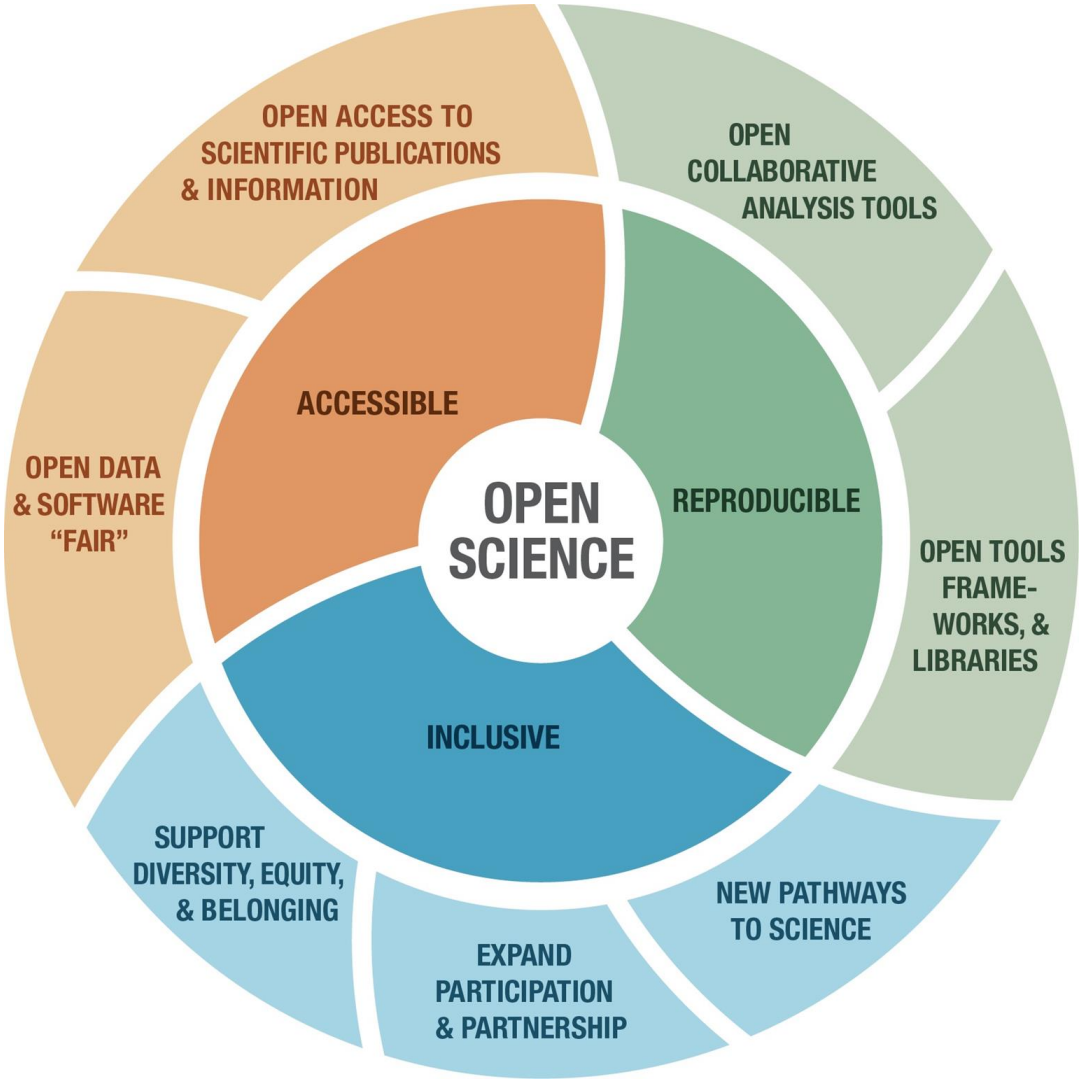
Jun 8, 2018 · 5 min read



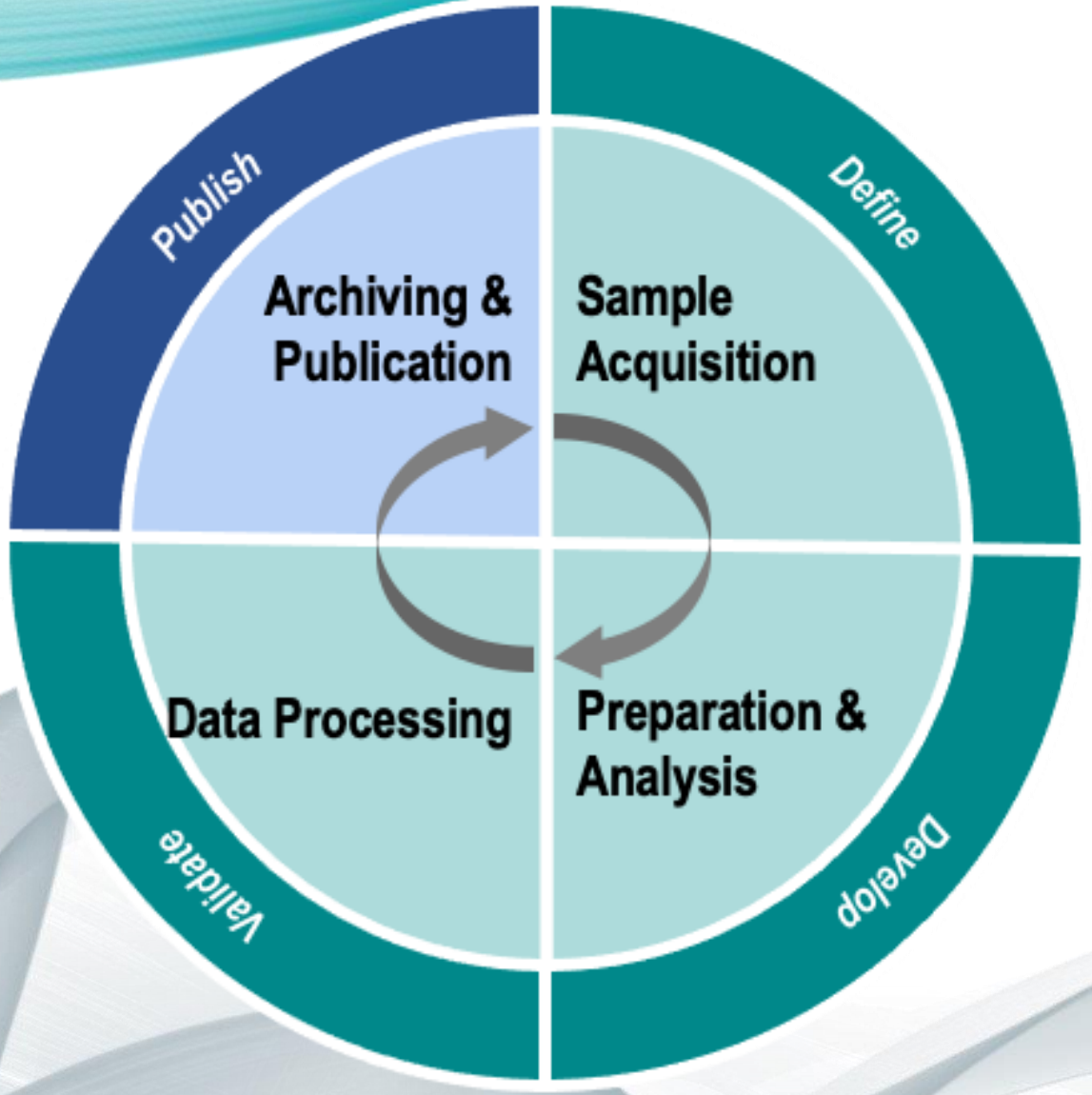
Open Data

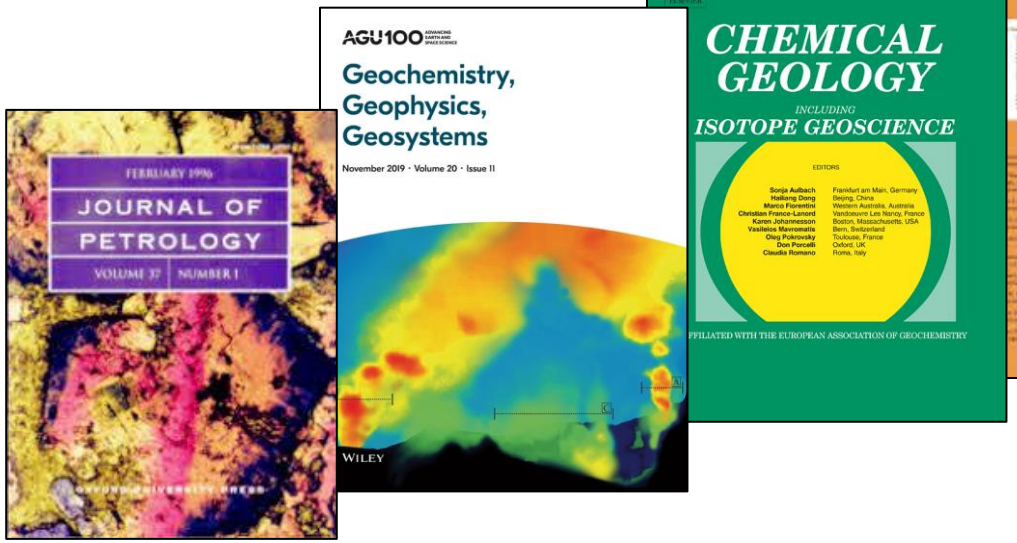
“data that can be freely used, modified, and shared by anyone for any purpose”
(Open Knowledge Foundation, 2015)

- must have an open license, or be freely available in the public domain;
- must be accessible at most at a one-time reproduction cost and should be downloadable from the internet;
- needs to be compiled in a way that makes it machine readable;
- should be shared in a nonproprietary format.



Traditional Research Life Cycle





https://d9-wret.s3.us-west-2.amazonaws.com/assets/palladium/production/s3fs-public/styles/info_block/public/thumbnails/image/Library%20GSA%20Bulletin.jpg?itok=fsSutlPQ

Klöcking, Lehnert & Wyborn (2024; Treatise on Geochemistry)

Full Research Data Life Cycle

Database dominated

Researcher dominated

How well are data being made accessible online, serviced, supported, and (re)used?

Quality attributes:

- Service accessibility and persistence
- Timeliness
- Security
- User support

What scientific questions will be addressed? How have samples and data been acquired and validated for intended use?

Quality attributes:

- Sample location attributes
- Sample preparation
- Instrument specifications
- Accuracy, Precision, Uncertainty, Validity
- Data Reduction

How well are data curated, preserved and made accessible?

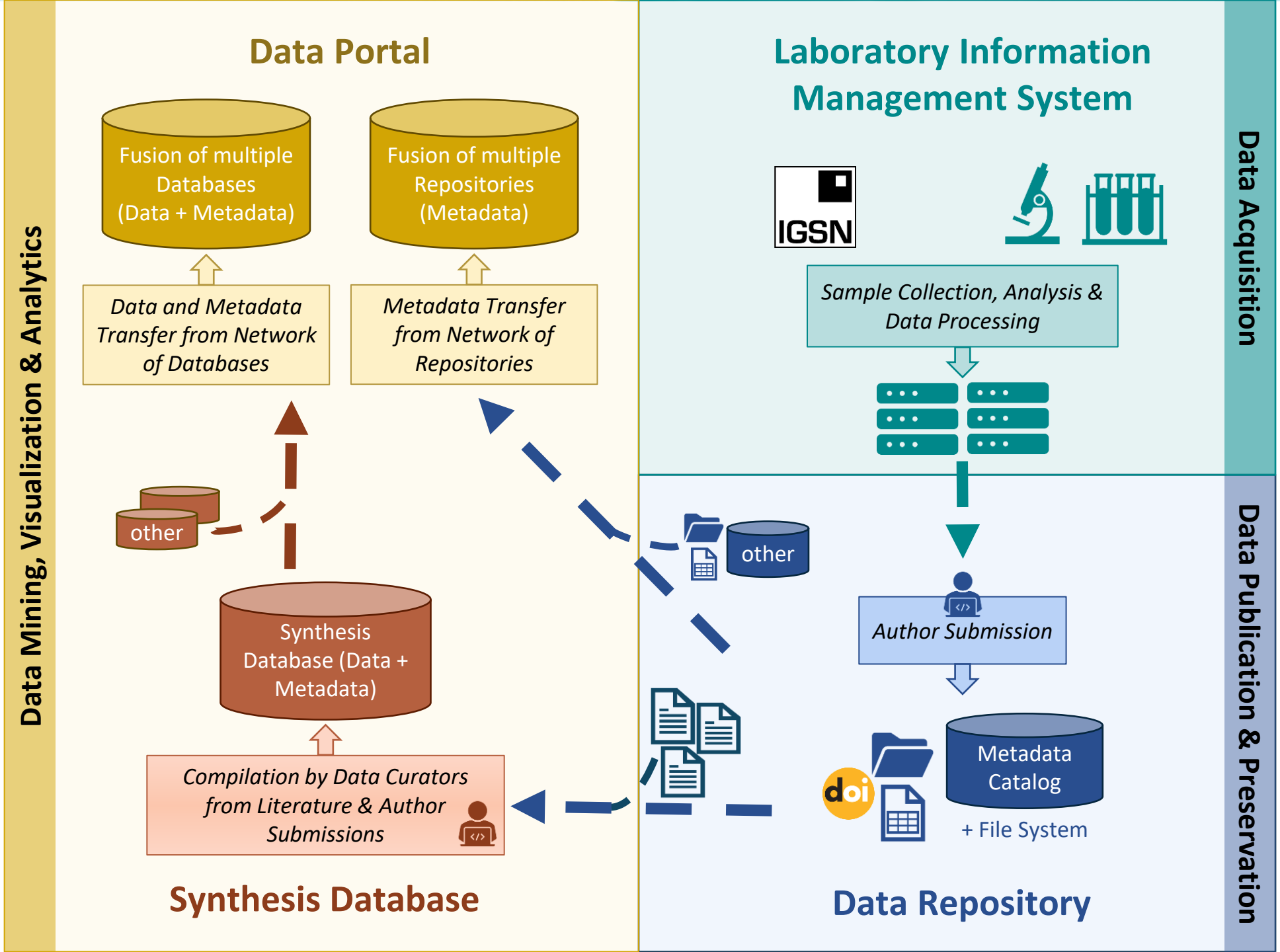
Quality attributes:

- domain-specific curation?
- Completeness of (meta)data
- ease of data accessibility
- Governance; are datasets version controlled?
- Compliance with (meta)data standards; TRUST

How was the final dataset produced, scientifically evaluated, and utilized?

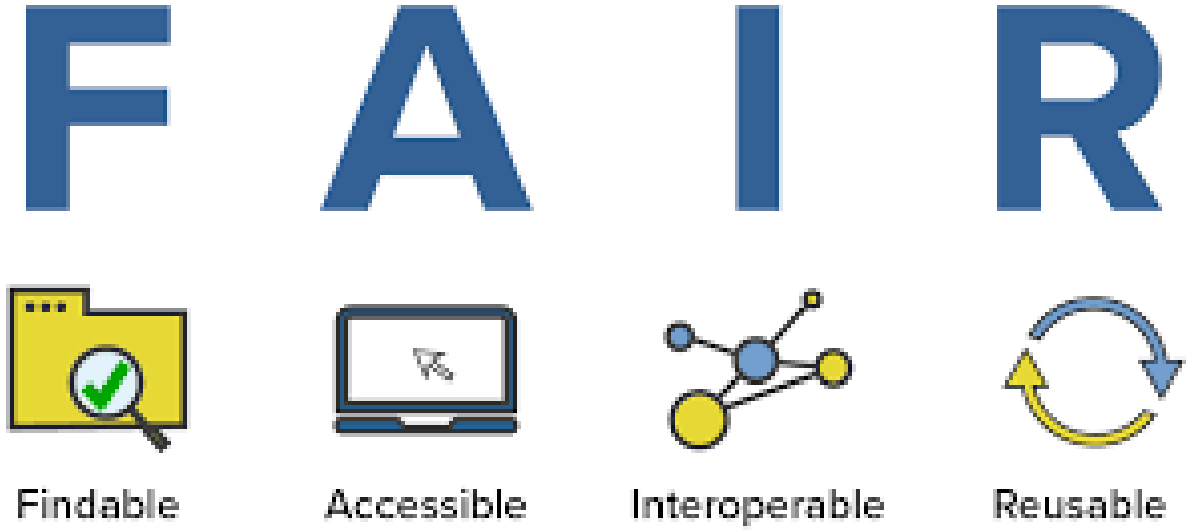
Quality attributes:

- Use of PIDs (e.g. DOI)
- Completeness of (meta)data
- Compliance with community standards
- Are (meta)data FAIR and CARE compliant?


Data Portal
Fusion of multiple Databases (Data + Metadata)
Fusion of multiple Repositories (Metadata)
Data and Metadata Transfer from Network of Databases
Metadata Transfer from Network of Repositories
other
Synthesis Database (Data + Metadata)
Compilation by Data Curators from Literature & Author Submissions
Synthesis Database
Laboratory Information Management System
IGSN
Sample Collection, Analysis & Data Processing
other
Author Submission
Metadata Catalog
+ File System
Data Repository
Data Mining, Visualization & Analytics
Data Acquisition
Data Publication & Preservation

Data need to be FAIR in order to fulfill the promises of Open Science.

Wilkinson, Mark D., et al. "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific data* 3.1 (2016).



- FAIR data are accessible and understandable to humans & machines.
- FAIR data are deposited and **curated** in certified trusted repositories to ensure reuse, persistent access & preservation.
- FAIR data are citable (registered with Persistent Identifier systems - DOI).
- FAIR data should have clear usage licenses.
- FAIR data are documented by rich metadata that support discovery and reuse.

Towards FAIR Geochemical Data

Comment | Published: 29 October 2021

Time to change the data culture in geochemistry

Katy J. Chamberlain, Kerstin A. Lehnert, Iona M. McIntosh, Dan J. Morgan & Gerhard Wörner

Nature Reviews Earth & Environment **2**, 737–739 (2021) | Cite this article

https://www.egu.eu/webinars/91/where-is-my-data-where-did-it-come-from-and-how-was-it-obtained-improving-access-to-geoanalytical-research-data/

GDB4

Where is my data, where did it come from and how was it obtained? Improving Access to Geoanalytical Research Data

Co-sponsored by AGU

Convener: Alexander Prent | Co-conveners: Marthe Klöcking[ECS], Geertje ter Maat[ECS], Lucia Profeta[ECS]

More Reading!!!

Results from a workshop held at the Goldschmidt Conference in Hawaii in 2022

ELSEVIER

Geochimica et Cosmochimica Acta

Volume 351, 15 June 2023, Pages 192-205

Community recommendations for geochemical data, services and analytical capabilities in the 21st century

https://doi.org/10.1016/j.gca.2023.04.024

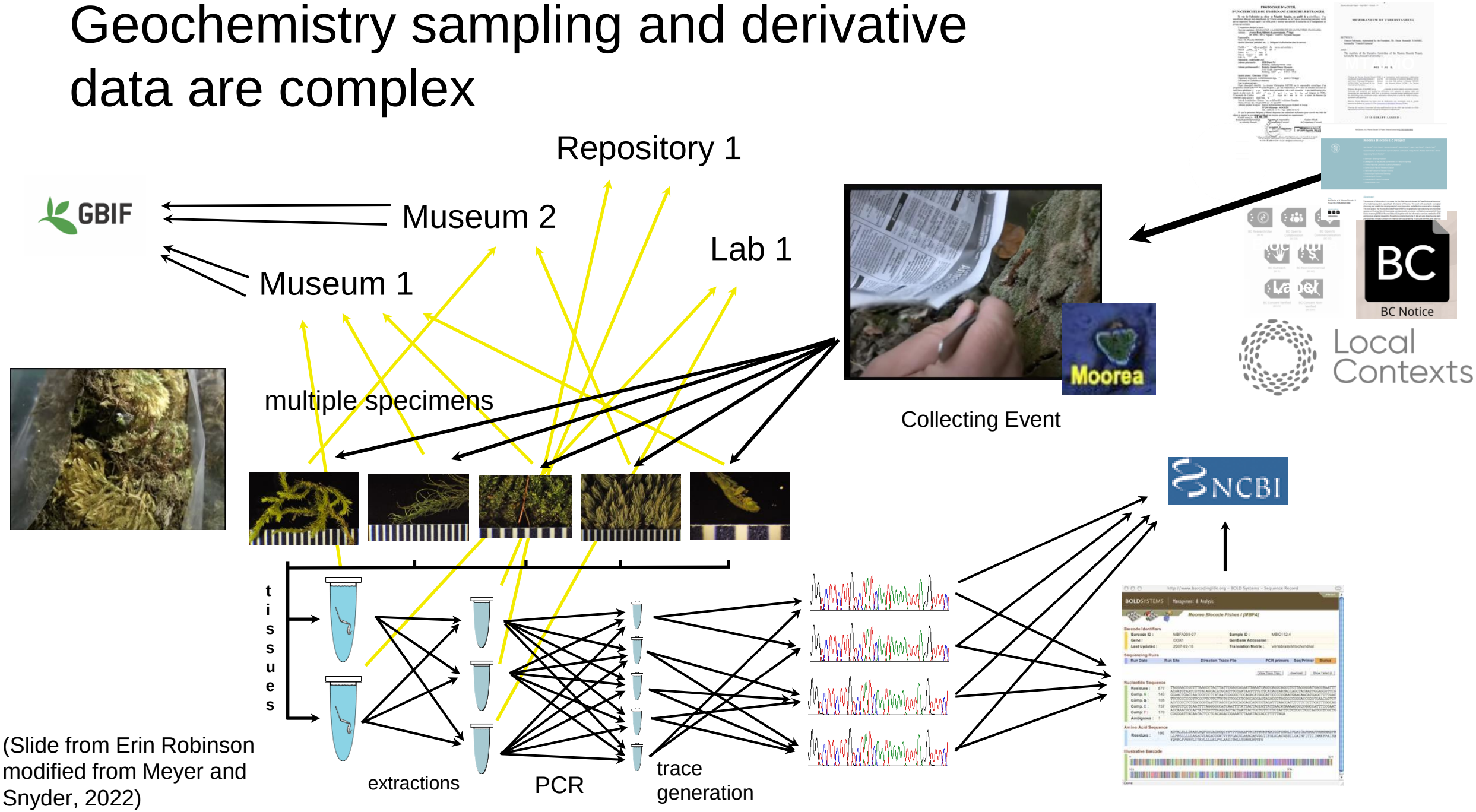
Marthe Klöcking [a], Lesley Wyborn [b], Kerstin A. Lehnert [c], Bryant Ware [d], Alexander M. Prent [e d f], Lucia Profeta [c], Fabian Kohlmann [g], Wayne Noble [g], Ian Bruno [h], Sarah Lambart [i], Halimulati Ananuer [j], Nicholas D. Barber [k l], Harry Becker [m], Maurice Brodbeck [n], Hang Deng [o], Kai Deng [p], Kirsten Elger [q], Gabriel de Souza Franco [r], Yajie Gao [b], Khalid Mohammed Ghasera [s]...Tengfei Zhou [ah]

We need CARE-compliant data

https://images.squarespace-cdn.com/content/v1/5d3799de845604000199cd24/1567592451970-2R8XFL670QNMAGDRV7ED/Be+FAIR+and+CARE.png?format=2500w

- But CARE builds on FAIR
- We are a long way from being able to do FAIR
- We are raising awareness of whether the source Samples are CARE compliant.
- If the samples are not CARE compliant then any derived analytical data is not as well.

Geochemistry sampling and derivative data are complex



(Slide from Erin Robinson modified from Meyer and Snyder, 2022)

For TRUST we need Domain Repositories

- Generic repositories can make data findable & accessible, but do not deliver the high-quality data services based on disciplinary expertise;
- Domain repositories align their services with scientific priorities through a combination of social & technical programs (e.g., community engagement & governance);
- Domain repositories develop & promote discipline-specific best practices & standards for data and software;

Need to follow the TRUST Principles:

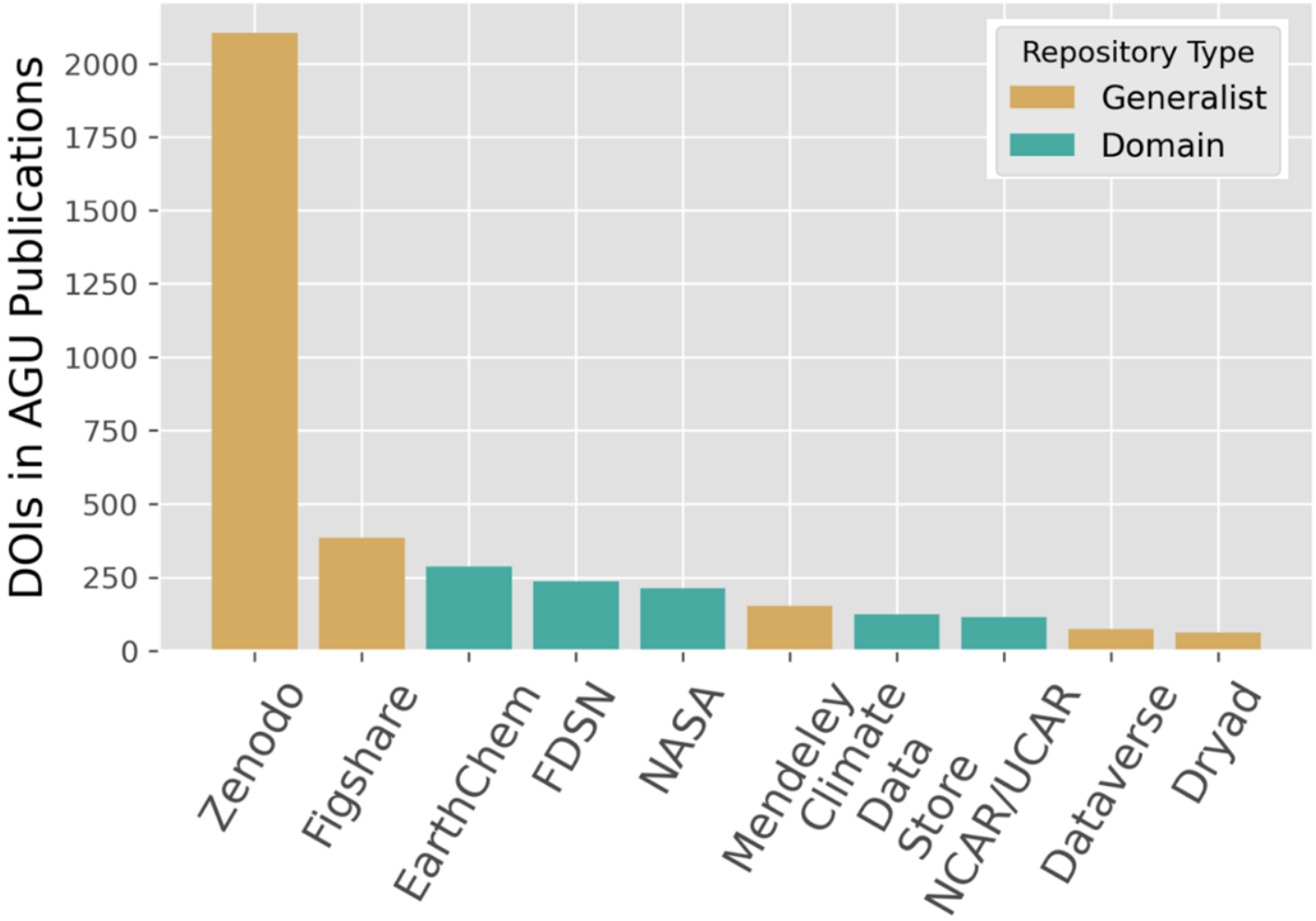
CoreTrustSeal certification does not mean there is curation

- Because of the lack of standards and the extensive fragmentation, basic domain specific curation is essential for interoperability, reuse and the aggregation of disparate datasets into synthesis databases
- CoreTrustSeal (CTS) does not mandate curation even for Domain Repositories
- There are 4 levels of curation accepted in CTS

1. Content distributed as deposited
2. Basic curation – e.g. brief checking, addition of basic metadata or documentation
3. Enhanced curation – e.g. conversion to new formats during ingest, enhancement of documentation and metadata
4. Data-level curation – as in C above, but with additional editing of deposited data

https://www.coretrustseal.org/why-certification/requirements/

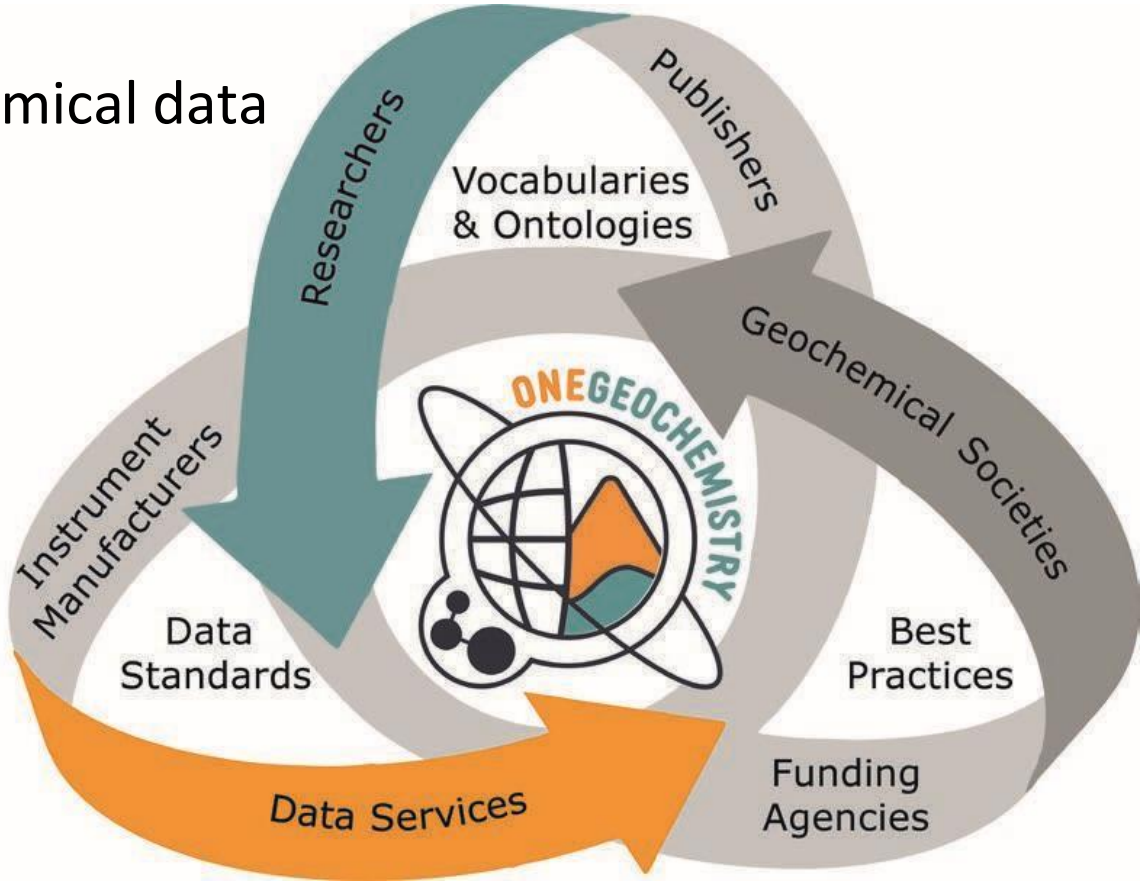
But the trend is towards researchers using Generic Repositories



Top 10 repositories used in primary research papers published in AGU journals in 2022. Modified from Vrouwenvelder & Stall (2024).

How do we advance standards for FAIR, CARE, TRUST and Open?

- We need to build community around geochemical data
  - Create awareness of what is needed
  - Broaden engagement from local
- Constitute a framework for standards
  - Establish governance
  - Create communication paths
  - Inform policy
- Advance technical solutions
  - Facilitate developments
  - Promote implementation



We have formed OneGeochemistry: An international collaboration:

https://onegeochemistry.github.io/

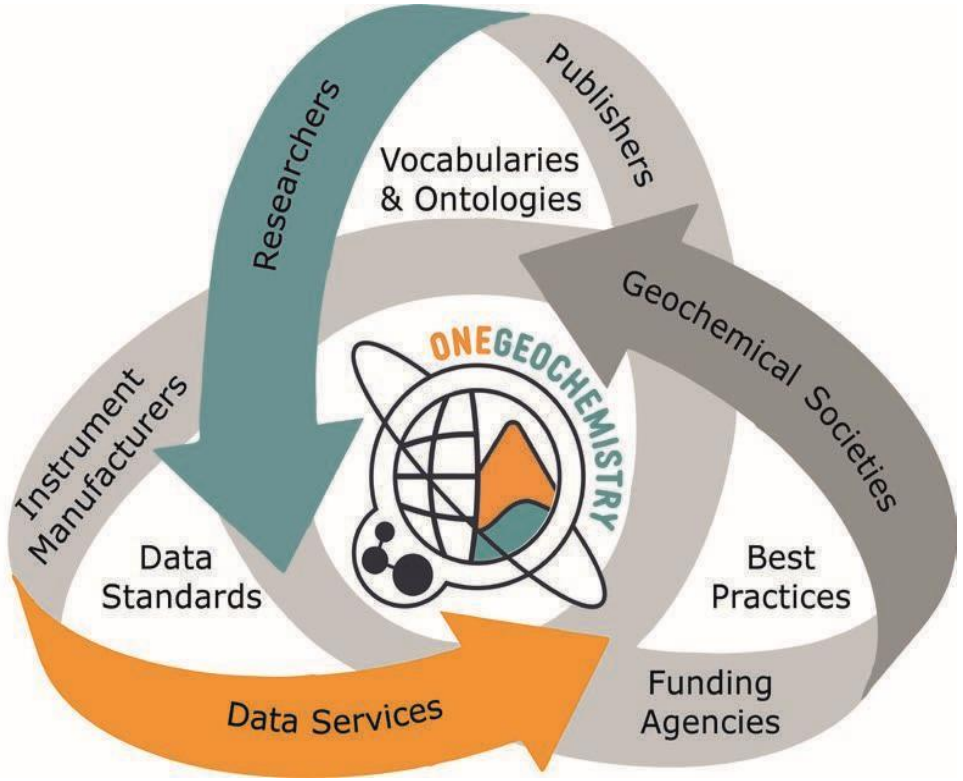
- Supported by

- Endorsed by:

Future Directions and Challenges

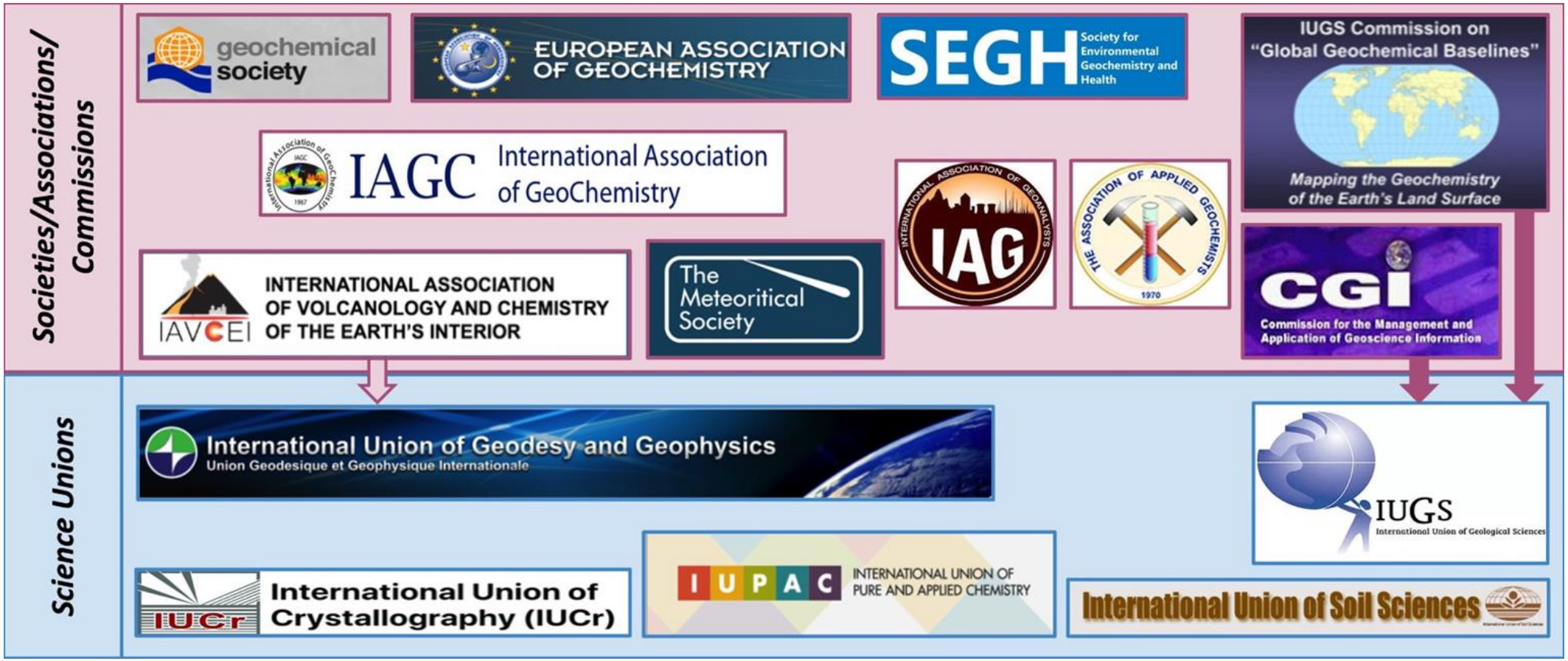
➔ Community action required

- Technical developments often disconnected from geochemistry community
- Data infrastructures chronically underfunded & underappreciated/-valued



- Culture change starting but sloooow!
  - Holding on to tradition
  - Fear of being scooped
  - Takes time & effort
  - Lack of credit & recognition in academic track record
- Question of authority
  - >10 different societies and unions
  - Increasing requirements from funders/publishers; but conflicting advice, no clear guidelines

Question of Authority



OneGeochemistry

OneGeochemistry is an international collaboration of organisations that support geochemistry capability and data production. Our focus is to better **coordinate global efforts in geochemical data standardisation**, facilitate communication between groups and lessen duplication of efforts.

It is our mission to advance geoscientific knowledge and discoveries by building and maintaining consensus-driven standards that make geochemistry research data globally findable and accessible, and truly interoperable and reusable to both humans and machines. OneGeochemistry seeks to create a global geochemical data network that facilitates and promotes data discovery and access through coordination and collaboration among international geochemical data providers.

Since December 2022, OneGeochemistry is acting as a CODATA Working Group under the International Science Council to bring together the disparate geochemistry initiatives across Scientific Unions, Associations, Societies and Commissions.

"We must, indeed, all hang together or, most assuredly, we shall all hang separately".
Benjamin Franklin

Find out More

RESOURCES

OneGeochemistry
http://onegeochemistry.org

Sample Registration for IGSN

SESAR
www.geosamples.org

Data Repositories

DIGIS Geochemical Data Repository
https://digis-repo.georoc.eu/

Astromaterials Data Archive
https://repo.astromat.org/

EarthChem Library
https://earthchem.org/ecl/

AusGeochem
https://ausgeochem.auscope.org.au/#/

Synthesis Databases

GEOROC
https://georoc.eu/

Astromaterials Data Synthesis
https://astromat.org/

PetDB (EarthChem Synthesis)
https://search.earthchem.org/

EarthChem Portal
http://portal.earthchem.org/

AusGeochem
https://ausgeochem.auscope.org.au/#/