NCI AUSTRALIA

One Geoscience: Providing FAIR global Access to all Geoscience Data - are we there yet?

Lesley Wyborn

GeoBerlin 2023, 7 September 2023

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

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

Acknowledgements

NCI/AuScope Geophysics and Data Team: Nigel Rees, Rui Yang, Hannes Hollmann, Jo Croucher, Rebecca Farrington, Yue Sun, Yiling Liu, Andrew Robinson, Ben Evans

ARDC National Information Infrastructure Team: Natasha SImons, Julia Martin, Melanie Barlow, Rowan Brownlee, Mingfang Wu, Adrian Burton

One Geochemistry Team: Kerstin Lehnert, Marthe Klöcking, Alexander Prent, Kirsten Elger, Dominik Hezel, Lucia Profeta, Rebecca Farrington

The 2006-2012 (?) One Geology/GeoSciML Teams: Ian Jackson, Francois Robida, Kristine Asch, Simon Cox, Tim Duffy, Boyan Broderic, Oliver Raymond

CC BY

Abstract

Geoscience deals with all fields of natural science related to understanding past, current and future states of the Earth and the terrestrial bodies. Geoscience has many sub-disciplines that have strong roots in other sciences such as chemistry, physics, geography, biology and mathematics. Each sub-discipline generally works in isolation and is governed by different science unions, societies and associations. Datasets generated by each are diverse, complex and heterogeneous: few fully comply with the FAIR principles. Hence it is hard to integrate datasets both within and across each sub-discipline. To enable efficient integration of geoscience data to effectively contribute to societal grand challenges will require conformance to agreed international standards and compliance with FAIR.

The first attempt at international digital integration of data within one of these sub-disciplines was OneGeology, which in 2008 harmonised geological map data globally. More recently, OneGeochemistry is emerging as a fledgling effort to unify geochemical data across multiple sample types and analytical techniques in the geochemical domain. Geophysicists have not yet taken up the call for OneGeophysics but the potential is there.

Rather than developing standards completely within each sub-discipline, a more holistic approach is to leverage the Observation, Measurement and Samples (OMS) Standard (IS0 19156: 2023) and break down the standards and vocabularies required into modules based around the feature of interest, instrument, procedures, event, place, properties value, result, etc.

This presentation will highlight potential ways to use OMS to accelerate development and convergence of standards and vocabularies required to enable OneGeoscience to become a reality.

cc BY

Welcome | Themes & Oral/Poster | Programme | Plenary | Field Trips Exkursionen | Registration | SOC | Venue & Hotel | Sponsoring | Join DGGV | Awardees | Contact | Press | Schulen & Lehrende

GeoBerlin 2023 –

Geosciences Beyond Boundaries – Research, Society, Future

150th PGLA (BGR) Anniversary and 175th DGGV Anniversary

Berlin | 3 – 7 September 2023

CC BY

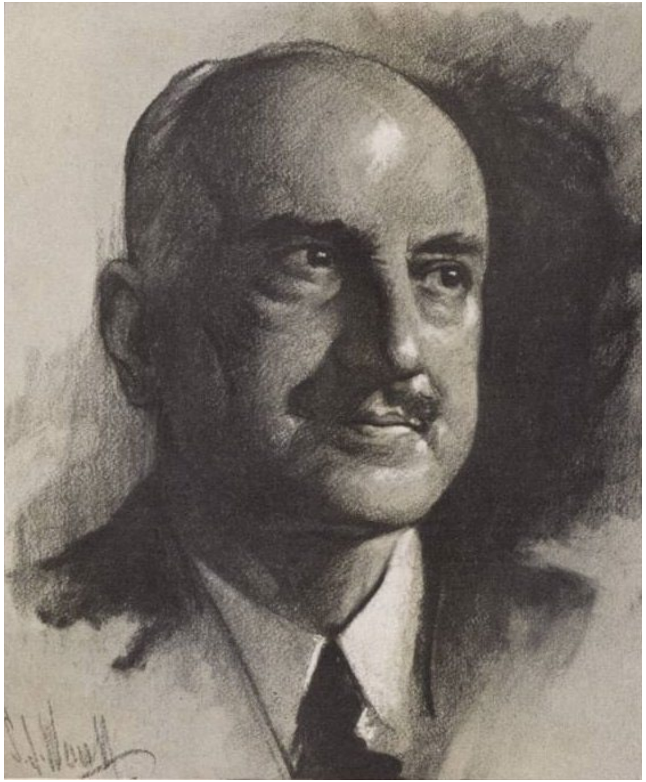
OneGeoscience: Providing FAIR global access to all Geoscience data - are we there yet?

This talk is about Science and in increasing need to share data to enable new and better science

1) Set the scene: what is science?
2) Then and now - what is driving change?
3) OneGeology
4) OneGeophysics
5) OneGeochemistry
6) Geosciences beyond the boundaries
7) What are our priorities?

CC BY

Setting the Scene: We are in a great time of change: What can we learn from history?



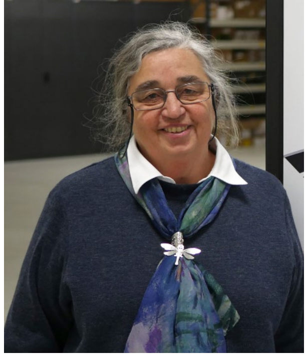
Those who cannot remember the past are condemned to repeat it.

George Santayana

Source: https://upload.wikimedia.org/wikipedia/commons/2/2a/George_Santayana.jpg

cc BY

Who am I?

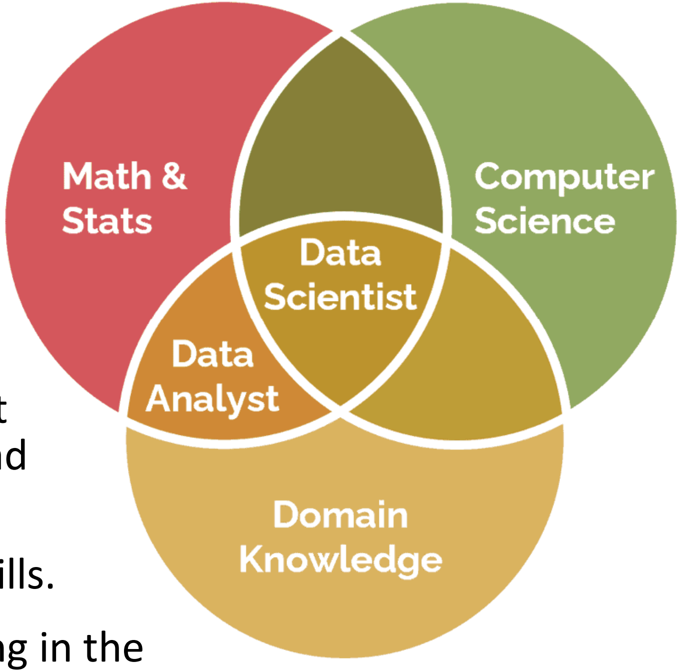


- I was born in the first half of the last century in the reign of King George VI.
- As an undergraduate in the late 1960's, computers were something in the postdoc/advanced research sections of the Physics department.
- The internet did not exist (neither did EXCEL): it was pen, paper and typewriter.
- There were no computer science departments: there was something called "Applied Mathematics".
- University science courses taught science and nothing else.
- My formal training is in research science (geochemistry and mineral systems)
- Today, I would be called a "Data Scientist"

CC BY

So what is a Data Scientist?

- In my beginnings there were no data scientists - there were nerdy scientists who cared about data, but there were no cheap tools to manage or store it.
- As the data deluge increased, computer experts were brought in to help researchers manage and process their data: they had no discipline knowledge.
- The majority of researchers had no data/computer science skills.
- Chaos ensued until it was recognised that there was something in the middle: researchers who understood data, and computer scientists with some discipline expertise - the data scientist was born.
- To me a data scientist is, and always sits, at the crossroads of science and computer science



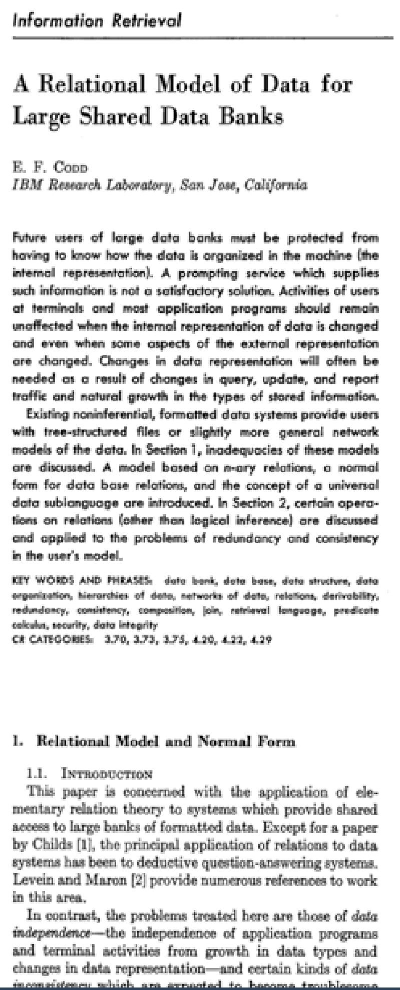
https://www.datascience-pm.com/data-analyst-vs-data-scientist/

CC BY

Codd's 1970 Seminal Paper on Relational Database Management Systems

1. Future users of large data banks need to be protected from having to know how the data is organised in the machine (the internal representation).
2. Activities of users at terminals … should remain unaffected when the internal representation of data is changed.
3. Changes in data representation will often be needed as a result of changes in query, update….and natural growth in the types of stored information.

Codd, E. F, 1970. A Relational Model of Data for Large Shared Data Banks. Communications of the ACM Volume 13 Issue 6, June 1970 pp 377–387 https://doi.org/10.1145/362384.362685


Information Retrieval

A Relational Model of Data for Large Shared Data Banks

E. F. Codd
IBM Research Laboratory, San Jose, California

Future users of large data banks must be protected from having to know how the data is organized in the machine (the internal representation). A prompting service which supplies such information is not a satisfactory solution. Activities of users at terminals and most application programs should remain unaffected when the internal representation of data is changed and even when some aspects of the external representation are changed. Changes in data representation will often be needed as a result of changes in query, update, and report traffic and natural growth in the types of stored information.

Existing noninferential, formatted data systems provide users with tree-structured files or slightly more general network models of the data. In Section 1, inadequacies of these models are discussed. A model based on n-ary relations, a normal form for data base relations, and the concept of a universal data sublanguage are introduced. In Section 2, certain operations on relations (other than logical inference) are discussed and applied to the problems of redundancy and consistency in the user's model.

KEY WORDS AND PHRASES: data bank, data base, data structure, data organization, hierarchies of data, networks of data, relations, derivability, redundancy, consistency, composition, join, retrieval language, predicate calculus, security, data integrity

CR CATEGORIES: 3.70, 3.73, 3.75, 4.20, 4.22, 4.29

1. Relational Model and Normal Form

1.1. Introduction

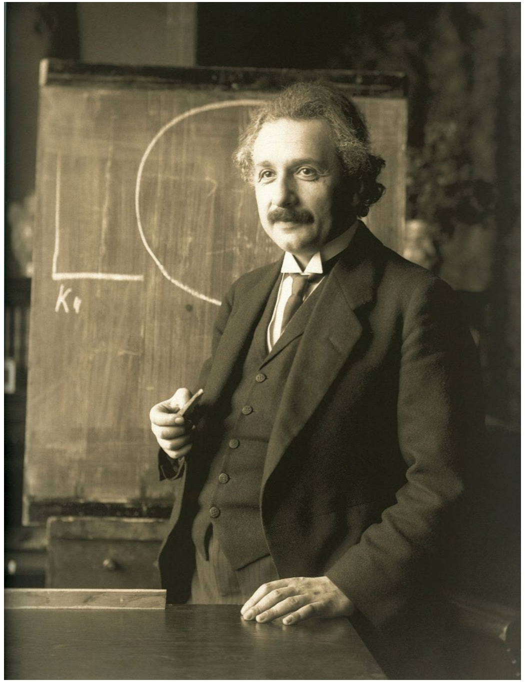
This paper is concerned with the application of elementary relation theory to systems which provide shared access to large banks of formatted data. Except for a paper by Childs [1], the principal application of relations to data systems has been to deductive question-answering systems. Levein and Maron [2] provide numerous references to work in this area.

In contrast, the problems treated here are those of *data independence*—the independence of application programs and terminal activities from growth in data types and changes in data representation—and certain kinds of data

CC BY

What is a scientist?

- A true scientist is one whose data and conclusions can be independently verified
- Therefore, input artefacts (samples, software, data, algorithms, etc) need to be transparent, open and accessible so we can validate the both the data and then assess the conclusions



https://upload.wikimedia.org/wikipedia/commons/thumb/3/3e/Einstein_1921_by_F_Schmutzer_-_restoration.jpg/1280px-Einstein_1921_by_F_Schmutzer_-_restoration.jpg

CC BY

Then and now:
What are the greatest changes I have seen?

CC BY


Welcome
Themes & Oral/Poster
Programme
Plenary
Field Trips Exkursionen
Registration
SOC
Venue & Hotel
Sponsoring
Join DGGV
Awardees
Contact
Press
Schulen & Lehrende
GeoBerlin 2023 –
Geosciences Beyond Boundaries – Research, Society, Future
150th PGLA (BGR) Anniversary and 175th DGGV Anniversary
Berlin | 3 – 7 September 2023

CC BY

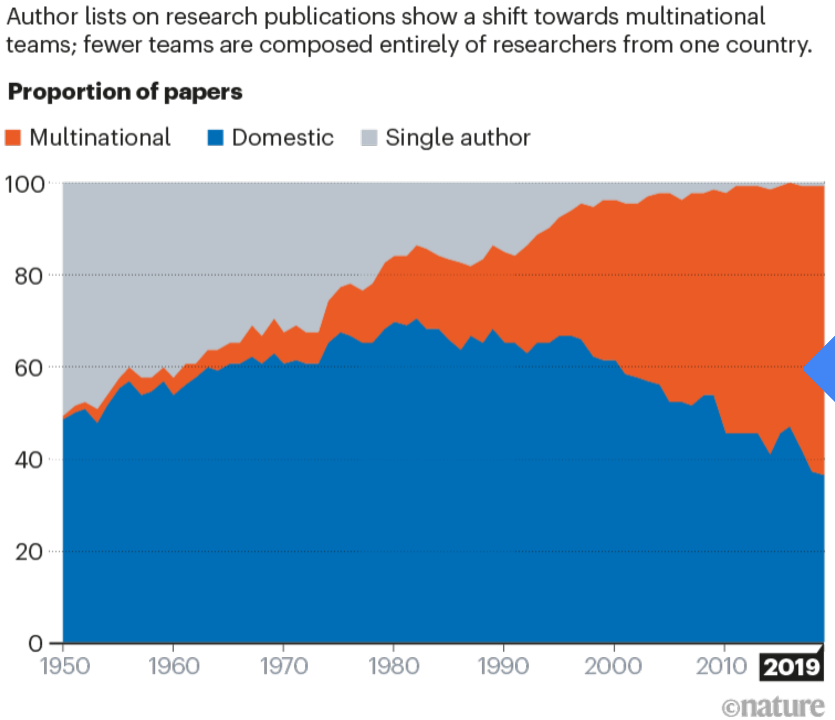
Our science today needs to address societal needs and contribute to grand challenges for sustaining our planet

https://sdgs.un.org/goals

CC BY

Our Research is increasingly reliant on International collaborations

To solve Global Grand Challenges, we need to collaborate internationally on data



Modern research papers increasingly result from International collaborations

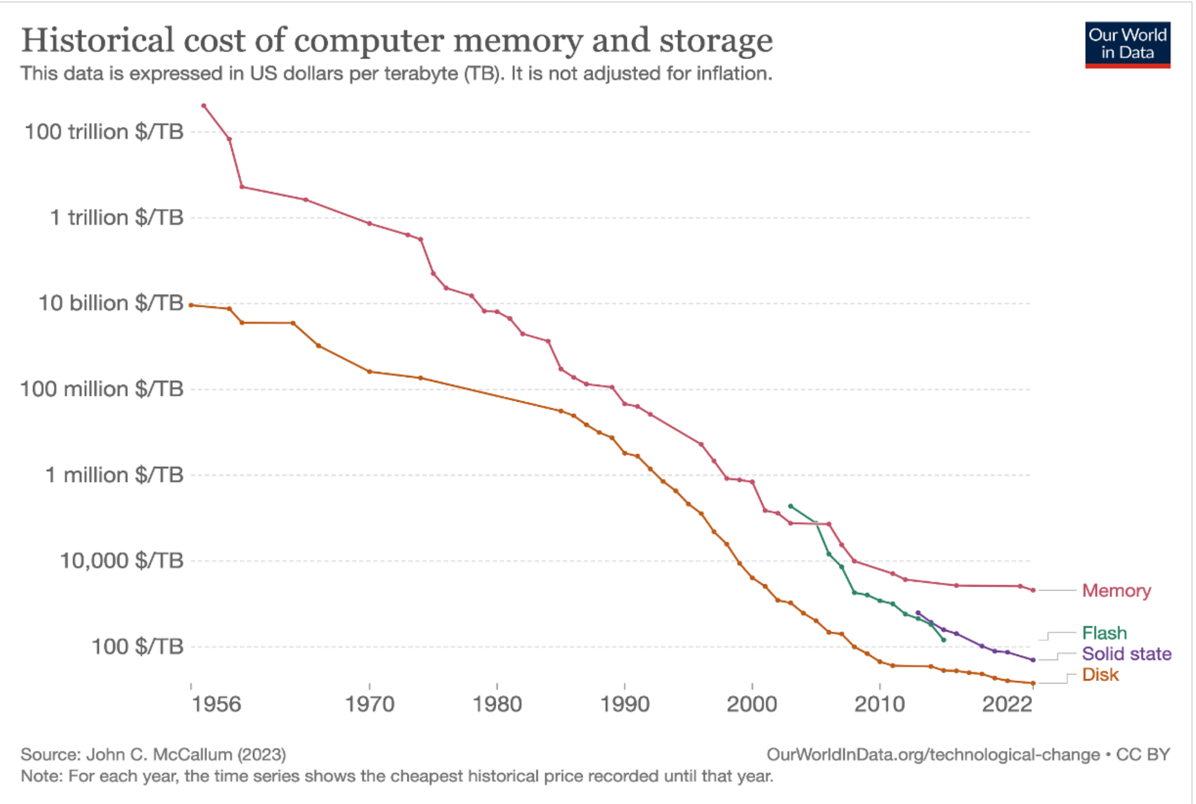
Monastersky, R., and Van Noorden, R., 2019. 150 years of *Nature*: a data graphic charts our evolution. Nature, 575(7781):22-23. https://doi.org/10.1038/d41586-019-03305

CC BY

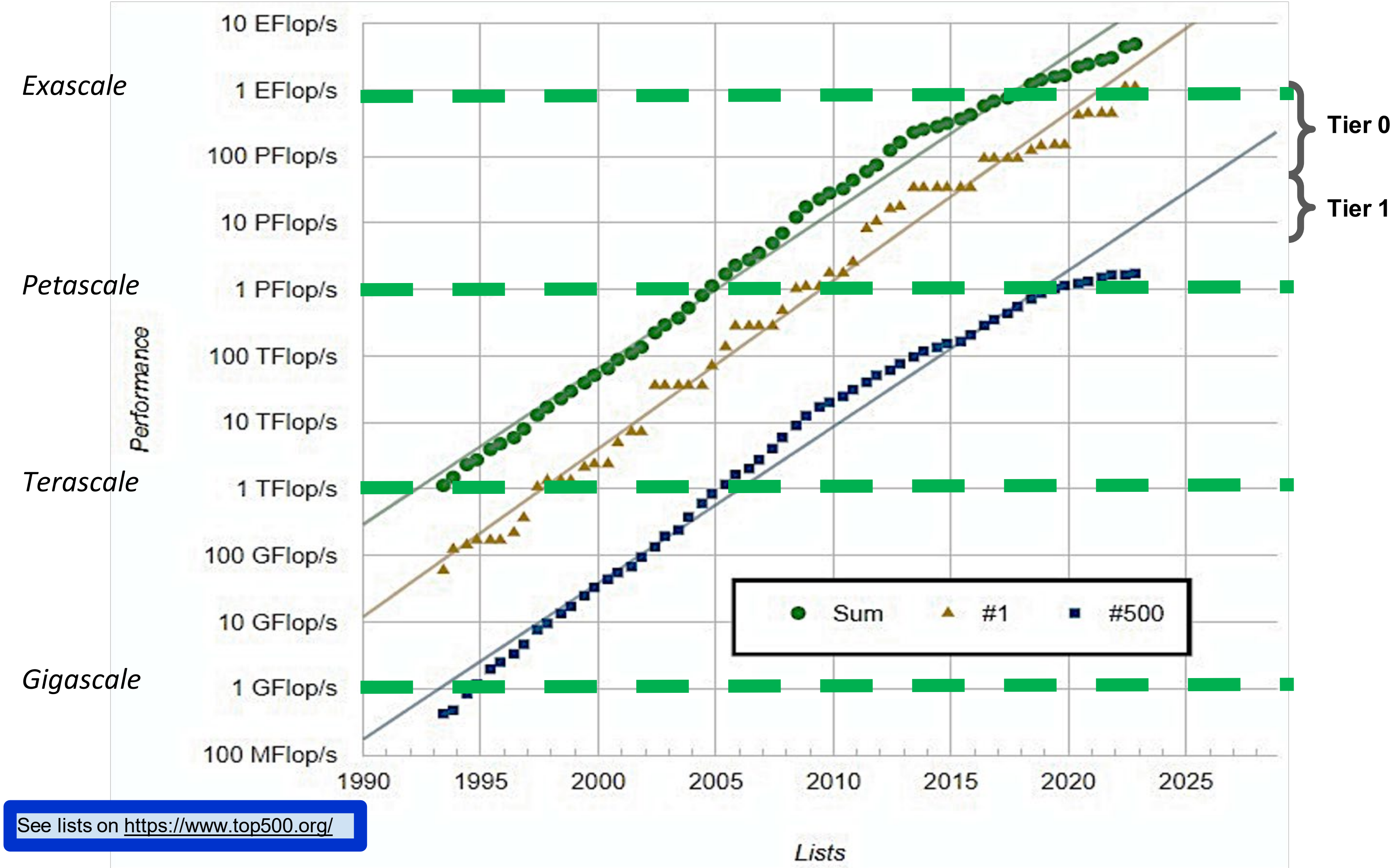
In the 1980's & 1990's storage and memory were expensive

- In 1980 storage was ~$1M per gigabyte: efficiency was key!
- Controlled vocabularies:
  - Were critical to efficient storage;
  - Kept the data clean: helped those geologists who could not spell (i.e., 90%!!!)
- Around the turn of the century, I joined the international GeoSciML project: worked on key vocabularies, including rock types.

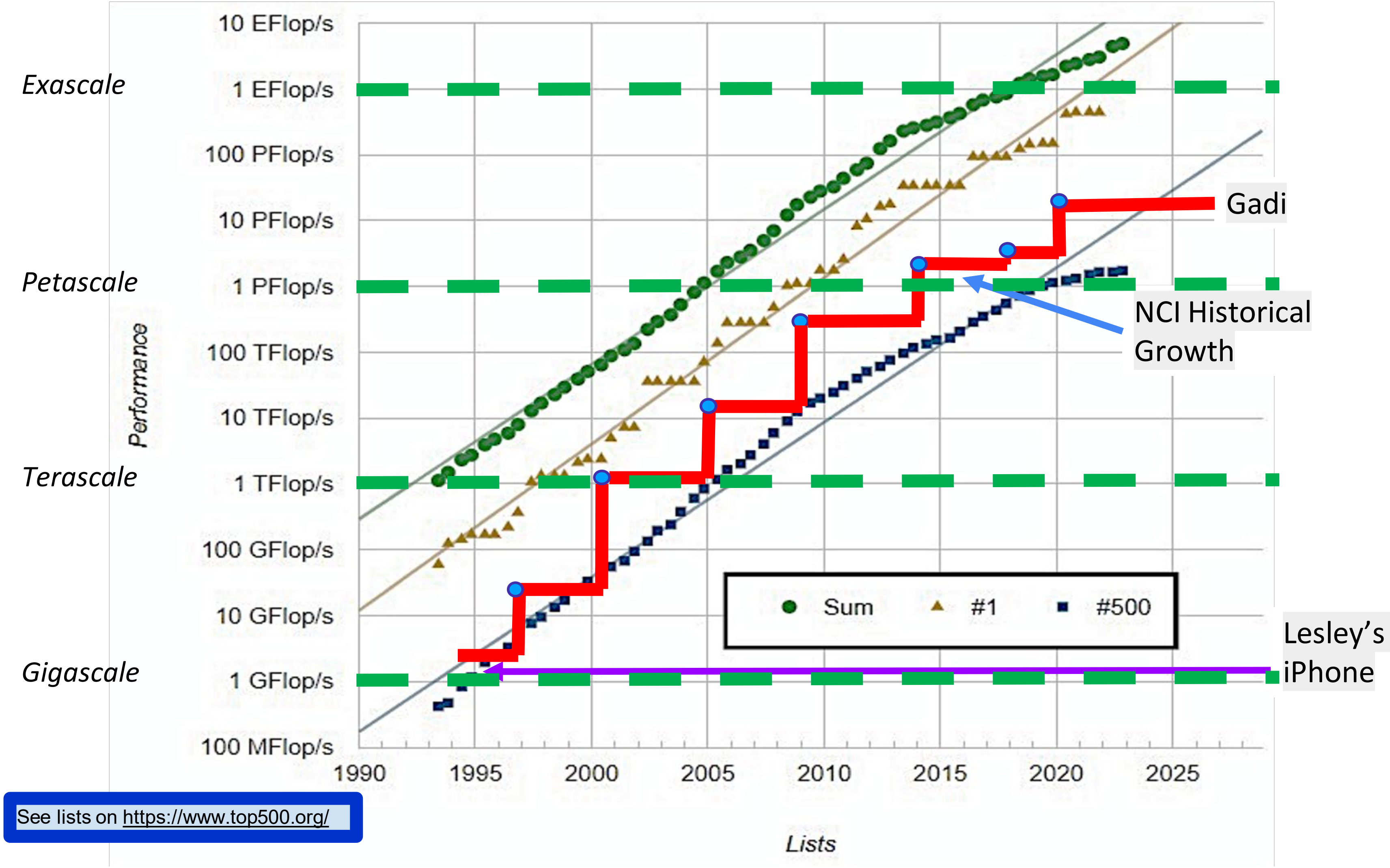
https://ourworldindata.org/technological-change



CC BY



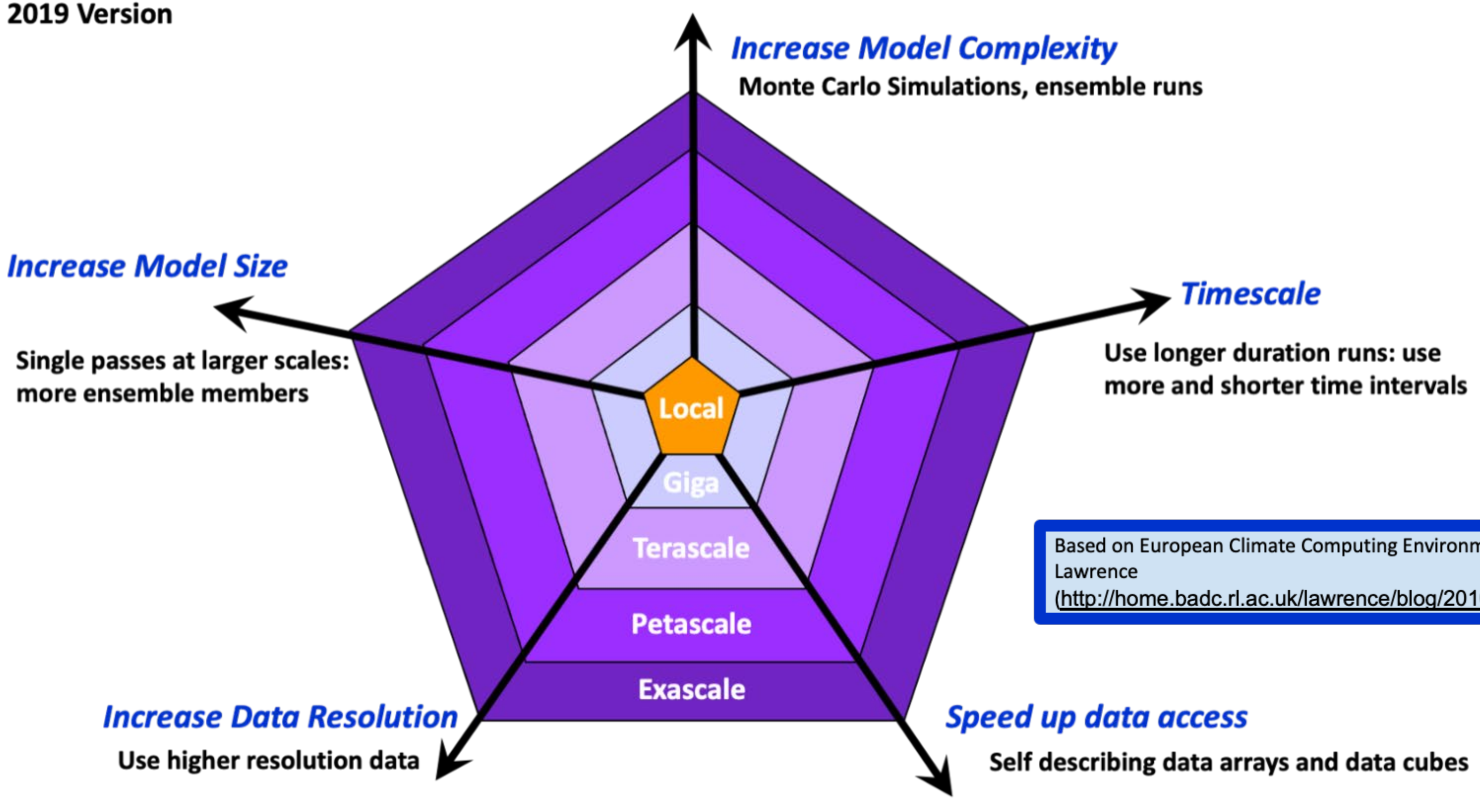
See lists on https://www.top500.org/



See lists on https://www.top500.org/

What Does More Computational Power Mean to our Science?

2019 Version

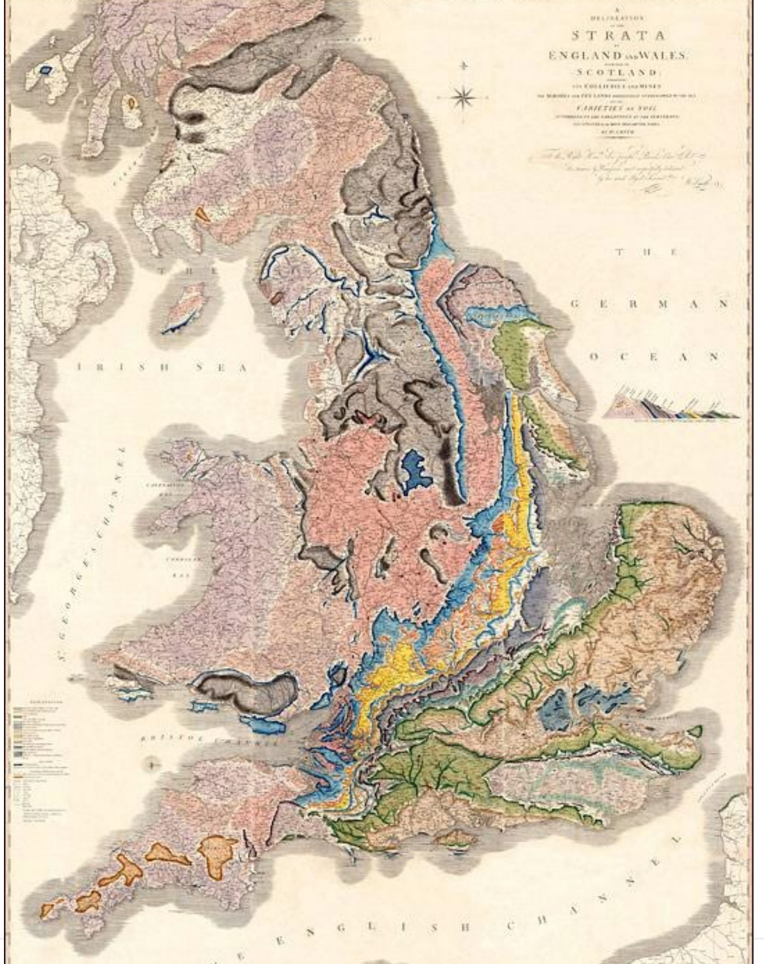


Based on European Climate Computing Environments, Bryan Lawrence (http://home.badc.rl.ac.uk/lawrence/blog/2010/08/02)

CC BY

The “Power of the Picture” has dominated since 1815

- The view/image/picture of data was key and for a long time has been the main means of communication of scientific data.
- This has been dominant in geoscience since William Smith’s first geological map in 1815.
- It is all about ‘Human readable” and has been since 1815.

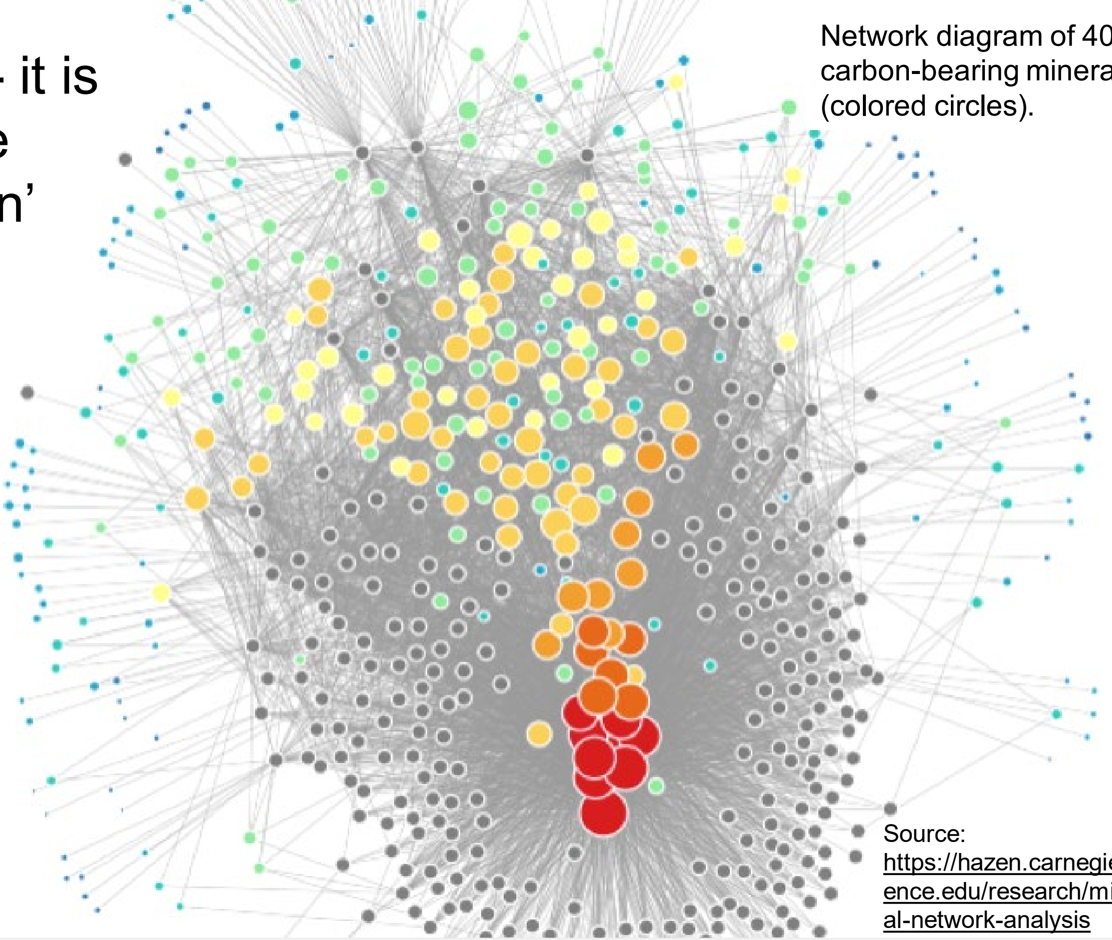


https://earthobservatory.n
asa.gov/images/8733/willi

CC BY

Science is changing - it is now the ‘Power of the Individual Observation’

- Through the semantic web we can now visualise millions of individual words (data points), not create visualisations of interpretations of millions of words (data points).
- It is now the ‘Power of the Each Observation’.
- New science is being enabled because we can look at data at full resolution



Network diagram of 400 carbon-bearing minerals (colored circles).

Source: https://hazen.carnegiescience.edu/research/mineral-network-analysis

CC BY

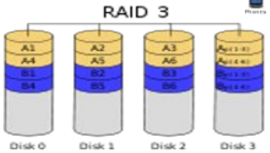
With each generation, we adapt new technologies to more data

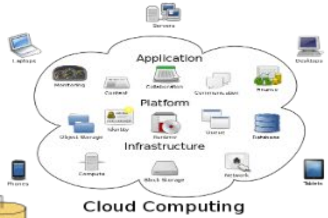
6th assessment 2020

5th assessment 2013

4th assessment 2007

3rd assessment 2001





50 gigabytes

35 terabytes

30 petabytes

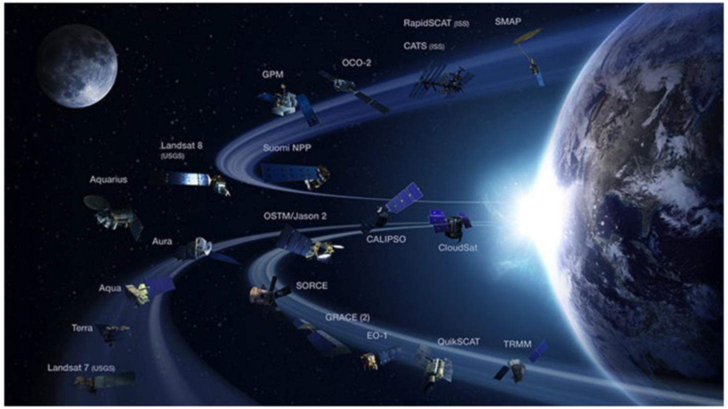
?? exabytes

ARC CENTRE OF EXCELLENCE FOR CLIMATE SYSTEM SCIENCE

CC BY

NCI AUSTRALIA

The NASA Processing Levels L0-L4: progressing from raw instrument data to multiple derivative products



Source of Graphic: https://earthobservatory.nasa.gov/blogs/earthmatters/2015/04/29/elusive-earthquake-imagery/

- Raw instrument data.
- L0 = Reconstructed, unprocessed instrument data at full resolution.
- L1 = L0 data time-referenced, annotated & processed to sensor units.
- L2 = Derived geophysical variables at the same resolution.
- L3 = Variables mapped onto uniform space-time grid scales.
- L4 = Model outputs or results from analyses of lower-level data.

Source: https://earthdata.nasa.gov/collaborate/open-data-services-and-software/data-information-policy/data-levels

CC BY

Scaling up MT time series processing

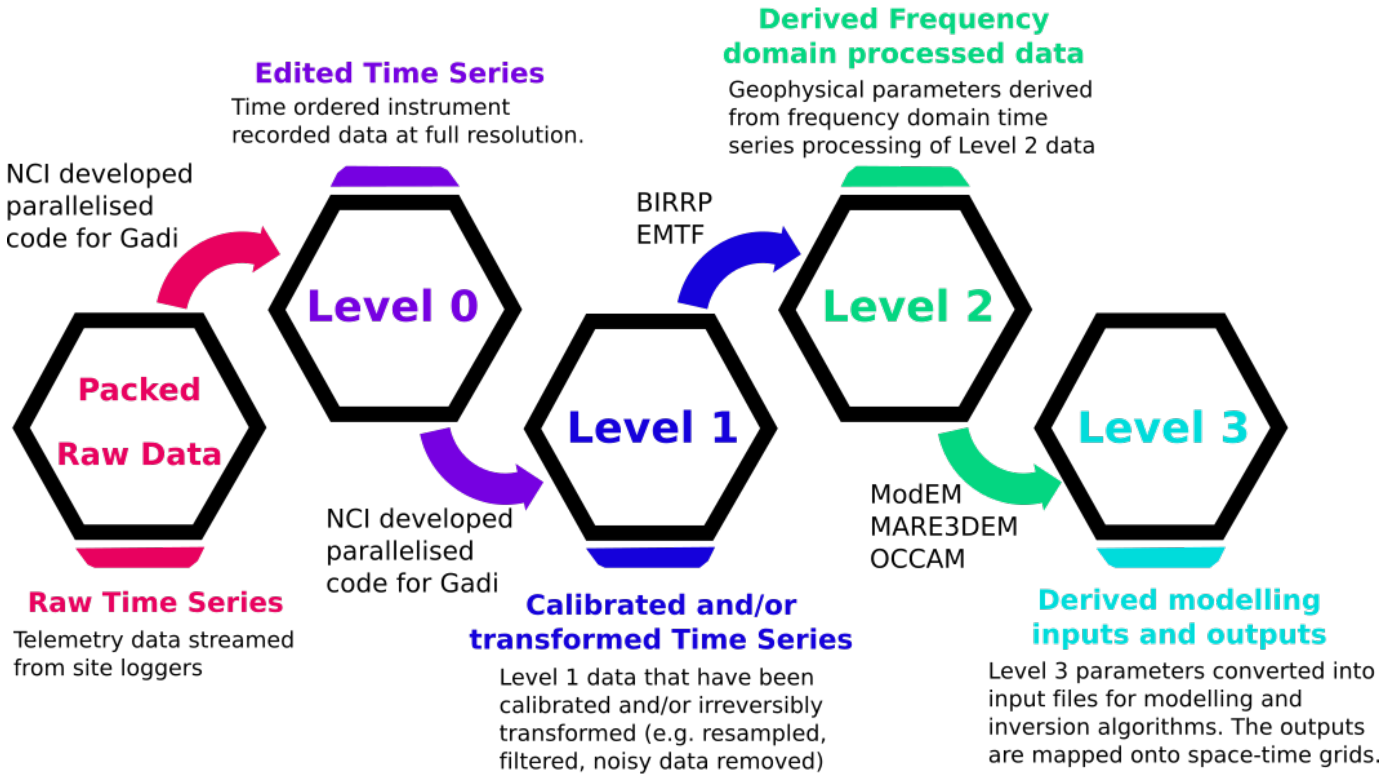
The **M**agneto**T**ellurics **t**ime **s**eries **d**ata **p**ublication (**MTtsdp**) codes: https://github.com/nci/MTtsdp

Processing Levels	Name	Description
Packed Raw Data	Raw Time Series	Telemetry data streamed from site loggers
Level 0	Edited Time Series	Time ordered instrument recorded data (e.g., raw voltages, counts) at full resolution
Level 1	Transformed Time Series	Level 0 data that have been transformed (e.g., calibrated, resampled, rotated, had noisy data removed, filters applied).
Level 2	Derived frequency domain processed data	Geophysical parameters (e.g., impedance tensors) derived from frequency domain time series processing of Level 1 data
Level 3	Derived modelling inputs and outputs	Level 2 parameters converted into input files for modelling and inversion algorithms with outputs mapped onto space-time grids.

Rees, N., Evans, B., Heinson, G., Conway, D., Yang, R., Thiel, S., Robertson, K., Druken, K., Goleby, B., Wang, J., Wyborn, L. & Seillé, H., 2019. The Geosciences DeVL Experiment: new information generated from old magnetotelluric data of The University of Adelaide on the NCI High Performance Computing Platform, ASEG Extended Abstracts, 2019:1, 1-6, DOI: 10.1080/22020586.2019.12073015

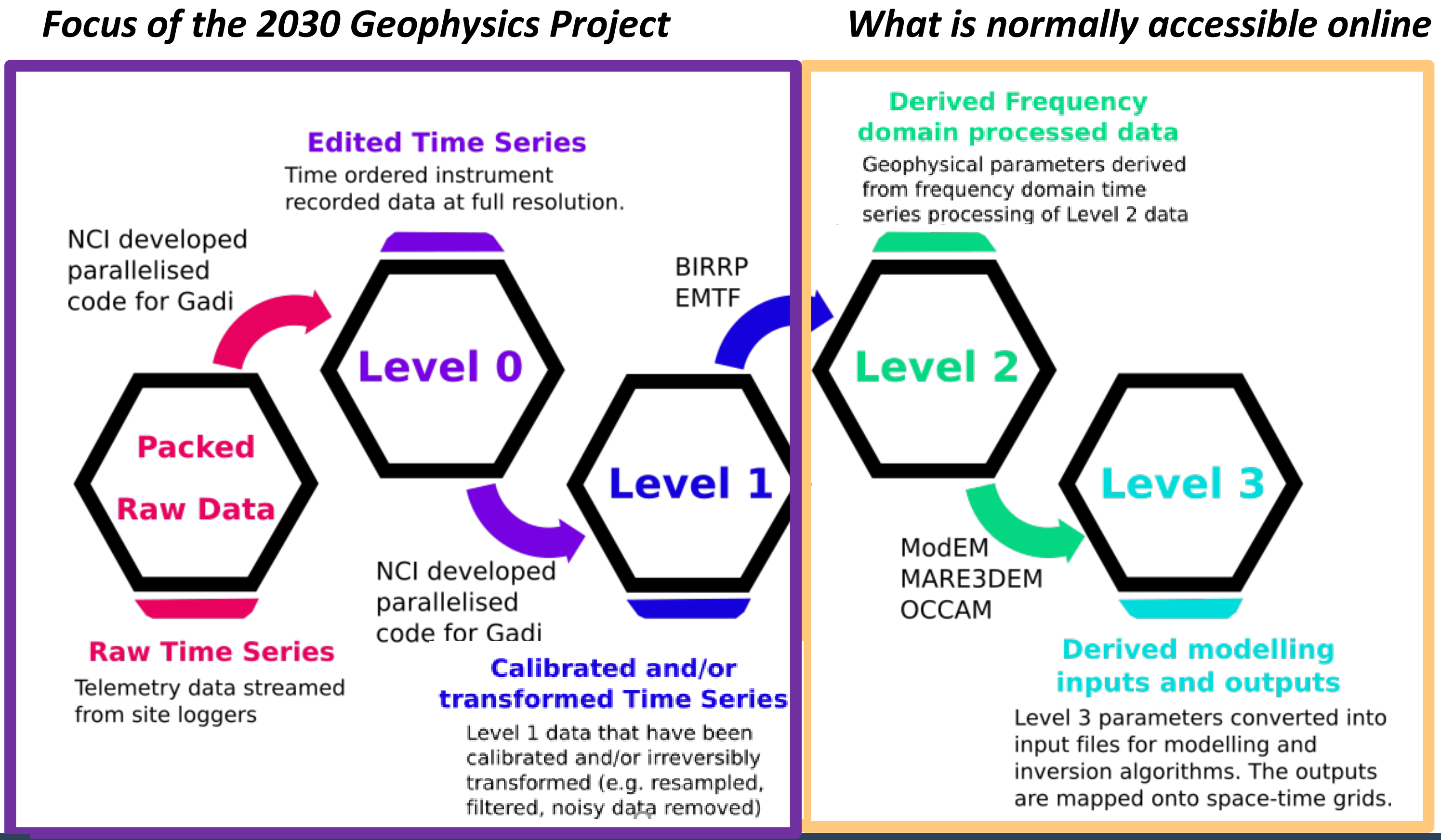
CC BY

The greatest impact is the ability to use less processed data



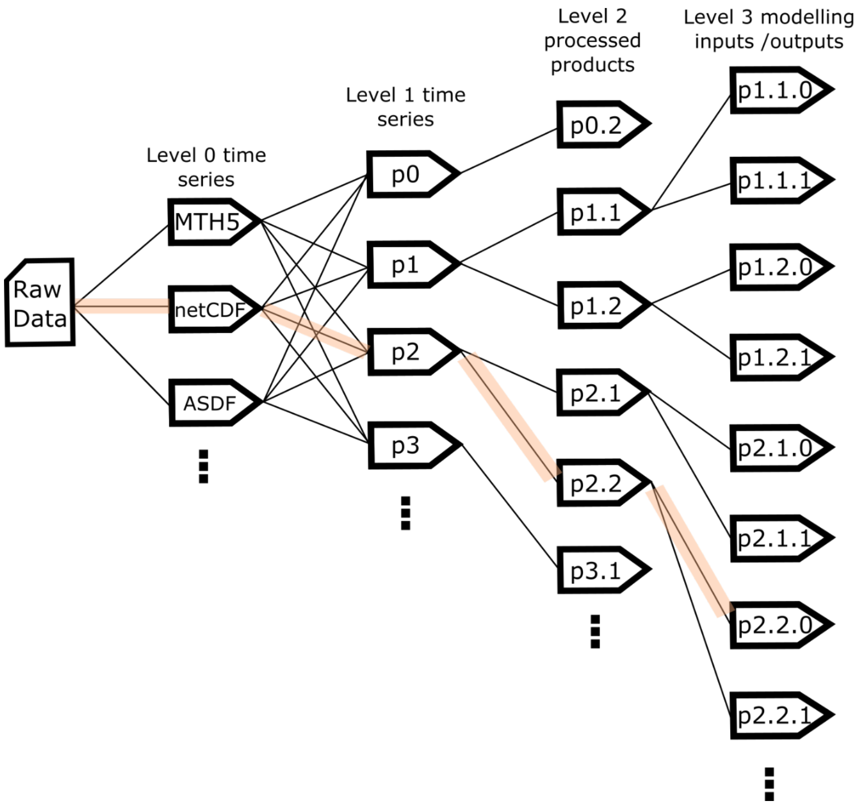
CC BY

Transparently developing your own products



CC BY

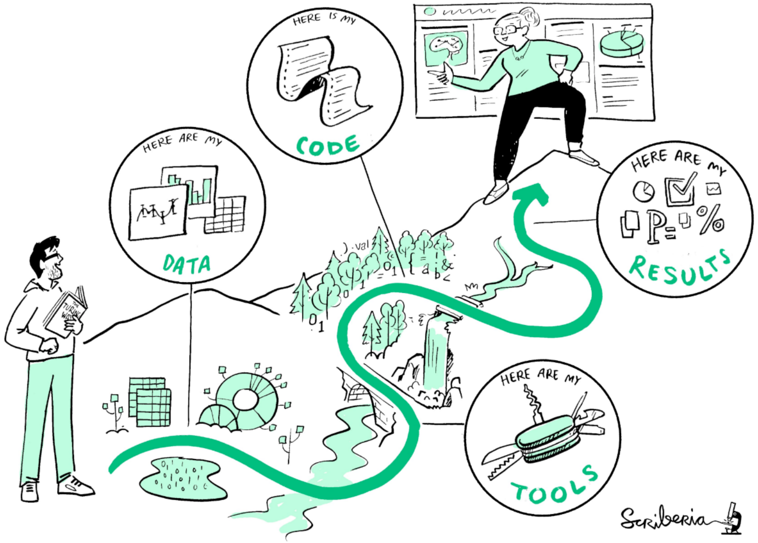
Transparent provenance between processing levels



CC BY

The Turing Way: Reproducibility in the computational era

Historically computationally based research involved an individual using their own data and processing it on their own private area, often using software they wrote or inherited from close collaborators.



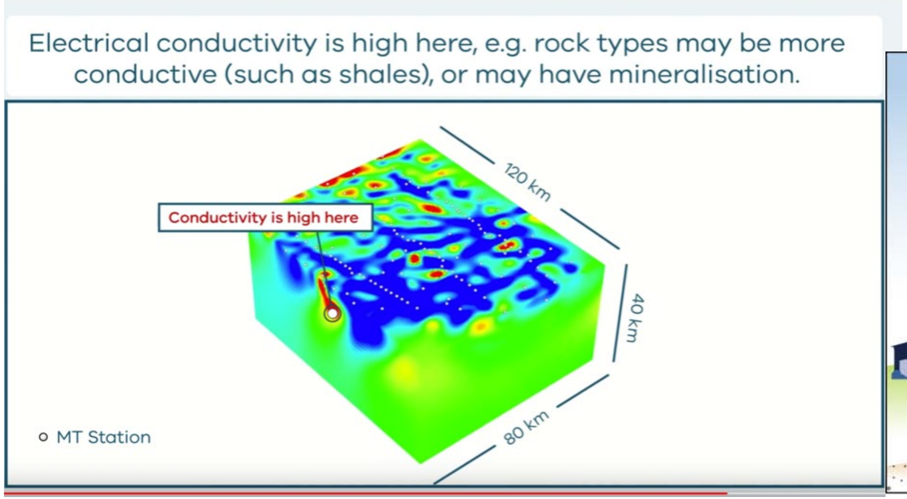
Today's research is likely to be part of a large team (could be global) that will;

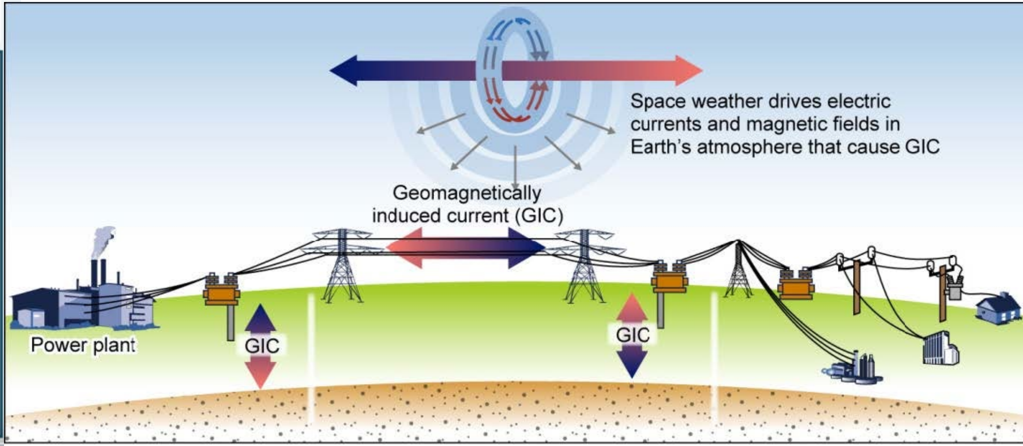
- ➢ use a subset of data from an external repository or multiple repositories
- ➢ Process the data on a public or private cloud or a large centralised supercomputer
- ➢ use a mixture of their own code, third party software, data services, libraries, stable global community codes

https://doi.org/10.5281/zenodo.3332807
https://the-turing-way.netlify.app/reproducible-research/reproducible-research.html

CC BY

Why is it important?: same data for different purposes





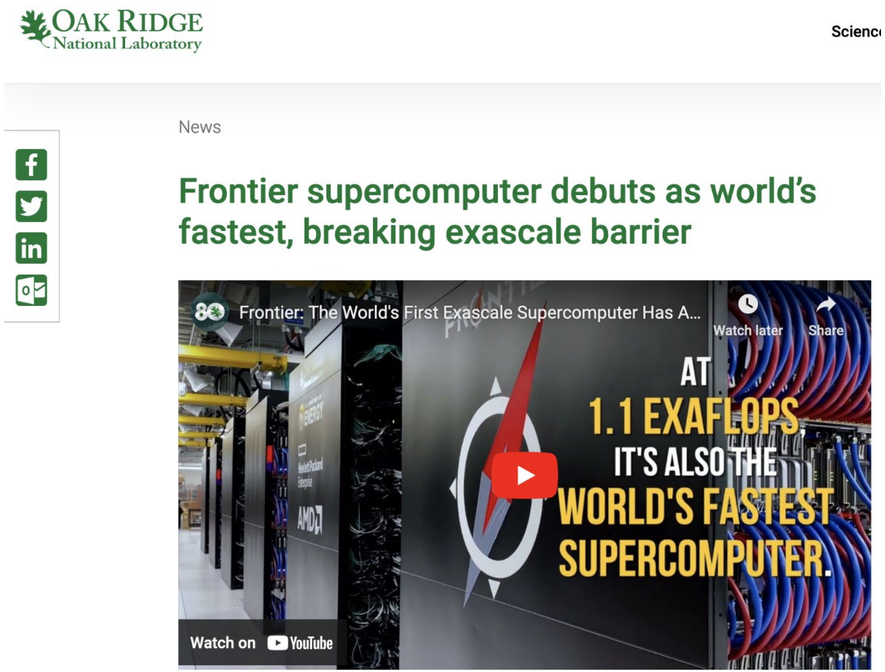
Sources: GAO (presentation); Art Explosion (images). | GAO-19-98

https://www.youtube.com/watch?v=I0c7rfaRm6g

https://www.gao.gov/products/gao-19-98

CC BY

Exascale computing is already a reality in the USA: Frontier



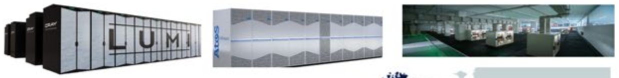
https://www.ornl.gov/news/frontier-supercomputer-debuts-worlds-fastest-breaking-exascale-barrier

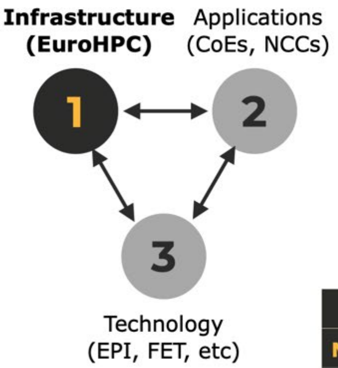
https://www.ornl.gov/news/ornl-celebrates-launch-frontier-worlds-fastest-supercomputer

CC BY

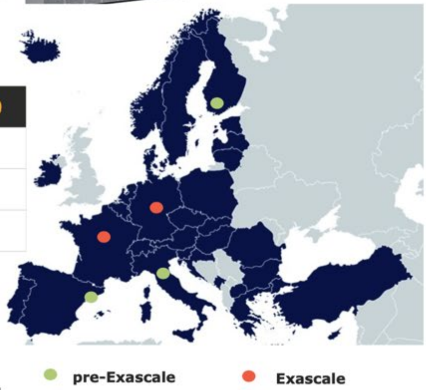
European Preparation for Exascale

The European HPC ecosystem (pillar 1)





tier-0 (pre-exascale)			T500
LUMI	Finland	500 PFlop/s	3
Leonardo	Italy	330 PFlop/s	4
MN-5	Spain	250 PFlop/s	-



tier-1 (petascale)			T500
MeluXina	Luxemburg	15 PFlop/s	52
Karolina	Txequia	7 PFlop/s	85
Discover	Bulgaria	5 PFlop/s	123
Vega	Slovenia	4 PFlop/s	140
Deucalio	Portugal	10 PFlop/s	-

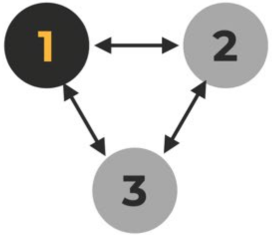
Slide from Arnau Folch, Centre of Excellence in Exascale computing for Solid Earth (ChEESE): https://cheese-coe.eu/

CC BY

European Preparation for Exascale

The European HPC ecosystem (pillar 1)

Infrastructure (EuroHPC) Applications (CoEs, NCCs)



Technology (EPI, FET, etc)

tier-1 (petascale)			T500
MeluXina	Luxemburg	15 PFlop/s	52
Karolina	Txequia	7 PFlop/s	85
Discover	Bulgaria	5 PFlop/s	123
Vega	Slovenia	4 PFlop/s	140
Deucalio	Portugal	10 PFlop/s	-

pre-Exascale Exascale

IEEE Spectrum FOR THE TECHNOLOGY INSIDER

Type to search

FEATURE COMPUTING

EUROPE GETS AN EXASCALE SUPERCOMPUTER

Germany will host JUPITER, Europe's entry into the exascale realm

https://spectrum.ieee.org/europe-s-exascale-supercomputer

Slide from Arnau Folch, Centre of Excellence in Exascale computing for Solid Earth (ChEESE): https://cheese-coe.eu/

CC BY

The Australian 2030 Geophysics Collection Project

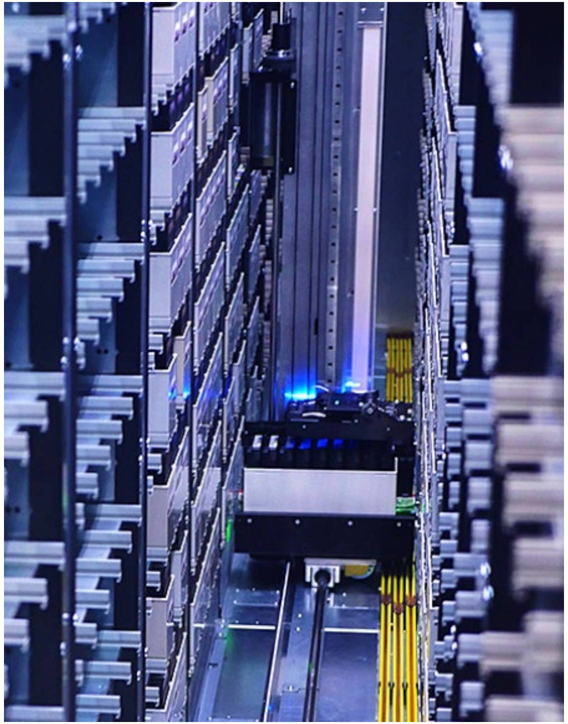
- 2030 is an R&D project funded through a collaboration between AuScope, National Computational Infrastructure (NCI), Terrestrial Ecosystems Research Network (TERN) and the Australian Research Data Commons (ARDC)

- The project seeks to:
  a. Make national-scale high-resolution geophysics datasets suitable for programmatic access in HPC environments;
  b. Lay the foundations for more rapid data processing by 2030 next-generation scalable, data-intensive computation including Artificial Intelligence (AI)/Machine Learning (ML) and data assimilation.
- The project is NOT about building systems for the infrastructures and stakeholder requirements of today.
- Rather it is about positioning Australian geophysical data collections to be capable of taking advantage of next generation technologies and computational infrastructures by 2030.
- Project website: (https://ardc.edu.au/project/2030-geophysics-collections/)

CC BY

What do we know about 2030 computing?

- High-end computational power will be at exascale
- Today's emerging collaborative platforms will continue to evolve as a mix of HPC and cloud
- Data volumes will be measured in Zettabytes (10^{21} bytes), which is about 10 times more than today
- **It will be mandatory for data discovery, accessibility, interoperability and reusability to be fully machine-to-machine as envisaged by the FAIR principles in 2016**



CC BY

HPC has reduced processing times significantly

Testing Parallel I/O via NCI-geophys environment on the AusLAMP Musgraves Province time series data: https://doi.org/10.25914/58gr-1550

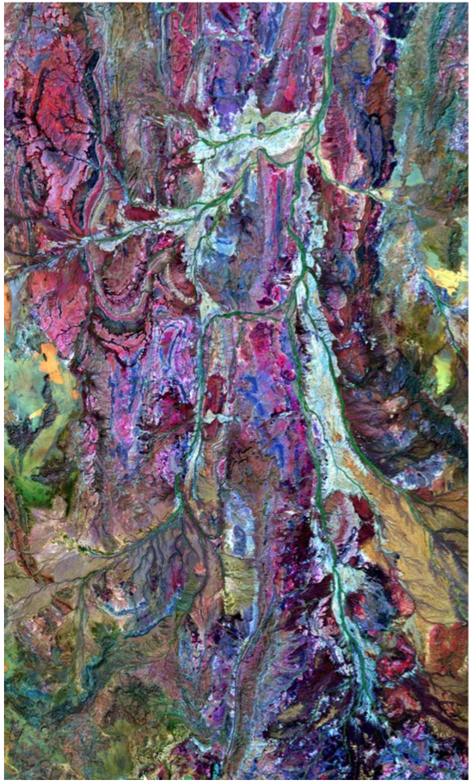
Dataset of 93 MT stations	Serial I/O	MPI based Parallel I/O (96 cores)
Level 0: one MTH5/mt_metadata file for all stations	~ 14 hours	~ 35 minutes
Level 0: one MTH5/mt_metadata file per station	~ 5 hours 47 minutes	~ 4 minutes
Level 1: one MTH5/mt_metadata file per station	~ 49 minutes	~ 1.2 minutes
Level 2: one EDI file per station	~ 2 hours 30 minutes	~ 2 minutes

CC BY

What are the opportunities of 2030 computing?

1. So often today's research is undertaken on pre-canned, analysis-ready datasets (ARD) that are tuned towards the highest common denominator as determined by the data owner/publisher.

1. By 2030:
   - Increased computational power co-located with fast-access storage systems will mean that geoscientists will be able to work on less processed data levels and then transparently develop their own derivative products.
   - Researchers will be able to see the quality of their algorithms more rapidly: there will be multiple versions of open source software used as researchers fine tune individual algorithms to suit their specific requirements.
   - We will be capable of more precise solutions and in hazards space and other relevant areas, analytics will be done in faster-than-real-time.



Landsat-8 image courtesy of the U.S. Geological Survey

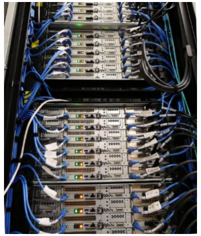
Wyborn, L., Rees, N., Klump, J., Evans, B., Rawling, T., and Druken, K.: The Known Knowns, the Known Unknowns and the Unknown Unknowns of Geophysics Data Processing in 2030 , EGU General Assembly 2022, Vienna, Austria, 23–27 May 2022, EGU22-11012, https://doi.org/10.5194/egusphere-egu22-11012 , 2022.

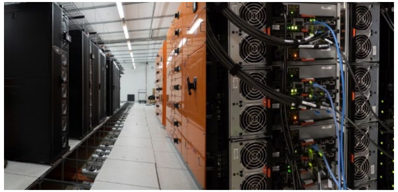
CC BY

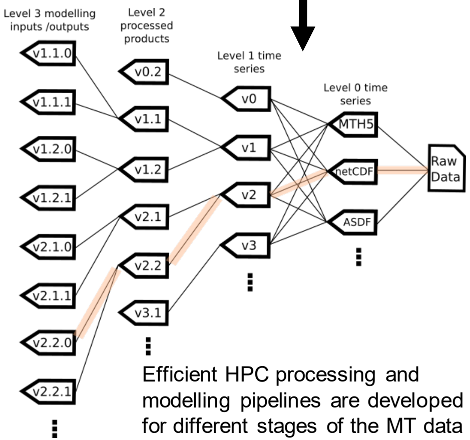
2030 Dirt to Desktop publishing pipelines

Raw data is collected in the field with internationally standardised metadata

Data is transferred to HPC facility as soon as there is a decent internet connection



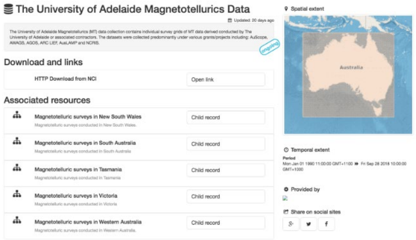




Efficient HPC processing and modelling pipelines are developed for different stages of the MT data life cycle

Each MT product is run through community agreed QA/QC procedures before publishing

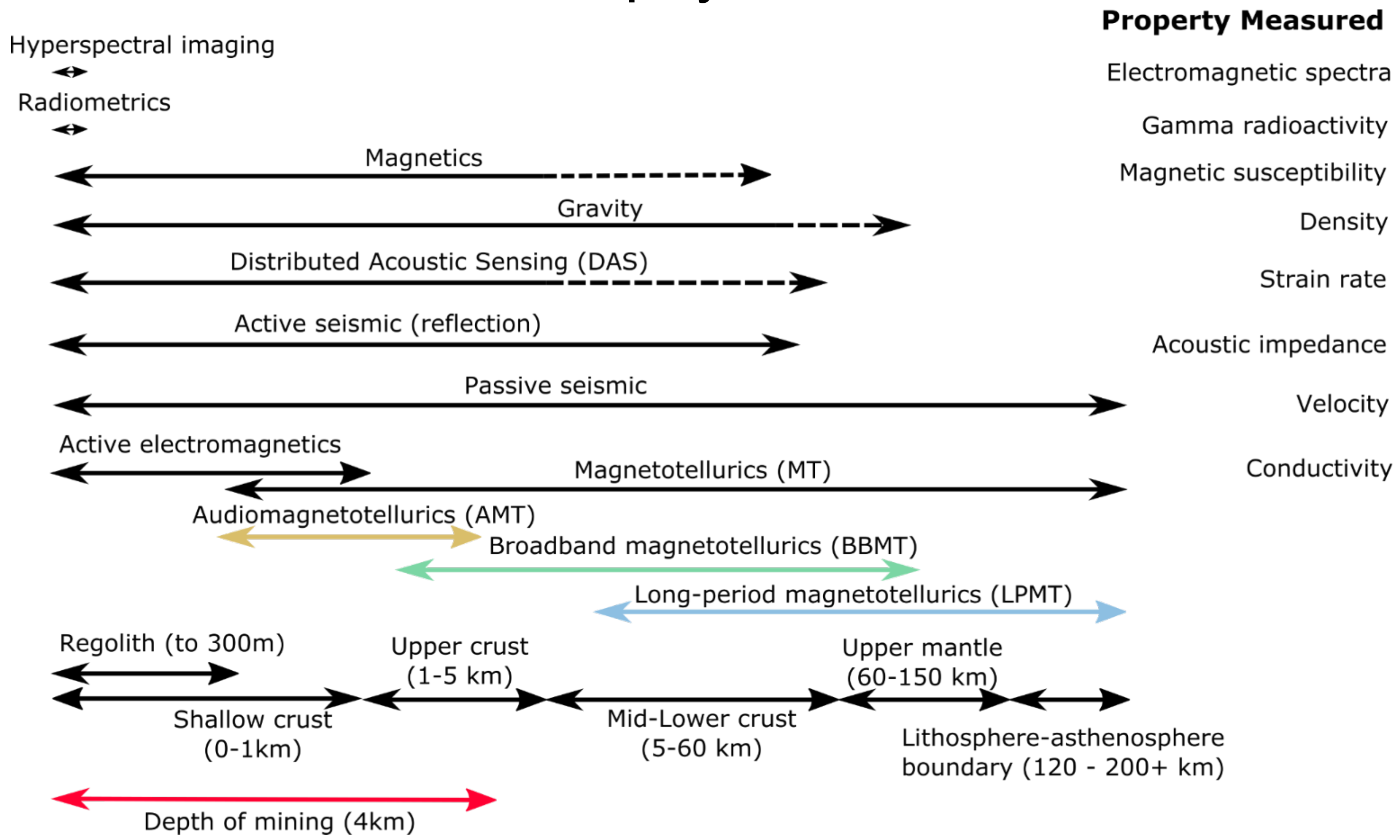
Curation and publication: persistent identifiers to help enable transparent data management and support reusability and reproducibility of workflows



Nigel Rees, Sheng Wang, Ben Evans, Lesley Wyborn, Tim Rawling, Bruce Goleby, Kelsey Druken, & Rui Yang. (2021). Using the NCI Gadi Supercomputer to revolutionise processing of MT time series data: results from the GeoDeVL experiment. Australian Society of Exploration Geophysicists Extended Abstracts, Volume 2021, 3rd Australasian Exploration Geoscience Conference, Brisbane, 2021. https://doi.org/10.5281/zenodo.7690550

CC BY

OneGeophysics



CC BY

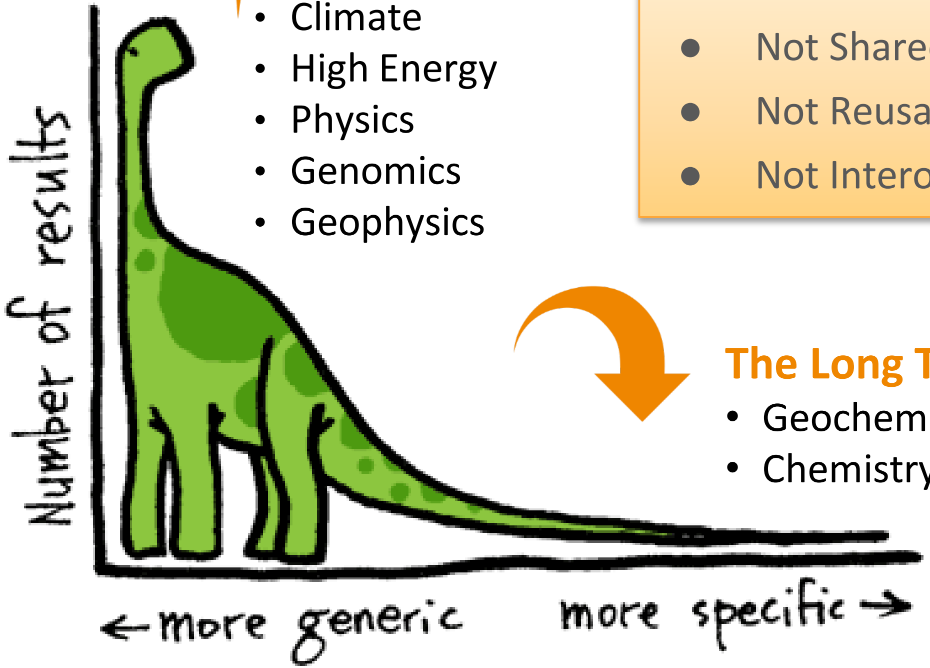
Geoscience Data is of two types - Really Big and Long Tail: but sharing is not usual

The Head:

- Astronomy
- Climate
- High Energy
- Physics
- Genomics
- Geophysics

But either is

- Not Shared
- Not Reusable
- Not Interoperable



The Long Tail:

- Geochemistry
- Chemistry

CC BY

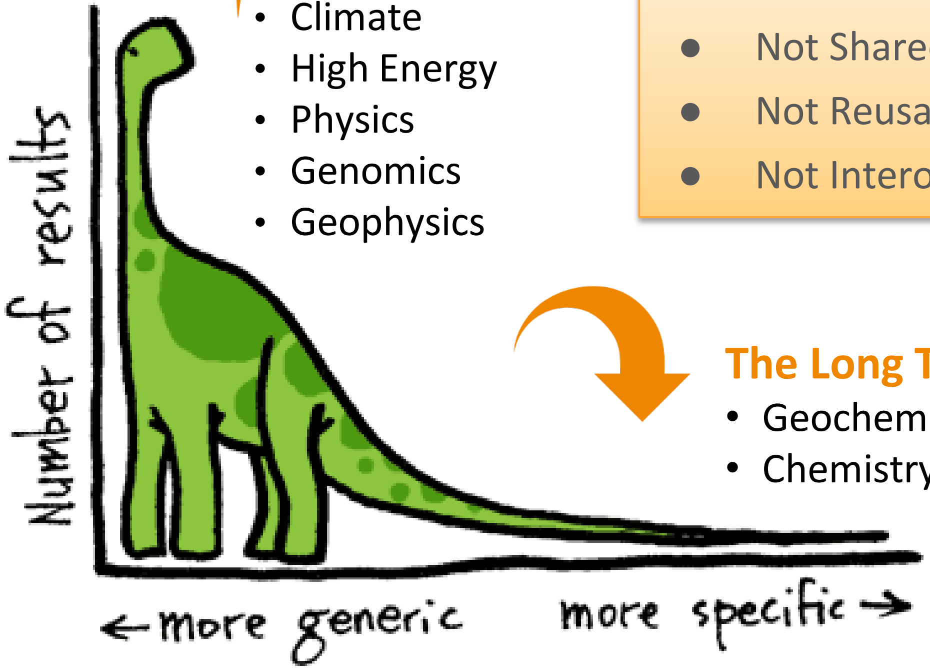
Geoscience Data is of two types - Really Big and Long Tail: but sharing is not usual

The Head:

- Astronomy
- Climate
- High Energy
- Physics
- Genomics
- Geophysics

But either is

- Not Shared
- Not Reusable
- Not Interoperable



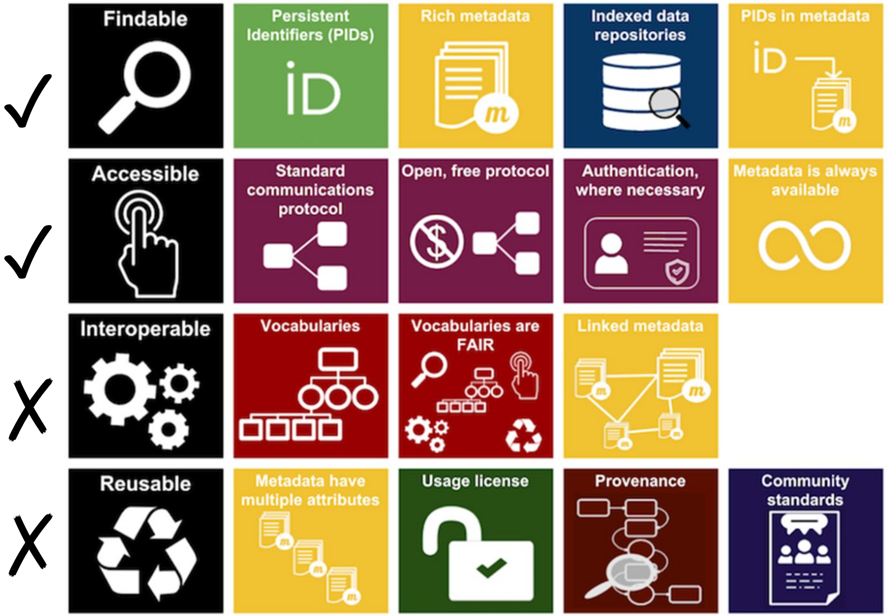
The Long Tail:

- Geochemistry
- Chemistry

https://sites.gatech.edu/admission-blog/2017/05/16/admission-its-not-fair/

CC BY

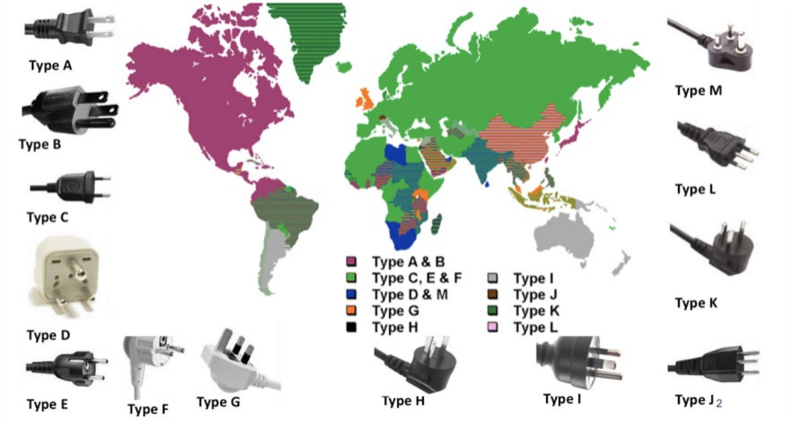
Are our data sets really FAIR?



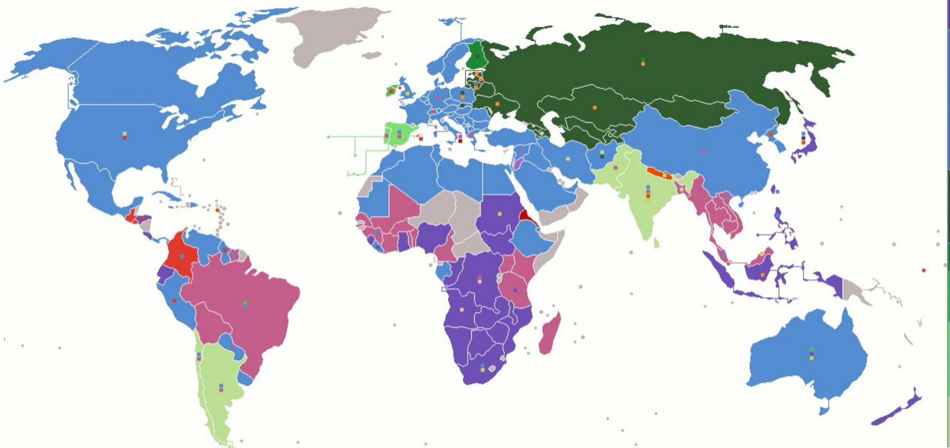
- FAIR principles published by Wilkinson et al (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3:160018 https://doi.org/10.1038/sdata.2016.18
- Data that is Findable, Accessible, Interoperable and Reusable by humans and machines
- Few realise that much that is described as fully FAIR compliant is not, because it is
    - NOT fully machine readable
    - NOT Interoperable or reusable

CC BY

Converging infrastructures internationally is not new



https://www.power-plugs-sockets.com/



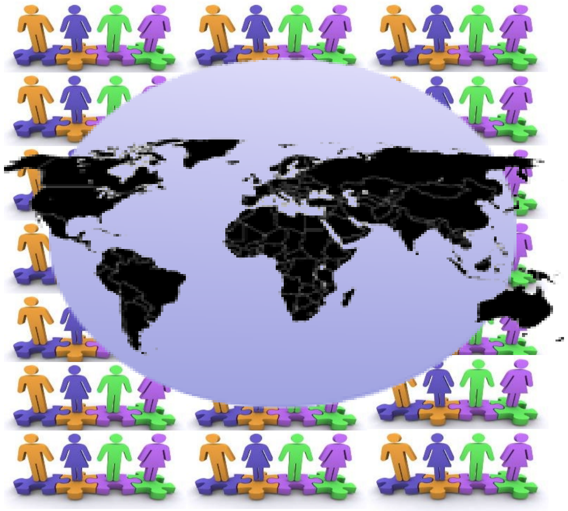
Gauge	Name
1,000 mm (3 ft 3$\frac{3}{8}$ in)	Metre gauge
1,067 mm (3 ft 6 in)	Three foot six inch gauge
1,435 mm (4 ft 8$\frac{1}{2}$ in)	Standard gauge
1,520 mm (4 ft 11$\frac{27}{32}$ in)	Five foot and 1520 mm gauge
1,524 mm (5 ft)	Finnish gauge
1,600 mm (5 ft 3 in)	Five foot three inch gauge
1,668 mm (5 ft 5$\frac{21}{32}$ in)	Iberian gauge
1,676 mm (5 ft 6 in)	Five foot six inch gauge

CC BY

NCI AUSTRALIA

The size of your community counts

- The growing need to share data, information and services across multiple disciplines and organisations
- This requires standards for machine readability
- Increasingly digital data collections need to be reused and repurposed by much broader communities
- The size of the community that you interoperate with is as large as the size of the community that uses your standard
- Where possible we need to converge on international standards



BY

Example: Variety of Analytical Methods

Credit: Kerstin Lehnert

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

EMPA	Raman	XCT	VLM	QRIS	GC-MS	LC-MS	VNMIR	NanoSIMS	SLS
μL^{2}MS	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						

CC BY

NCI AUSTRALIA

Common words we can use across disciplines

Remote sensing
Sensor
Value
Parameter
Scene

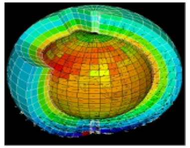
Geophysics
Algorithm, code, simulator
Model, field
Variable
Volume, grid

Metrology
Instrument
Value
Measurand
Sample

Geochemistry
Instrument, analytical process
Analysis
Analyte
Sample

Environmental monitoring
Gauge, sensor
Value, time-series
Parameter
Station

Observations & Measurements
procedure
result
observed property
feature of interest



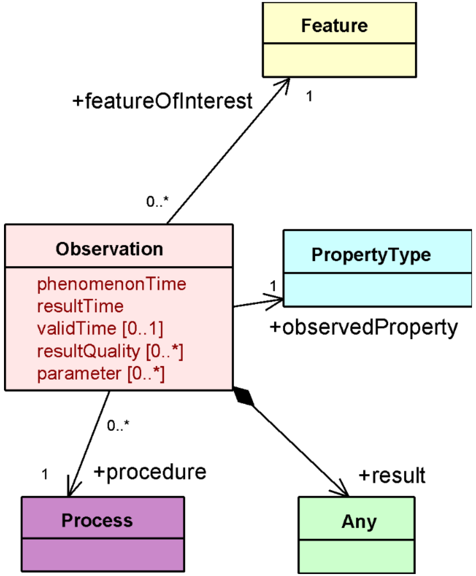
Credit: Simon Cox

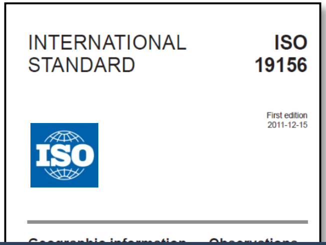
CC BY

NCI AUSTRALIA

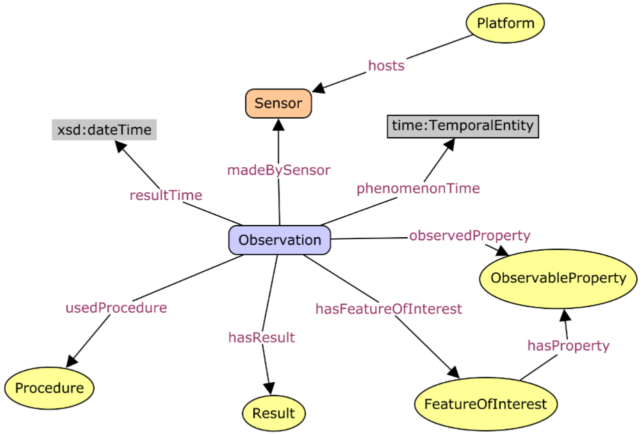
Introducing the Observation and Measurement Standard

Observation & Measurement Model in XML



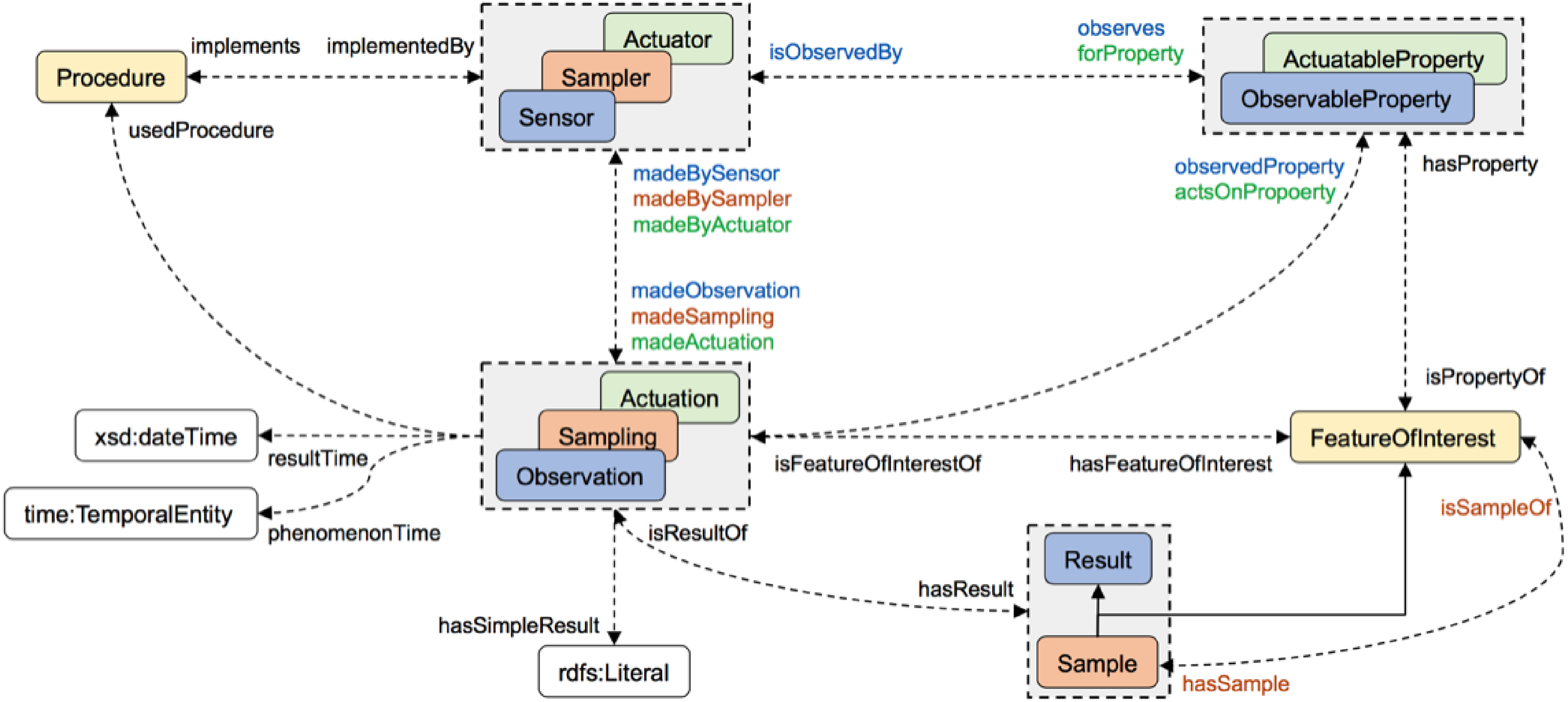


Observation & Measurement Model in RDF (Linked Data)



https://www.w3.org/TR/vocab-ssn/

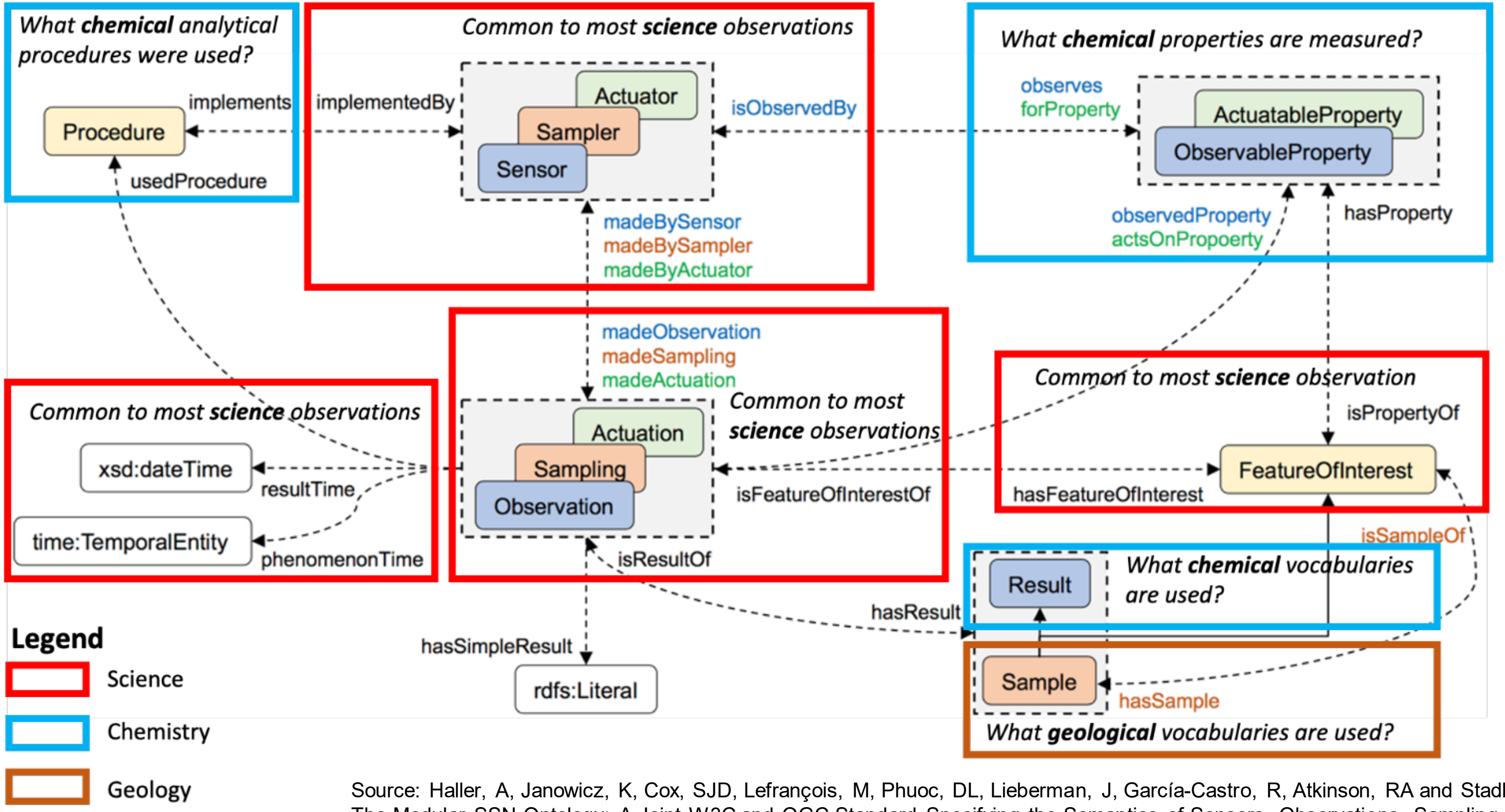
CC BY



Source: Haller, A, Janowicz, K, Cox, SJD, Lefrançois, M, Phuoc, DL, Lieberman, J, García-Castro, R, Atkinson, RA and Stadler, C. 2019. The Modular SSN Ontology: A Joint W3C and OGC Standard Specifying the Semantics of Sensors, Observations, Sampling, and Actuation. *Semantic Web*, 10(1): 9–32. https://doi.org/10.3233/SW-180320

CC BY

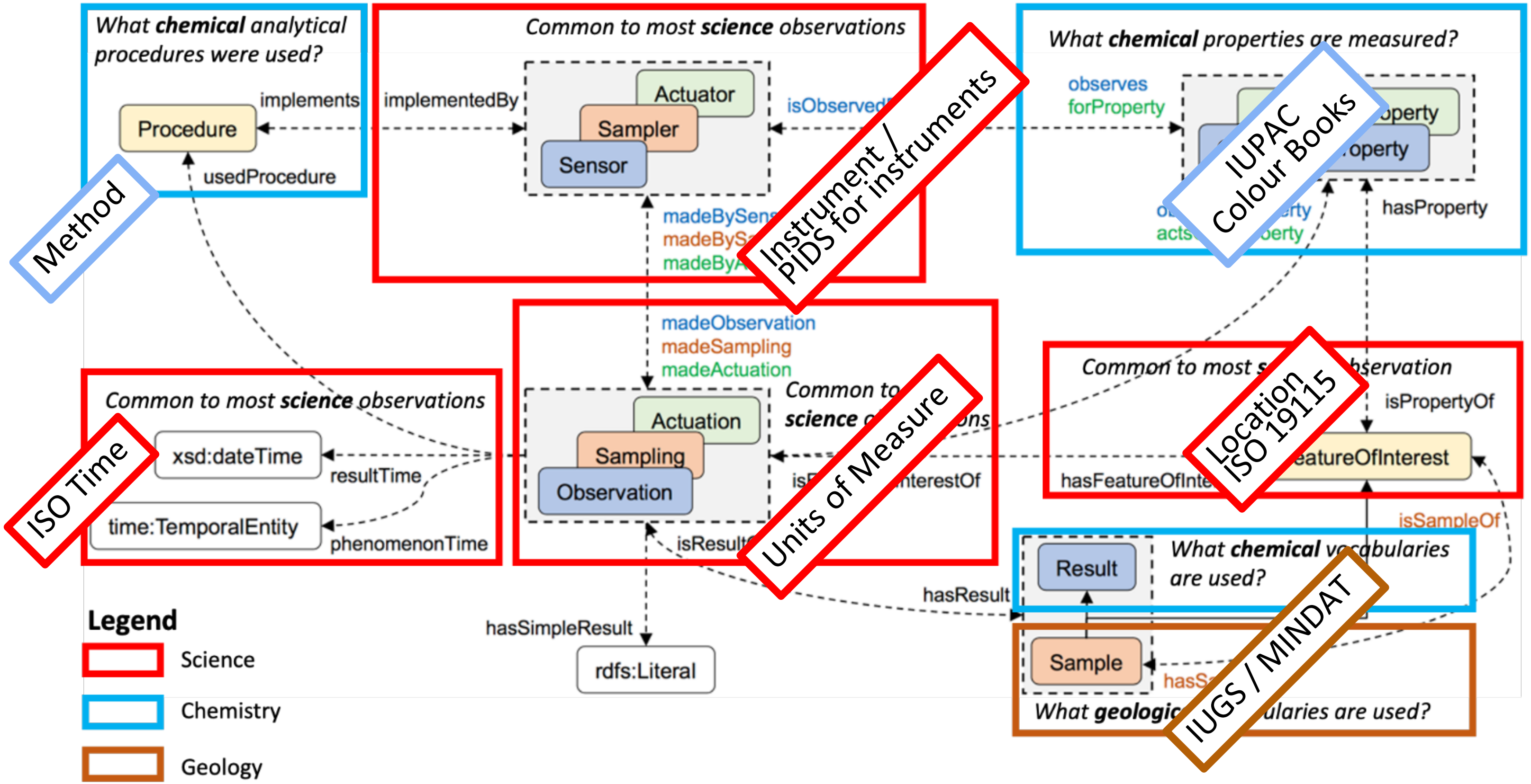
Components to Address in Geochemistry



Source: Haller, A, Janowicz, K, Cox, SJD, Lefrançois, M, Phuoc, DL, Lieberman, J, García-Castro, R, Atkinson, RA and Stadler, C. 2019. The Modular SSN Ontology: A Joint W3C and OGC Standard Specifying the Semantics of Sensors, Observations, Sampling, and Actuation. *Semantic Web*, 10(1): 9–32. https://doi.org/10.3233/SW-180320

CC BY

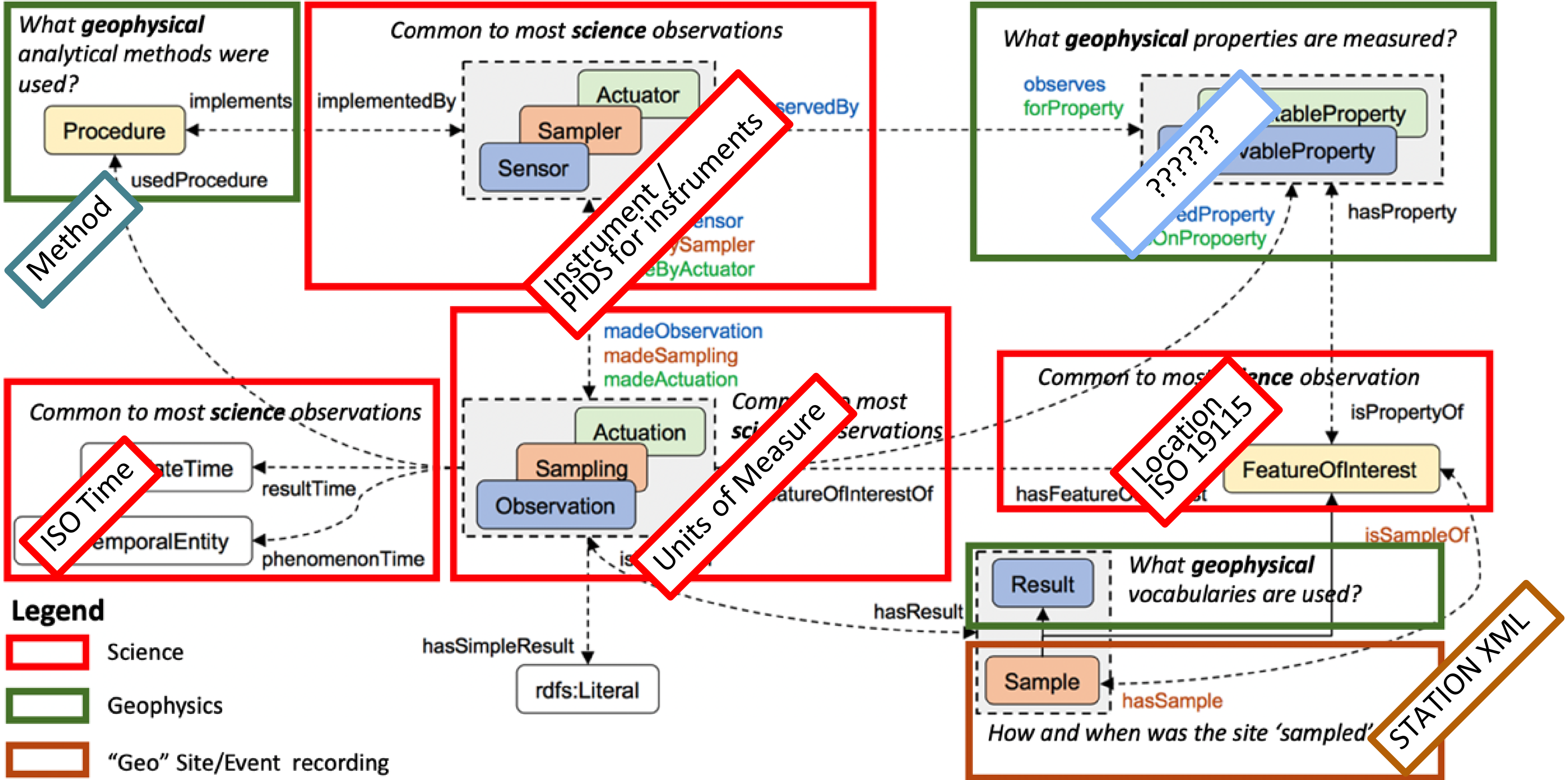
Summary of components that could be repurposed



CC BY

Source: Haller, A, et al. 2019. The Modular SSN Ontology: A Joint W3C and OGC Standard Specifying the Semantics of Sensors, Observations, Sampling, and Actuation. *Semantic Web*, 10(1): 9–32. https://doi.org/10.3233/SW-180320

Repurposing for Geophysics

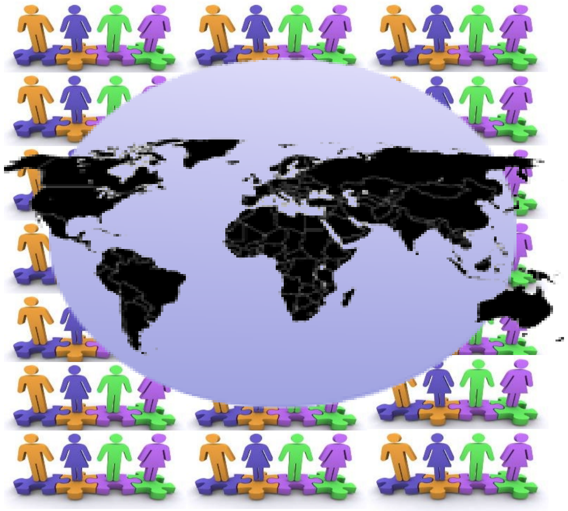


CC BY

Source: Haller, A, et al. 2019. The Modular SSN Ontology: A Joint W3C and OGC Standard Specifying the Semantics of Sensors, Observations, Sampling, and Actuation. *Semantic Web*, 10(1): 9–32. https://doi.org/10.3233/SW-180320

The size of your community you share data with counts

- The growing need to share data, information and services across multiple disciplines and organisations
- This requires standards for machine readability
- Increasingly digital data collections need to be reused and repurposed by much broader communities
- The size of the community that you interoperate with is as large as the size of the community that uses your standard
- Where possible we need to converge on international standards

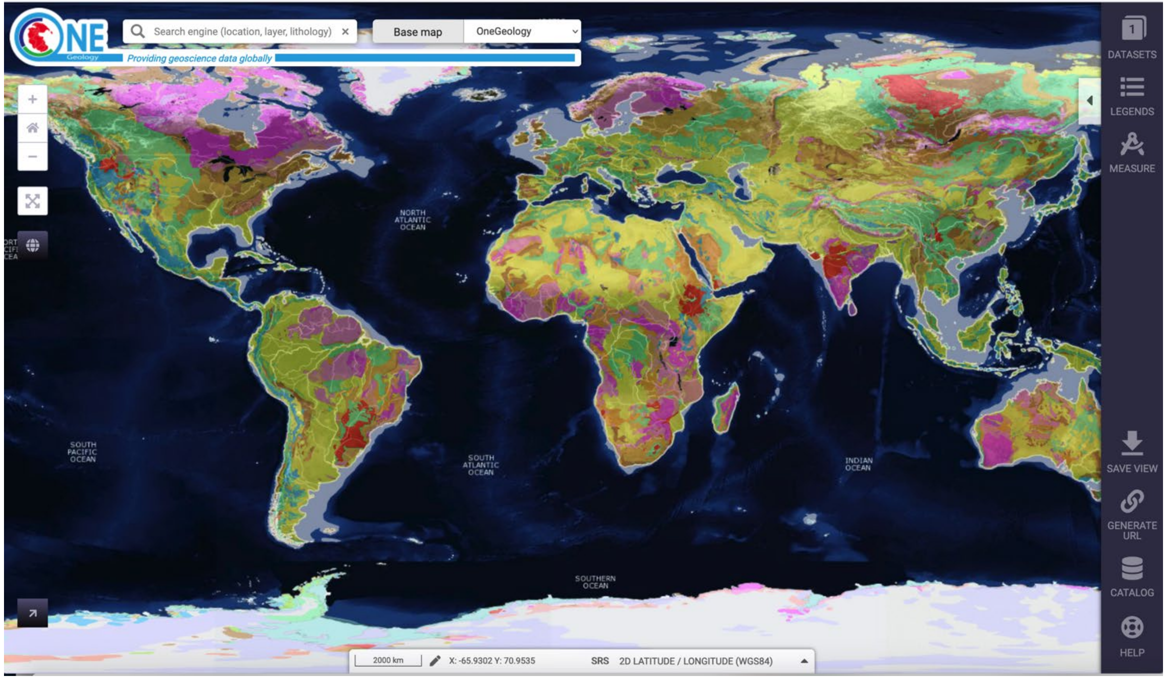


CC BY

NCI AUSTRALIA

https://www.onegeology.org/

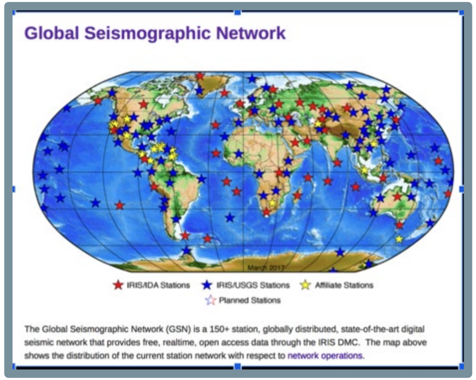
It can be done: OneGeology



OneGeology proved that distributed frameworks with geological map data globally accessible using standard web services and agreed standards and vocabularies would work.

CC BY

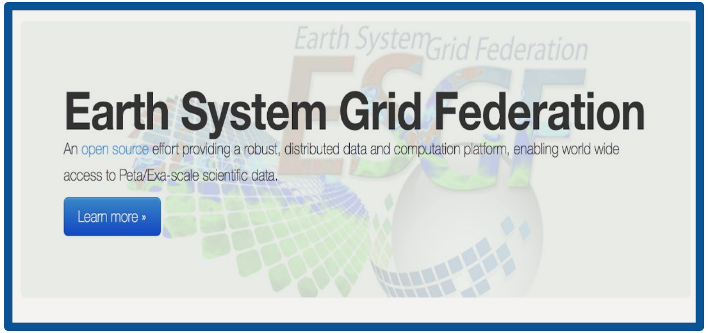
Other disciplines have created global networks...

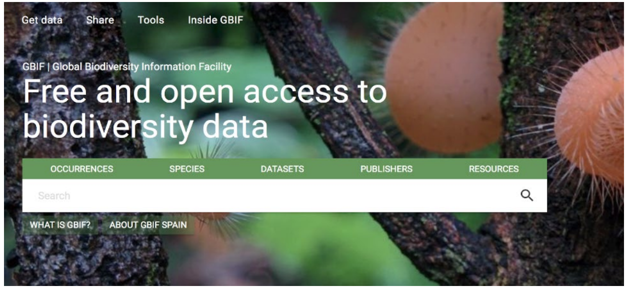


Crystallographic Information Framework

The International Union of Crystallography is the sponsor of the **Crystallographic Information Framework**, a standard for information interchange in crystallography.

The acronym CIF is used both for the *Crystallographic Information File*, the data exchange standard file format of Hall, Allen & Brown (1991) (see Documentation), and for the *Crystallographic Information Framework*, a broader system of exchange protocols based on data dictionaries and relational rules expressible in different machine-readable manifestations, including, but not restricted to, Crystallographic Information File and XML.





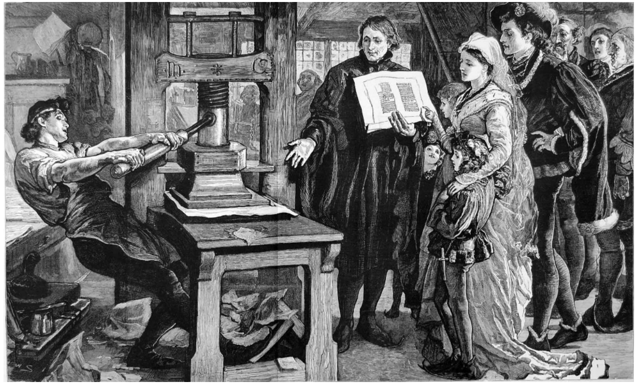
IUPAC Color Books
An authoritative resource for chemical nomenclature, terminology, and symbols. Terminology definitions are drafted by international committees of experts in the appropriate chemistry subdisciplines

CC BY

Parallels in History

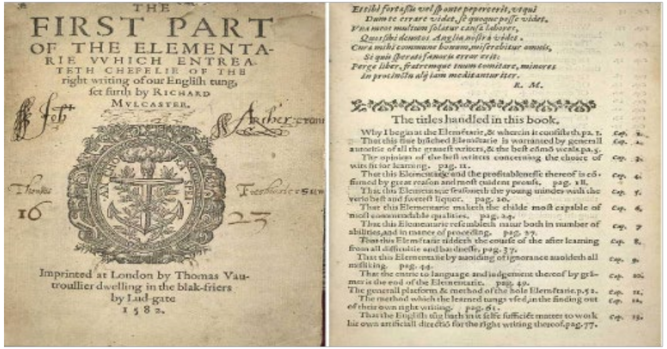
Discover how Johannes Gutenberg's printing press increased the literacy and education of people in Europe

https://www.britannica.com/video/171689/history-printing-press-work-discussion-Johannes-Gutenberg



https://www.ripleys.com/weird-news/gutenberg-printing-press/

https://www.bl.uk/learning/timeline/external/mulcaster-tl.jpg



THE FIRST PART OF THE ELEMENTARIE VVHICH ENTREATETH CHEFELIE OF THE right writing of our English tung, set furth by RICHARD MVLCASTER.

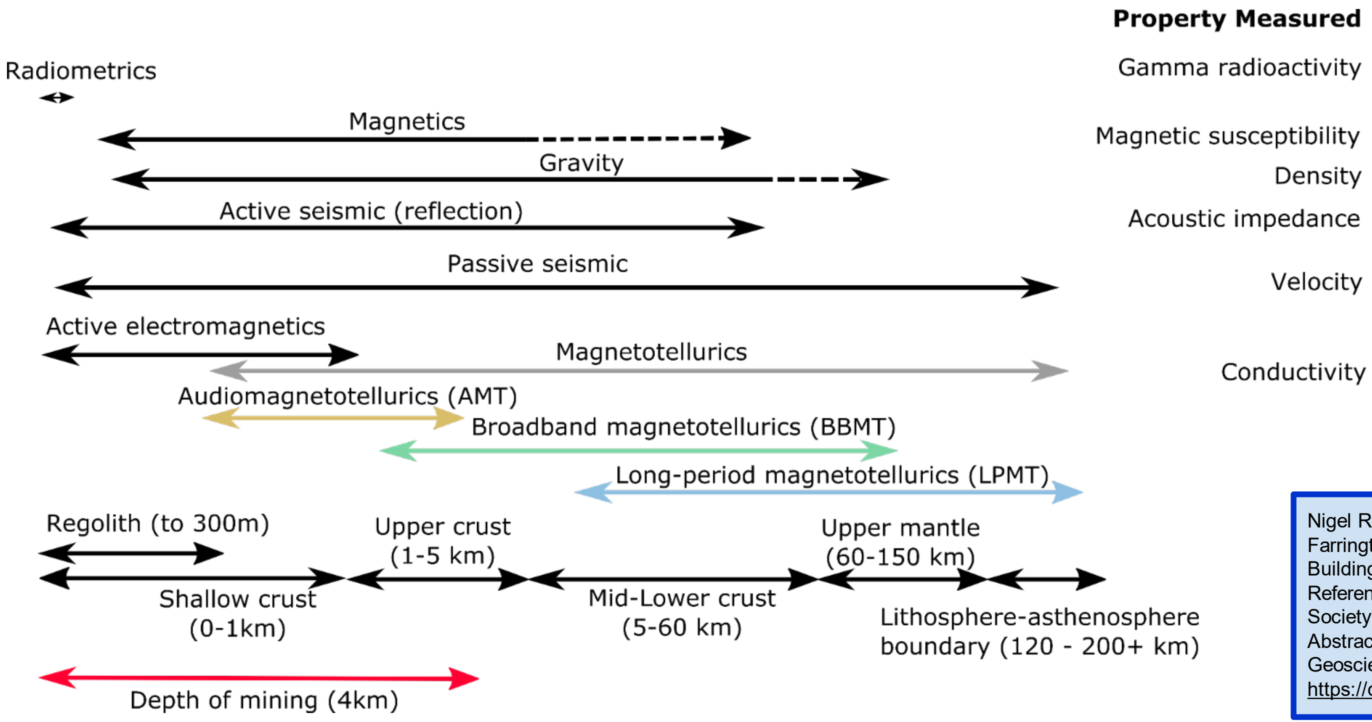
Imprinted at London by Thomas Vautroullier dwelling in the blak-friers by Lud-gate 1582.

The titles handled in this book.

The printing press in 1436, but did not take off until dictionaries that sorted out spelling around 1604. Note: the first modern English dictionary was 1755

CC BY

Multiphysics Analysis but how do we harmonise the data...

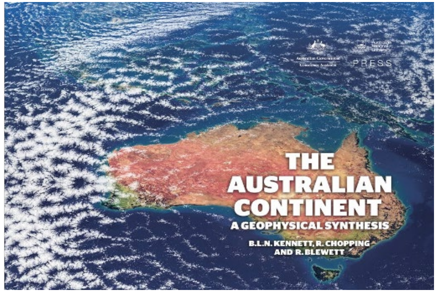


Nigel Rees, Lesley Wyborn, Ben Evans, Rebecca Farrington, Tim Rawling, Rui Yang, & Yue Sun. (2023). Building a National High-Resolution Geophysics Reference Collection for 2030 Computation. Australian Society of Exploration Geophysicists Extended Abstracts, Volume 2023, 4th Australasian Exploration Geoscience Conference, Brisbane, 2023. https://doi.org/10.5281/zenodo.7980192

Types of geophysical data collected in Australia, the physical property measured and the depth of the crust that is sampled: also shown is the depth of current mining. Figure modified from original of Richard Chopping (GSWA).

CC BY

Can we create a collaborative National High Resolution Geophysics Data Platform from all this data? How do I access it?



Brian Kennett, Richard Chopping and Richard Blewett, 2018. The Australian Continent, a Geophysical synthesis. Available on https://press.anu.edu.au/publications/australian-continent#tabanchor

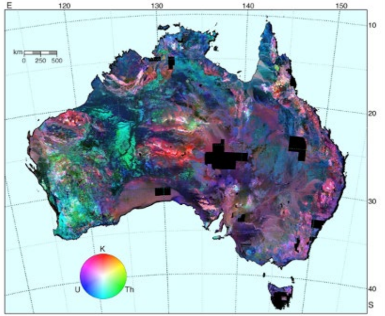


Figure 3.1: Gamma-ray spectrometric map for the continent of Australia created by

Radiometrics

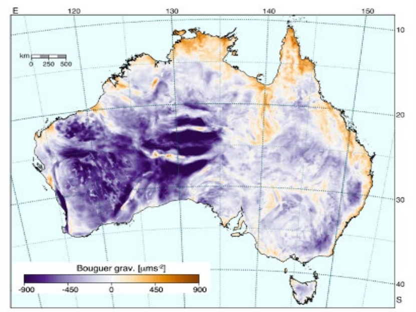


Figure 5.7: Bouguer gravity anomaly.

Bouguer Gravity

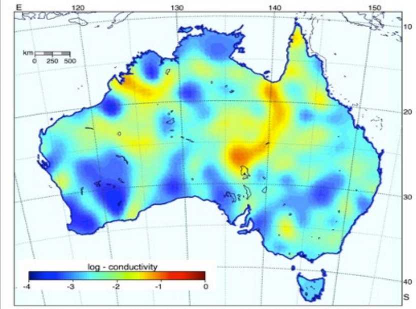


Figure 10.3: Logarithm of electrical conductivity [S/m] at a depth of 52 km.

Electrical Conductivity at 52 km

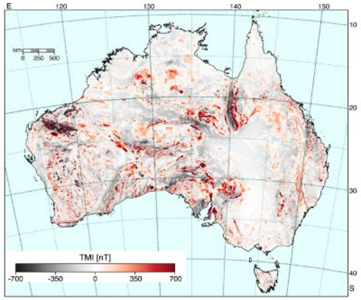


Figure 4.6: Total magnetic intensity across the Australian continent.

Total Magnetic Intensity

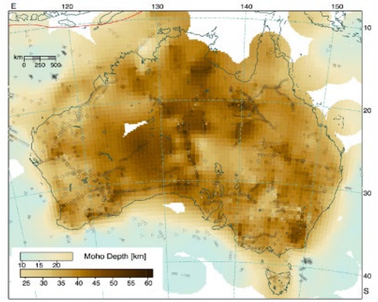


Figure 7.16: Moho surface across Australia utilising the full range of seismic information

Moho Depth

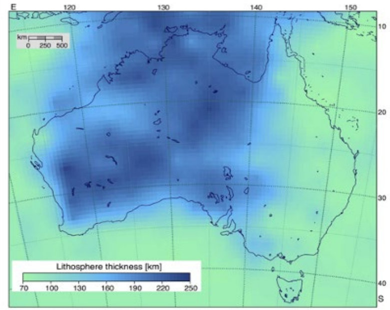


Figure 8.9: Estimate of thickness of the lithosphere across the Australian region.

Thickness of the lithosphere

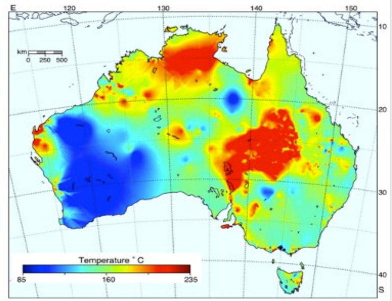


Figure 11.2: OZtemp estimate of the temperature distribution at 5 km depth across the Australian continent.

Temperature at 5km

CC BY

Converting MT time series (meta)data to modern international self-describing standards - it is starting to happen…

ELSEVIER

Computers & Geosciences

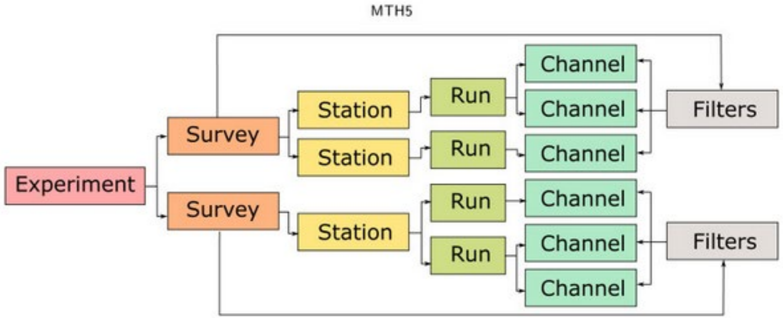
Volume 162, May 2022, 105102

Research paper

MTH5: An archive and exchangeable data format for magnetotelluric time series data

Jared Peacock [a], Karl Kappler [b], Lindsey Heagy [c], Timothy Ronan [d], Anna Kelbert [e], Andrew Frassetto [d]

https://doi.org/10.1016/j.cageo.2022.105102



CC BY

Associations of the International Union of Geodesy & Geophysics (IUGG): can they help coordinate and endorse?

International Union of Geodesy and Geophysics
IUGG
UGGI
Union Géodésique et Géophysique Internationale

IACS
INTERNATIONAL ASSOCIATION OF CRYOSPHERIC SCIENCES (IACS)

IAG
INTERNATIONAL ASSOCIATION OF GEODESY (IAG)

IAGA
INTERNATIONAL ASSOCIATION OF GEOMAGNETISM AND AERONOMY (IAGA)

IAHS AISH
INTERNATIONAL ASSOCIATION OF HYDROLOGICAL SCIENCES (IAHS)

IAMAS
INTERNATIONAL ASSOCIATION OF METEOROLOGY AND ATMOSPHERIC SCIENCES (IAMAS)

IAPSO
INTERNATIONAL ASSOCIATION FOR THE PHYSICAL SCIENCES OF THE OCEANS (IAPSO)

IASPEI
INTERNATIONAL ASSOCIATION OF SEISMOLOGY AND PHYSICS OF THE EARTH'S INTERIOR (IASPEI)

IAVCEI
INTERNATIONAL ASSOCIATION OF VOLCANOLOGY AND CHEMISTRY OF THE EARTH'S INTERIOR (IAVCEI)

CC BY

Our current world is imploding

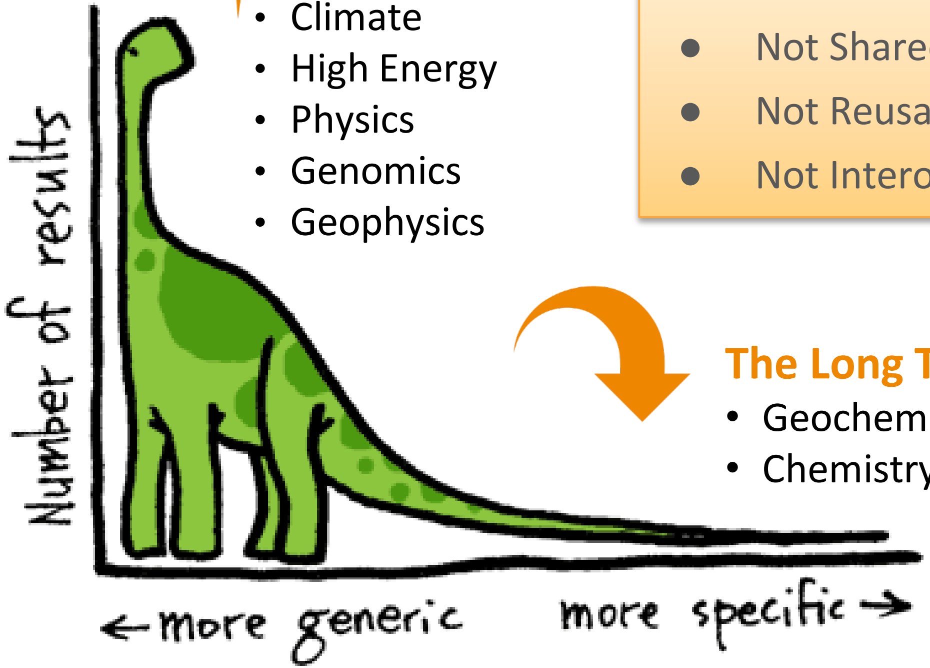
In summary, we are living in a time of great change

1. We have more computation than we can use
   - *But is it necessary - I am happy with what I have?*
2. We are changing to data intensive computing that requires FAIR data and self describing formats
   - *But do we need to use data - can't we just use images?*
   - *Why do we need the actual data?*
3. We can move to working more and more on high resolution data sets
   - *Why - large, high resolution data files are too big for my online GIS system?*
4. Standards are changing to machine actionable
   - *So what - my software can't read the new standards, so I can't change*

https://www.istockphoto.com/vector/earth-like-bomb-with-mathch-in-fire-gm1043315054-279285220?phrase=world%20exploding

CC BY

Issues of the Long Tail



The Head:

- Astronomy
- Climate
- High Energy
- Physics
- Genomics
- Geophysics

But either is

- Not Shared
- Not Reusable
- Not Interoperable

The Long Tail:

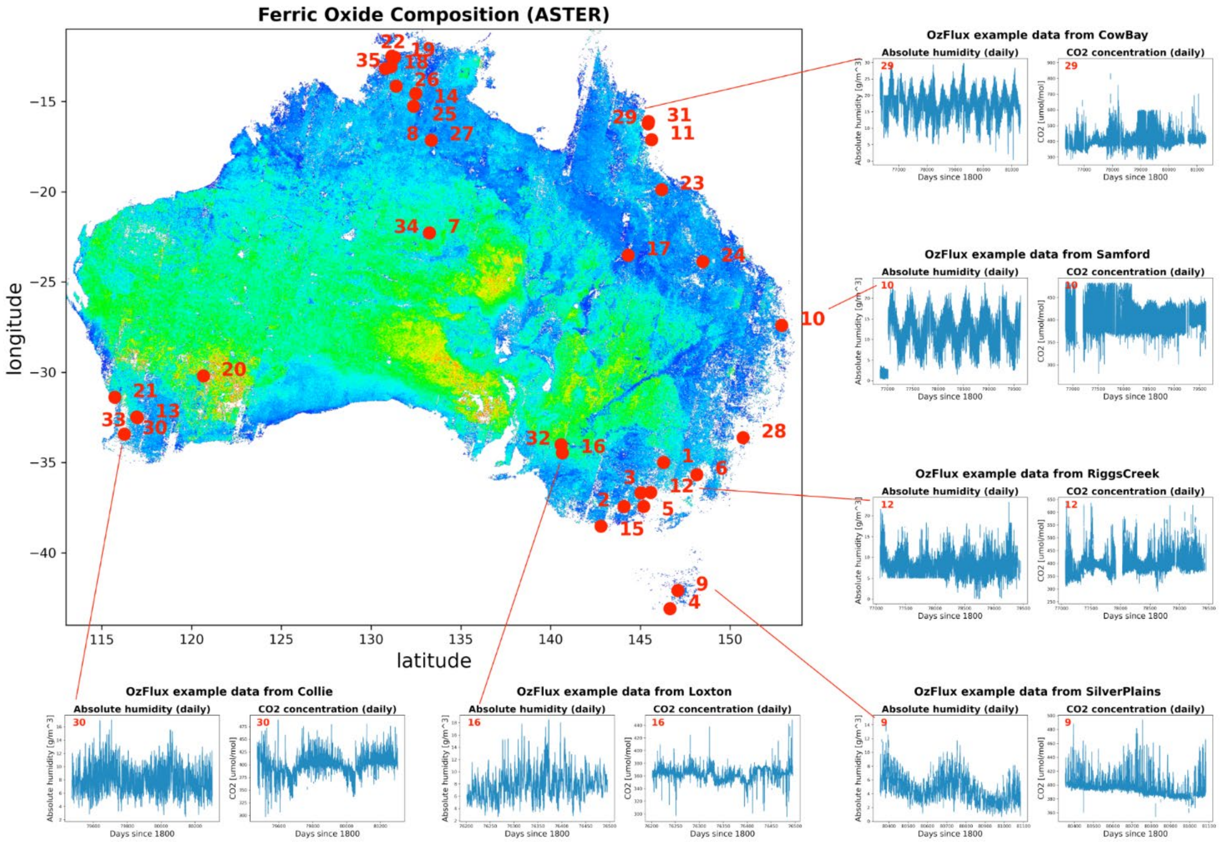
- Geochemistry
- Chemistry

Common Excuses for not sharing

- Often specialized
- Low volume
- Collected by many people
- Heterogeneous
  - In purpose (many disciplines)
  - In provenance (many methods)
- Fragmented data landscape
- Un-curated (on C drives)
- Hard to find & access
- Not persistent
- Difficult to integrate
- Not machine-readable

CC BY

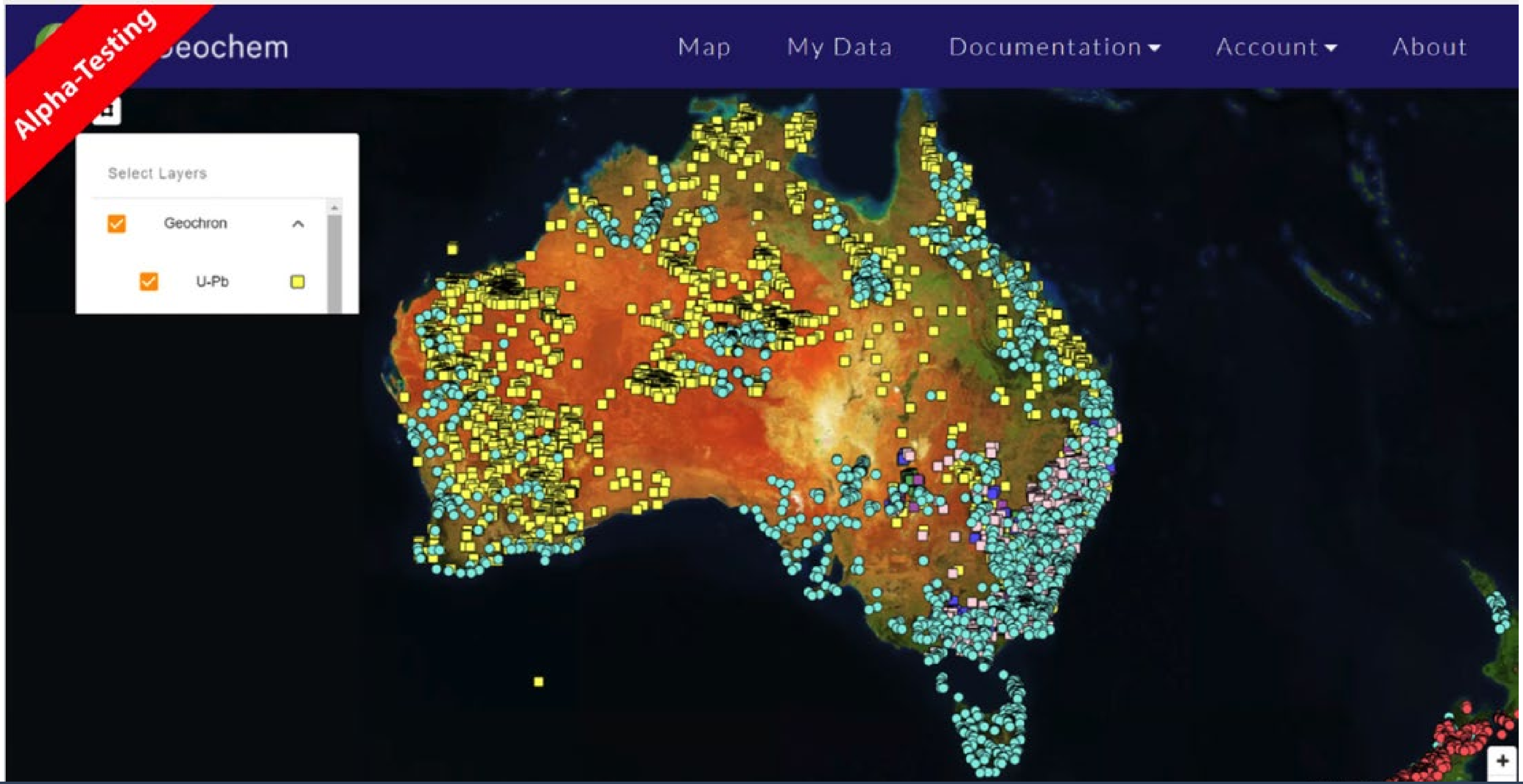
The power of point-located observations



OzFlux tower site locations (red dots) are plotted on the ASTER Ferric Oxide Composition of Australia layer. Time series data for CO_2 and absolute humidity are shown for six OzFlux tower sites. ASTER data were sourced from NCI's Gadi gdata file system and the OzFlux data were drawn from the TERN THREDDS Data Server.

CC BY

Point located observational data are vital for accurate calibration of remotely sensed data



CC BY

Moving Toward FAIR: Best Practice Papers are emerging

Geochron Workshop reports sponsored by EarthChem and EARTHTIME

Walker, Douglas J.; Condon, Daniel; Thompson, William; Renne, Paul; Koppers, Anthony; Hodges, Kip; Reiners, Peter; Stockli, Daniel; Schmitz, Mark; Bowring, Samuel; Gehrels, George

This report compilation describes the outcomes of a series of workshops sponsored by EarthChem and EARTHTIME. The goal of the workshops was to establish community requirements regarding data reporting and the approach employed for getting data into the database are used with the Geochron database. The reports are grouped by technique, and inclu…

ELSEVIER

Quaternary Geochronology
Volume 52, June 2019, Pages 77-87

Guidelines for reporting and archiving ^{210}Pb sediment chronologies to improve fidelity and extend data lifecycle

Colin J. Courtney Mustaphi[a, b, c], Janice Brahney[d], Marco A. Aquino-López[e], Simon Goring[f], Kiersten Orton[d], Alexandra Noronha[g], John Czaplewski[f], Quinn Asena[h], Sarah Paton[i], Johnny Panga Brushworth[j]

Show more

ELSEVIER

Quaternary Geochronology
Volume 39, April 2017, Pages 142-149

Research paper

Data reporting standards for publication of U-series data for geochronology and timescale assessment in the earth sciences

A. Dutton[a], K. Rubin[b], N. McLean[c], J. Bowring[d], E. Bard[e], R.L. Edwards[f], G.M. Henderson[g], M.R. Reid[h], D.A. Richards[i], K.W.W. Sims[j], J.D. Walker[c], Y. Yokoyama[k]

Show more

+ Add to Mendeley   Share   Cite

https://doi.org/10.1016/j.quageo.2017.03.001   Get rights and content

Under a Creative Commons license   open access

Abstract

"Reporting and Interpretation of Geochronologic Data": Invitation for special papers in the *Geological Society of America Bulletin*

The science editors of the *Geological Society of America Bulletin* are encouraging a series of invited special papers to be published during the next 3+ years that highlight current best practices in the reporting and interpretation of geochronologic data and metadata.

The need for such reviews has become acute as the number of manuscripts submitted to *GSA Bulletin* (and other journals) whose conclusions are founded on radioisotopic dating of minerals, rocks, and organic materials has grown explosively during the past decade. Justification includes the following. (1) Many radioisotopic dating methods are in use, including some relatively new methods. (2) Data and metadata sets are commonly large and often contain dozens to thousands of isotopic measurements. (3) With increasing precision of many dating methods, the interpretation of large complex data sets is not always straightforward. It is becoming a challenge for many readers of *GSA Bulletin* to appreciate how a large set of isotopic dates is used to determine a geologically meaningful age for a rock or process. (4) Standardization of many methods has been greatly improved, but a plethora of standard values often requires re-calibration of published data sets.

The papers will be written by specialists, but the aim is that they will be presented in a form that is readily accessible to the broad readership of *GSA Bulletin*. Ideally, each paper would: (1) present a brief overview of the chronometer including the fundamental underlying assumptions and sources of uncertainty; (2) discuss what materials are most suited for analysis and how these are collected in the field and prepared in the laboratory; (3) address the limits on the age range of applicability; (4) review the principal applications; (5) recommend minimum requirements for data and metadata reporting about samples and the isotopic information derived from them with an eye toward archiving the data in a format that will be accessible for the foreseeable future; and (6) provide examples of the challenges in determining ages from complex sets of dates.

We hope to strengthen the role of *GSA Bulletin* in its function of archiving metadata that are of sufficient quality and completeness so that any potential future user has all the information needed to use it—or update it—as standards, decay constants, or production rates, interpretive strategies, etc., evolve.

We intend that these papers will become a valuable resource for authors who generate, organize, and interpret geochronologic data for publication, as well as for editors and associate editors who must evaluate the integrity of geochronologic data sets and the conclusions based upon them.

We seek papers that address these issues for the most widely used geo- and thermochronometers including:

1. U-Pb ID TIMS
2. In situ U-Pb laser ablation ICPMS and SIMS
   a. U-Pb and Lu-Hf laser ablation split stream (LASS)-MC-ICPMS
3. ^{40}Ar/^{39}Ar (and K-Ar)
4. U-series disequilibrium
6. Cosmogenic surface exposure methods
   a. Noble gas methods (^{3}He, ^{21}Ne)
   b. AMS-based methods (^{10}Be, ^{26}Al, ^{36}Cl)
7. Optical luminescence
8. Radiocarbon
9. Fission track
10. Re-Os
11. Other isochron methods (Rb-Sr, Sm-Nd, etc.)

GSA Bulletin

Archive   Content   GSA Publications   Themed Issues   GSA Member Sign In

RESEARCH ARTICLE | JULY 01, 2020

Interpreting and reporting ^{40}Ar/^{39}Ar geochronologic data

Allen J. Schaen; Brian R. Jicha; Kip V. Hodges; Pieter Vermeesch; Mark E. Stelten; Cameron M. Mercer; David Phillips; Tiffany A. Rivera; Fred Jourdan; Erin L. Matchan; Sidney R. Hemming; Leah E. Morgan; Simon P. Kelley; William S. Cassata; Matt T. Heizler; Paulo M. Vasconcelos; Jeff A. Benowitz; Anthony A.P. Koppers; Darren F. Mark; Elizabeth M. Niespolo; Courtney J. Sprain; Willis E. Hames; Klaudia F. Kuiper; Brent D. Turrin; Paul R. Renne; Jake Ross; Sebastien Nomade; Hervé Guillou; Laura E. Webb; Barbara A. Cohen; Andrew T. Calve… Adán Ramirez; J…

GSA Bulletin (20…

https://doi.org/1…

DATA SCIENCE JOURNAL

Reading: Sample Identifiers and Metadata to Support Data Management and Reuse in Multidisciplinary Ec...   Share:

Research Papers

Sample Identifiers and Metadata to Support Data Management and Reuse in Multidisciplinary Ecosystem Sciences

Authors: Joan E. Damerow, Charuleka Varadharajan, Kristin Boye, Eoin L. Brodie, Madison Burrus, K. Dana Chadwick, Robert Crystal-Ornelas, Hesham Elbashandy, Ricardo J. Eloy Alves, Kim S. Ely, Amy E. Goldman, Ted Haberman, Valerie Hendrix, Zarine Kakalia, Kenneth M. Kemner, Annie B. Kersting, Nancy Merino, Fianna O'Brien, Zach Perzan, Emily Robles, Patrick Sorensen, James C. Stegen, Ramona L. Walls, Pamela Weisenhorn, Mavrik Zavarin, Deborah Agarwal

Abstract

Paleoceanography and Paleoclimatology

Feature Article   Free Access

PaCTS 1.0: A Crowdsourced Reporting Standard for Paleoclimate Data

D. Khider, J. Emile-Geay, N. P. McKay, Y. Gil, D. Garijo, V. Ratnakar, M. Alonso-Garcia, S. Bertrand, O. Bothe, P. Brewer, A. Bunn, M. Chevalier, L. Comas-Bru, A. Csank, E. Dassié, ... See all authors

First published: 03 September 2019 | https://doi.org/10.1029/2019PA003632 | Citations: 9

This article was corrected on 21 FEB 2020. See the end of the full text for details.

SECTIONS   PDF   TOOLS   SHARE

Abstract

The progress of science is tied to the standardization of measurements, instruments, and

---

IAG   GEOSTANDARDS and GEOANALYTICAL RESEARCH

Original Article   Open Access

Community-Derived Standards for LA-ICP-MS U-(Th-)Pb Geochronology – Uncertainty Propagation, Age Interpretation and Data Reporting

Matthew S. A. Horstwood, Jan Košler, George Gehrels, Simon E. Jackson, Noah M. McLean, Chad Paton, Norman J. Pearson, Keith Sircombe, Paul Sylvester, Pieter Vermeesch, ... See all authors

First published: 01 February 2016 | https://doi.org/10.1111/j.1751-908X.2016.00379.x | Citations: 279

SECTIONS   PDF   TOOLS   SHARE

Abstract   EN FR

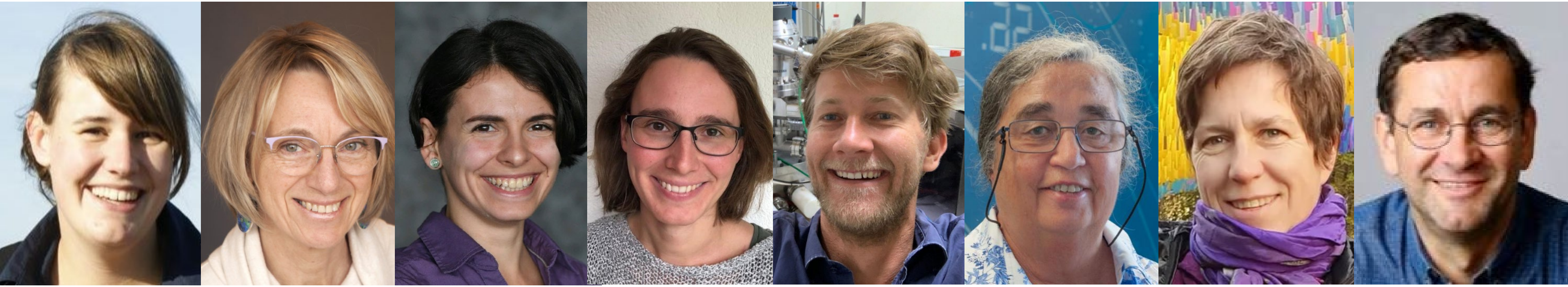
The LA-ICP-MS U-(Th-)Pb geochronology international community has defined new standards for the determination of U-(Th-)Pb ages. A new workflow defines the appropriate propagation of uncertainties for these data, identifying random and systematic components. Only data with uncertainties relating to random error should be used in weighted mean calculations of population ages; uncertainty components for systematic errors are propagated after this stage, preventing their erroneous reduction. Following this improved uncertainty propagation protocol, data can be compared at

cc BY

NCI AUSTRALIA

OneGeochemistry is an international collaboration

ONEGEOCHEMISTRY



EPOS EUROPEAN PLATE OBSERVING SYSTEM

EARTHCHEM

ASTROMAT

DIGIS GEOROC

AusGeochem

ARDC Australian Research Data Commons

GFZ DATA SERVICES GEOSCIENCES DATA PUBLISHER

NFDI4Earth

CC BY

But no single authority: 5 Unions and ~40 Societies

NCI AUSTRALIA

ONEGEOCHEMISTRY

Societies/Associations

geochemical society

EUROPEAN ASSOCIATION OF GEOCHEMISTRY

SEGH Society for Environmental Geochemistry and Health

IUGS Commission on "Global Geochemical Baselines"

Mapping the Geochemistry of the Earth's Land Surface

IAGC International Association of GeoChemistry

IAVCEI INTERNATIONAL ASSOCIATION OF VOLCANOLOGY AND CHEMISTRY OF THE EARTH'S INTERIOR

IAG INTERNATIONAL ASSOCIATION OF GEOANALYSTS

THE ASSOCIATION OF APPLIED GEOCHEMISTS 1970

The Meteoritical Society

CGI Commission for the Management and Application of Geoscience Information

Deep-time Digital Earth IUGS Big Science Program

Science Unions

International Union of Geodesy and Geophysics

Union Geodesique et Geophysique Internationale

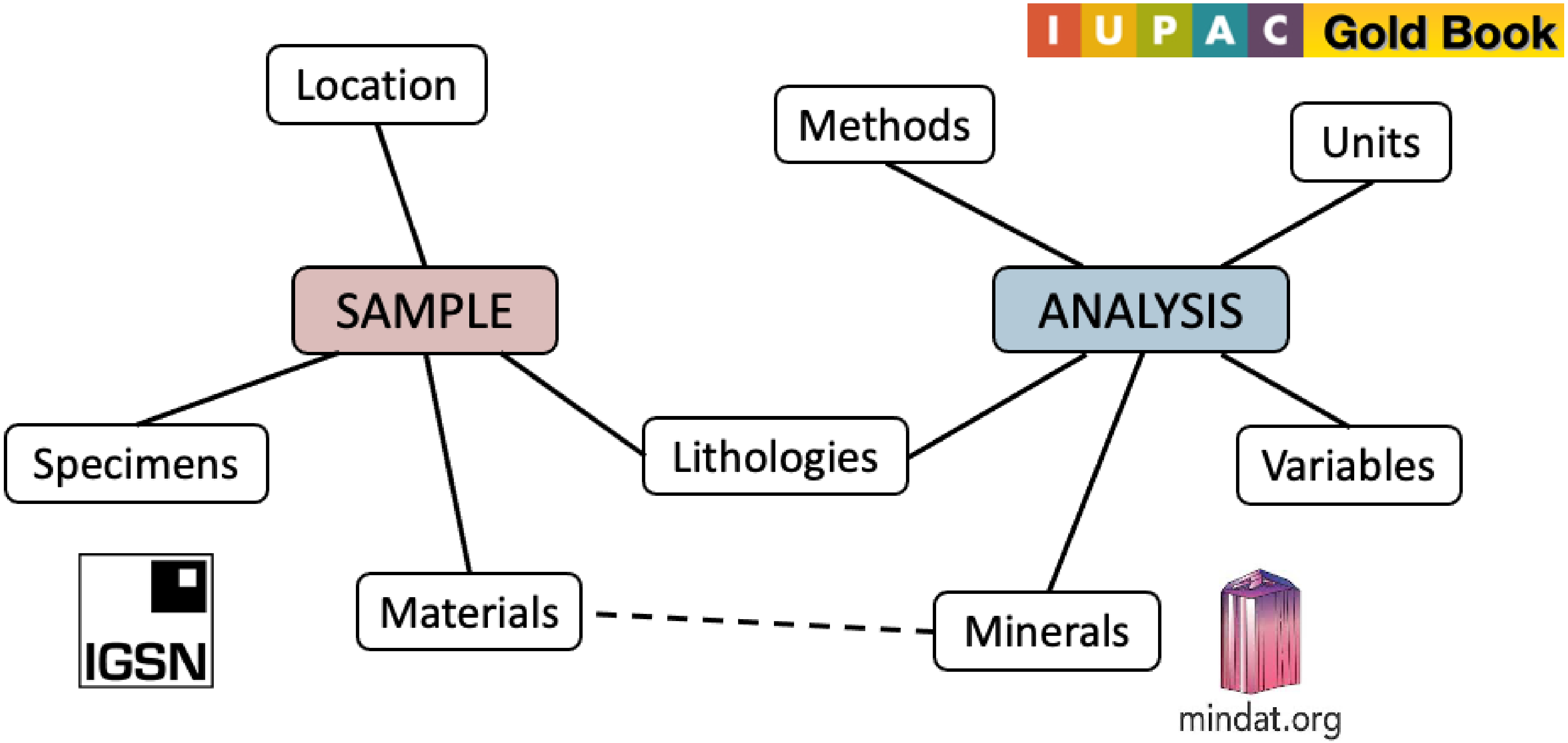
IUGS International Union of Geological Sciences

International Union of Crystallography (IUCr)

IUPAC INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY

International Union of Soil Sciences

Here comes O&M again



CC BY

The IUPAC Gold Book – A digital aggregation of IUPAC Terminologies

IUPAC Gold Book

Searchable → Search (three chars min) Resources

https://goldbook.iupac.org

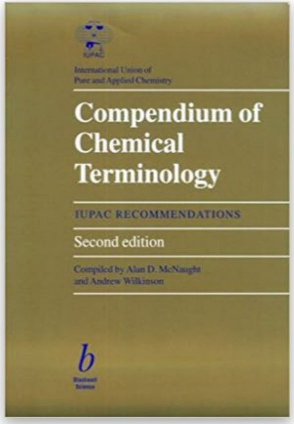
Compendium of Chemical Terminology

Alphabetical Index

A	B	C	D	E	F
G	H	I	J	K	L
M	N	O	P	Q	R
S	T	U	V	W	XYZ

Additional Indexes

- Physical Constants
- Units of Measure
- Physical Quantities
- SI Prefixes
- Ring Index
- General Formulae
- Exact Formulae
- Source Documents



Welcome to the new interactive ... Chemical Terminology, informa... these pages you will find a ... publication. Start by:

- browsing the alphabetical ...
- using one of the many the...
- using the search in the na...

To learn more about this new in... (Updated July 1st, 2019).

Expanded Search! New Download Formats!

absorbance, A

DOI → https://doi.org/10.1351/goldbook.A00028

Logarithm of the ratio of incident to ... adiant power through a sample (excluding the effe... ...ing on the base of the logarithm a decadic and Napierian absorba... , A_{10}, A_e. This quantity is sometimes called extinction, althoug... ...er called attenuance, is reserved for the quantity which takes into account ... effects of luminescence and scattering as well.

Provenance

Source:
Green Book, 2nd ed., p. 32 [Terms] [Book]

The Gold Book API Alpha API v1.0 (5/31/19) ← API

While we expect a lot of humans to stop by the Gold Book, its about time that the vocabulary be friendly towards computers and have set up an application programming interface (API) so they may download a bunch of stuff. Here is the overview of the API and we are working on additional documentation. (click the headers below to toggle whats visible.

Terms

Endpoint/Notes	Example(s)
`/terms/index/[scope]/[format]/[download]` ***List of terms in the Gold Book*** [scope]: (all), A-W, XYZ (returns to referring page if no data) [format]: (html), xml, json (rest are ignored) [download]: (""), download (rest are ignored)	/terms/index/all (just "terms" works too) /terms/index/C/xml /terms/index/XYZ/json/download
`/terms/view/[identifier]/[format]/[download]` ***A term from the Gold Book*** [identifier]: term code or id [format]: (html), xml, json (rest are ignored) [download]: (""), download (rest are ignored)	/terms/view/A00001 /terms/view/P04409/json /terms/view/ZT07132/xml/download

Sources (Click to show)

IUPAC are currently developing the processes and practices needed to ensure definitions are born digital as part of Digital Chemistry

Ack: S. Chalk

Welcome | Themes & Oral/Poster | Programme | Plenary | Field Trips Exkursionen | Registration | SOC | Venue & Hotel | Sponsoring | Join DGGV | Awardees | Contact | Press | Schulen & Lehrende

GeoBerlin 2023 –

Geosciences Beyond Boundaries – Research, Society, Future

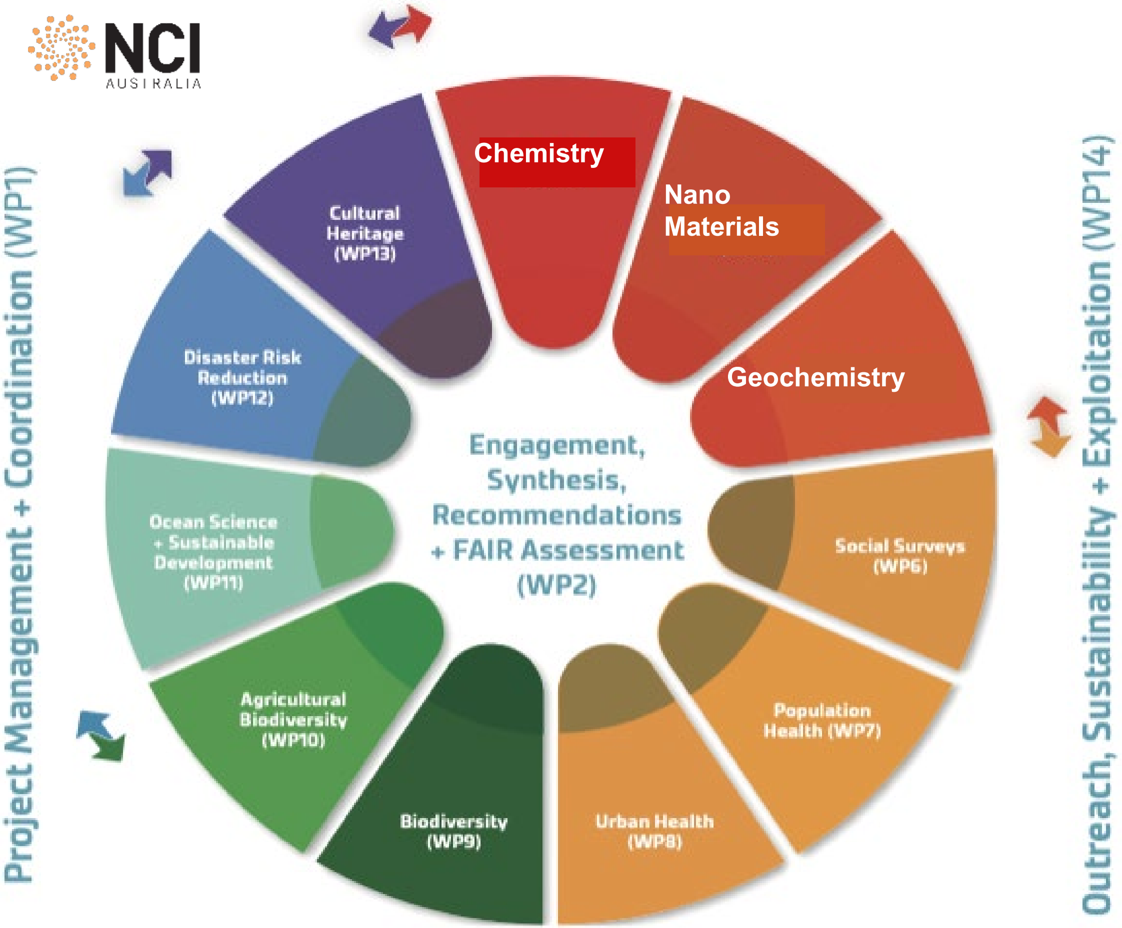
150th PGLA (BGR) Anniversary and 175th DGGV Anniversary

Berlin | 3 – 7 September 2023

CC BY

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



The 11 WorldFAIR case studies are:

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

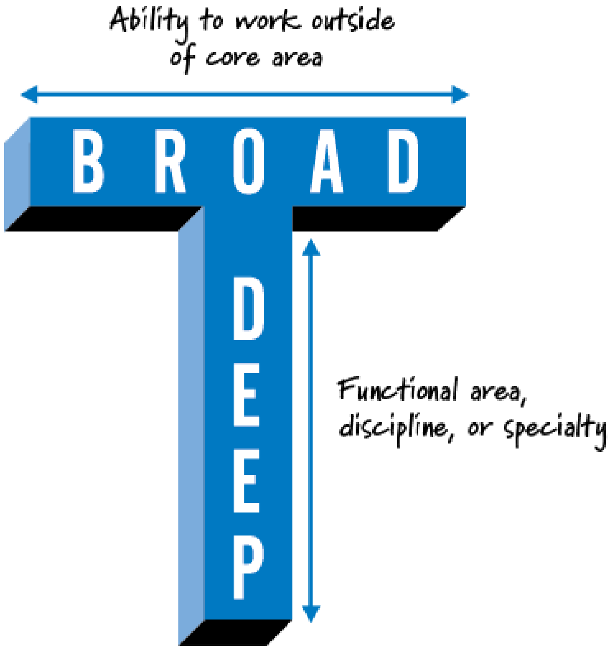
Theme 1: 1–3
Theme 2: 4–5
Theme 3: 6–8
Theme 4: 8–10
Theme 5: 11

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

CC BY

The T-Bone effect:

1. We need to be aware of tensions between broad and deep
2. The tension between developing user-friendly processes that enable integration of datasets across disparate disciplinary areas in ways that do not impact on the quality and integrity of deeper disciplinary research



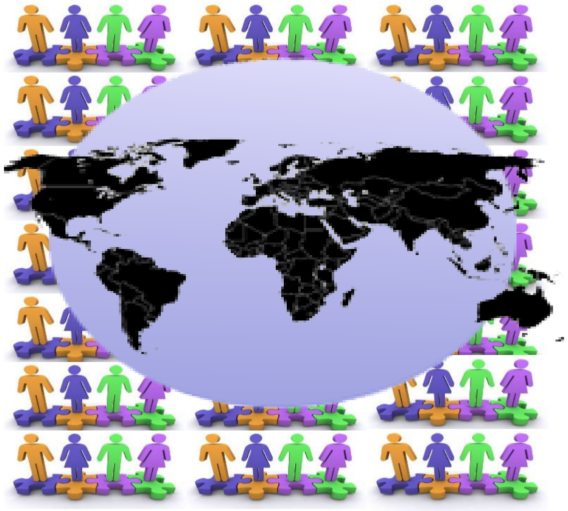
Copyright © 2012, Kenneth S. Rubin and Innolution, LLC. All Rights Reserved.

https://innolution.com/resources/glossary/t-shaped-skills

CC BY

The size of your community you share data with counts

- The growing need to share data, information and services across multiple disciplines and organisations
- This requires standards for machine readability
- Increasingly digital data collections need to be reused and repurposed by much broader communities
- The size of the community that you interoperate with is as large as the size of the community that uses your standard
- Where possible we need to converge on international standards



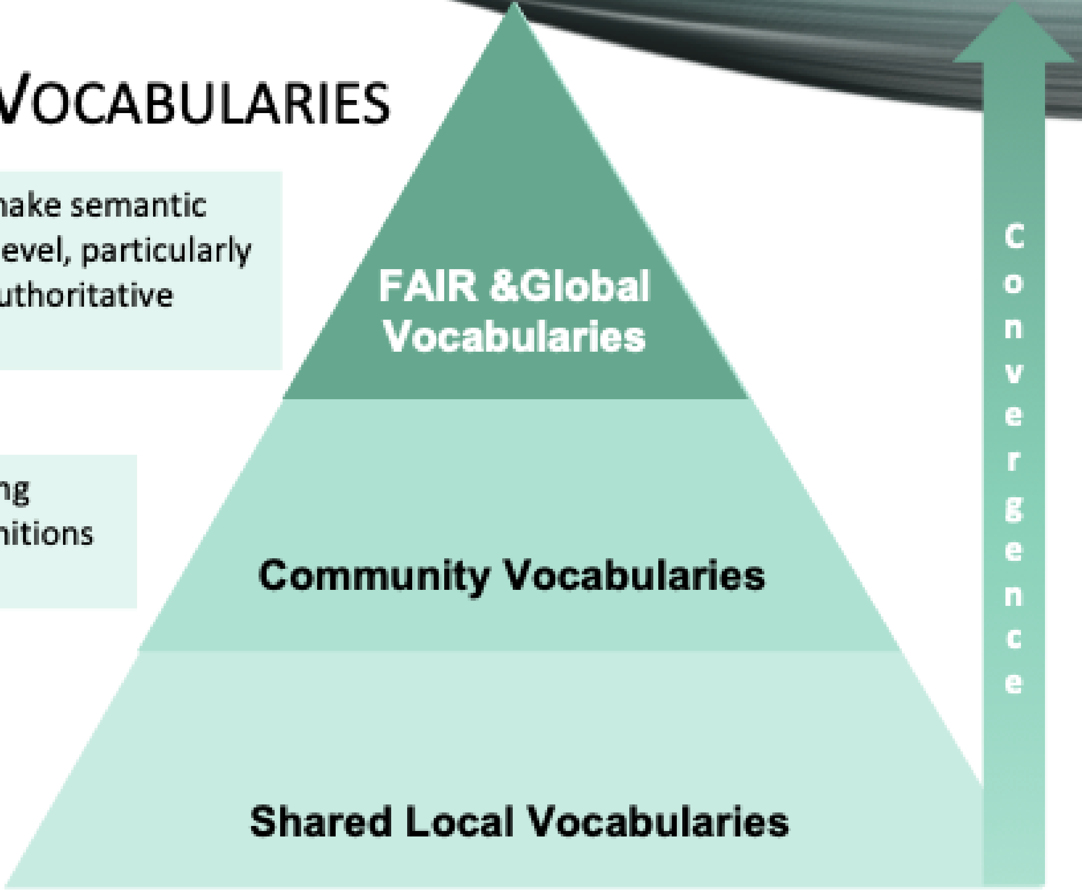
CC BY

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).

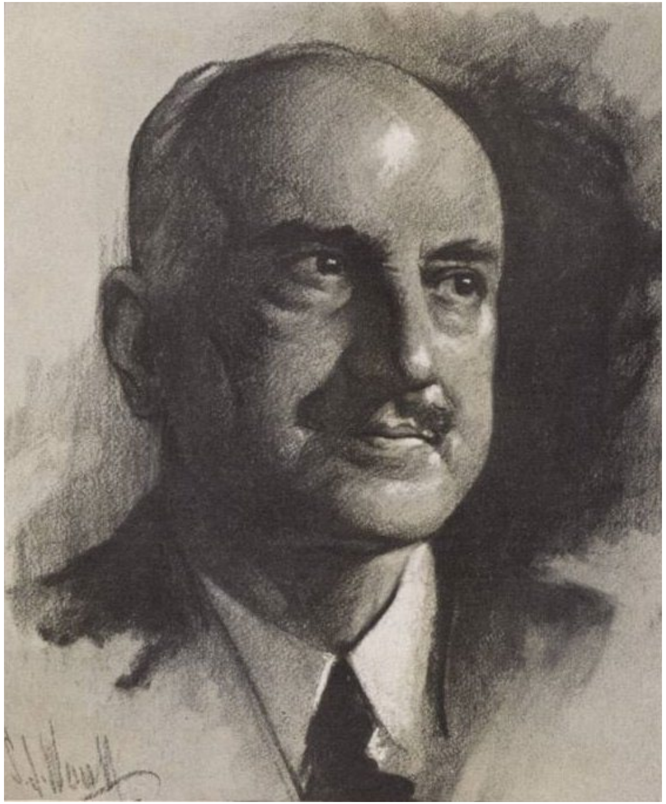
Groups with similar topics to begin harmonizing across multiple locally-derived concepts/ definitions and publish these as community resources;

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



Credit: Kerstin Lehnert

Let's look at history to see how others dealt with change



Those who cannot remember the past are condemned to repeat it.

George Santayana

CC BY

Codd's 1970 Rules of Relational Database Management Systems (RDBMS)

1. Future users of large data banks need to be protected from having to know how the data is organised in the machine (the internal representation).

2. Activities of users at terminals … should remain unaffected when the internal representation of data is changed.

3. Changes in data representation will often be needed as a result of changes in query, update….and natural growth in the types of stored information.

Codd, E. F, 1970. A Relational Model of Data for Large Shared Data Banks. Communications of the ACM Volume 13 Issue 6, June 1970 pp 377–387 https://doi.org/10.1145/362384.362685

A Relational Model of Data for Large Shared Data Banks

E. F. Codd
IBM Research Laboratory, San Jose, California

Future users of large data banks must be protected from having to know how the data is organized in the machine (the internal representation). A prompting service which supplies such information is not a satisfactory solution. Activities of users at terminals and most application programs should remain unaffected when the internal representation of data is changed and even when some aspects of the external representation are changed. Changes in data representation will often be needed as a result of changes in query, update, and report traffic and natural growth in the types of stored information.

Existing noninferential, formatted data systems provide users with tree-structured files or slightly more general network models of the data. In Section 1, inadequacies of these models are discussed. A model based on *n*-ary relations, a normal form for data base relations, and the concept of a universal data sublanguage are introduced. In Section 2, certain operations on relations (other than logical inference) are discussed and applied to the problems of redundancy and consistency in the user's model.

KEY WORDS AND PHRASES: data bank, data base, data structure, data organization, hierarchies of data, networks of data, relations, derivability, redundancy, consistency, composition, join, retrieval language, predicate calculus, security, data integrity
CR CATEGORIES: 3.70, 3.73, 3.75, 4.20, 4.22, 4.29

1. Relational Model and Normal Form

1.1. Introduction

This paper is concerned with the application of elementary relation theory to systems which provide shared access to large banks of formatted data. Except for a paper by Childs [1], the principal application of relations to data systems has been to deductive question-answering systems. Levein and Maron [2] provide numerous references to work in this area.

In contrast, the problems treated here are those of *data independence*—the independence of application programs and terminal activities from growth in data types and changes in data representation—and certain kinds of *data*

CC BY

OneGeoscience = OneGeology + OneGeochemistry + OneGeophysics + ????

Photo by Nathan Dumlao on Unsplash

1. We no longer have the gift of time.
2. When you publish a dataset with a paper, ask yourself how many other people are using your standards, formats and vocabularies - that will tell you the size of the community that can read and understand your data.
3. Although there is a the focus on distribution of images and derived products, we still need to find all the rawer forms of critical geoscience datasets and make less processed forms of these data more FAIR compliant and able to be aggregated into seamless national/global high-resolution datasets.
4. Ensure that whatever we do, it is always scalable to the future and can maximise benefits from new compute, data and software technologies as they come on line. Things will always change.

CC BY

NCI AUSTRALIA

NCI Contacts

General enquiries: +61 2 6125 9800

Support: help@nci.org.au

Email: nigel.rees@anu.edu.au

Address

NCI, ANU Building 143
143 Ward Road
The Australian National University
Canberra ACT 2601

License

CC BY SA