Modernising Datasets: Revitalising the South Australia Magnetotellurics Collection for Current and Future Data Intensive Analysis

Hannes Hollmann[1], Nigel Rees[1], Jo Croucher[1], Bruce Goleby[2], Ben Evans[1], Rebecca Farrington[3], Yiling Liu[1], Lesley Wyborn[1]

[1]National Computational Infrastructure, [2]OPM Consulting Pty Ltd, [3]AuScope Ltd

29 October 2024

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

Since the 1950s, large volumes of Australian geophysics data have been acquired by universities, industry, and government agencies. However, in many geophysical disciplines, only higher-level data products are accessible online, while valuable original time-series data and/or metadata remains largely inaccessible. This data is often stored offline on various types of media and as institutional knowledge of these datasets fades, the risk of permanent loss increases. The research impact is exacerbated by the interest in machine learning applications that require comprehensive collections of datasets adhering to the FAIR principles. Open Science mandates increasingly require access to all artefacts of the scientific workflow, not just the processed products.

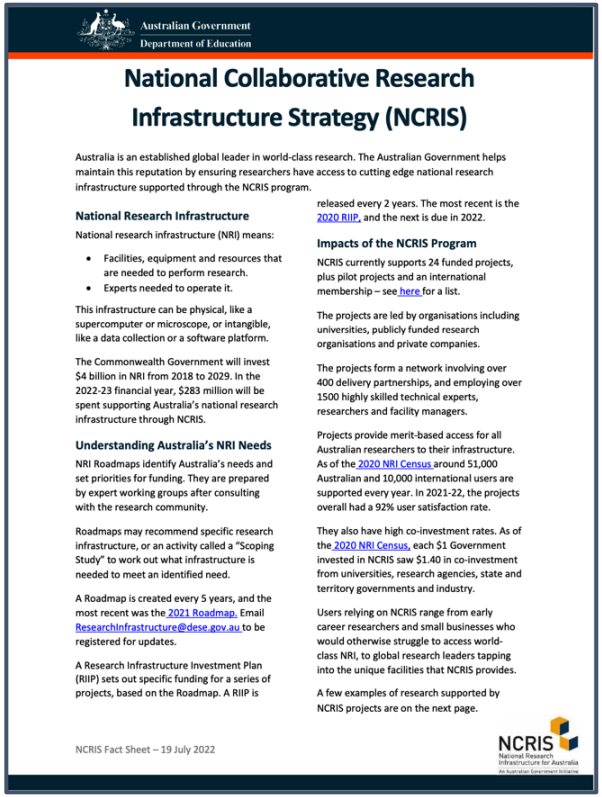
Ensuring that raw time-series data is publicly accessible preserves scientific legacy and enhances future research capabilities by providing a rich foundation for advanced analytical techniques. Data rescue efforts to locate, preserve, and digitise these valuable datasets are crucial for preventing their permanent loss and ensuring their availability for future advancements in analysis techniques.

To address these issues, the National High-resolution Geophysics Reference Collections for 2030 Computation Project has focused on modernising the South Australian AusLAMP Magnetotellurics Collection by consolidating MT datasets and metadata resources into a single archive, available as part of AuScope's data collection at NCI for either in-situ HPC/cloud access or file downloads. Through collaboration with current and former survey personnel, the project has increased the preservation of survey sites from 50% to 96% and added value through data curation and enhanced metadata, thereby fostering a more open and transparent scientific environment.

AuScope/NCI are part of NCRIS: the Australian Research Infrastructure Strategy

- It seeks to ensure Australia is an established global leader in world-class research.
- The Australian Government helps maintain this reputation by **ensuring researchers have access to cutting edge national research infrastructure** supported through the NCRIS program.
- There are currently 26 NCRIS facilities, including:
  - AuScope (Solid Earth Sciences)
  - IMOS (Integrated Marine Observing System)
  - TERN (Terrestrial Ecosystem Research Network)
  - NCI (National Computational Infrastructure)
  - ARDC (Australian Research Data Commons)

https://www.education.gov.au/ncris



Australian Government
Department of Education

National Collaborative Research Infrastructure Strategy (NCRIS)

Australia is an established global leader in world-class research. The Australian Government helps maintain this reputation by ensuring researchers have access to cutting edge national research infrastructure supported through the NCRIS program.

National Research Infrastructure

National research infrastructure (NRI) means:

- Facilities, equipment and resources that are needed to perform research.
- Experts needed to operate it.

This infrastructure can be physical, like a supercomputer or microscope, or intangible, like a data collection or a software platform.

The Commonwealth Government will invest $4 billion in NRI from 2018 to 2029. In the 2022-23 financial year, $283 million will be spent supporting Australia's national research infrastructure through NCRIS.

Understanding Australia's NRI Needs

NRI Roadmaps identify Australia's needs and set priorities for funding. They are prepared by expert working groups after consulting with the research community.

Roadmaps may recommend specific research infrastructure, or an activity called a "Scoping Study" to work out what infrastructure is needed to meet an identified need.

A Roadmap is created every 5 years, and the most recent was the 2021 Roadmap. Email ResearchInfrastructure@dese.gov.au to be registered for updates.

A Research Infrastructure Investment Plan (RIIP) sets out specific funding for a series of projects, based on the Roadmap. A RIIP is released every 2 years. The most recent is the 2020 RIIP, and the next is due in 2022.

Impacts of the NCRIS Program

NCRIS currently supports 24 funded projects, plus pilot projects and an international membership – see here for a list.

The projects are led by organisations including universities, publicly funded research organisations and private companies.

The projects form a network involving over 400 delivery partnerships, and employing over 1500 highly skilled technical experts, researchers and facility managers.

Projects provide merit-based access for all Australian researchers to their infrastructure. As of the 2020 NRI Census around 51,000 Australian and 10,000 international users are supported every year. In 2021-22, the projects overall had a 92% user satisfaction rate.

They also have high co-investment rates. As of the 2020 NRI Census, each $1 Government invested in NCRIS saw $1.40 in co-investment from universities, research agencies, state and territory governments and industry.

Users relying on NCRIS range from early career researchers and small businesses who would otherwise struggle to access world-class NRI, to global research leaders tapping into the unique facilities that NCRIS provides.

A few examples of research supported by NCRIS projects are on the next page.

NCRIS Fact Sheet – 19 July 2022

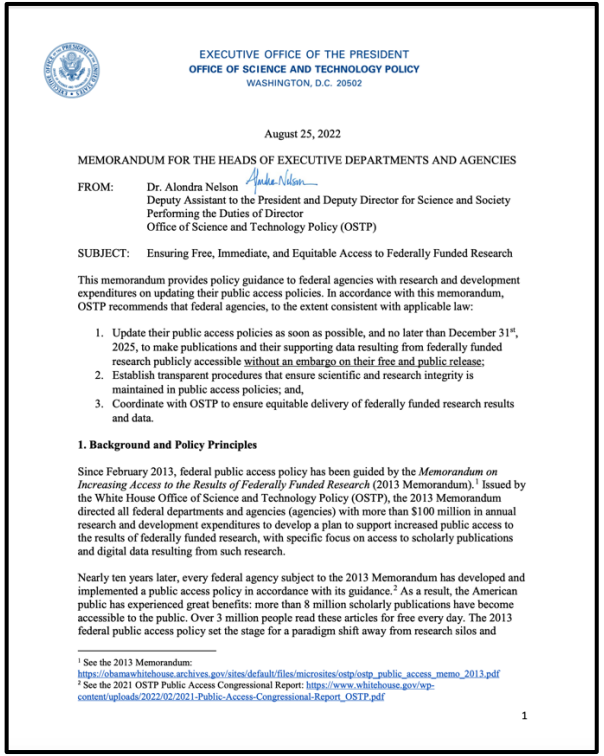
NCRIS National Research Infrastructure for Australia An Australian Government Initiative

AuScope/NCI Vision is built on the Open Science Paradigm

Open science aims to ensure transparent and accessible knowledge, for the scientific community and the general public alike, to the whole scientific workflow, rather than just the final research product (Gentlemann, 2023)

The US Government Office of Science and Technology (OSTP) policy has mandated that all federally funded agencies with research and development expenditures update their public access policies (Including NSF) to:

1. Update their public access policies as soon as possible, and no later than Dec 31, 2025, to make publications and their supporting data resulting from federally funded research publicly accessible without an embargo on their free and public release;
2. Establish transparent procedures that ensure scientific and research integrity is maintained in public access policies; and,
3. Coordinate with OSTP to ensure equitable delivery of federally funded research results and data.



EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF SCIENCE AND TECHNOLOGY POLICY
WASHINGTON, D.C. 20502

August 25, 2022

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: Dr. Alondra Nelson
Deputy Assistant to the President and Deputy Director for Science and Society
Performing the Duties of Director
Office of Science and Technology Policy (OSTP)

SUBJECT: Ensuring Free, Immediate, and Equitable Access to Federally Funded Research

This memorandum provides policy guidance to federal agencies with research and development expenditures on updating their public access policies. In accordance with this memorandum, OSTP recommends that federal agencies, to the extent consistent with applicable law:

1. Update their public access policies as soon as possible, and no later than December 31st, 2025, to make publications and their supporting data resulting from federally funded research publicly accessible without an embargo on their free and public release;
2. Establish transparent procedures that ensure scientific and research integrity is maintained in public access policies; and,
3. Coordinate with OSTP to ensure equitable delivery of federally funded research results and data.

1. Background and Policy Principles

Since February 2013, federal public access policy has been guided by the *Memorandum on Increasing Access to the Results of Federally Funded Research* (2013 Memorandum).[1] Issued by the White House Office of Science and Technology Policy (OSTP), the 2013 Memorandum directed all federal departments and agencies (agencies) with more than $100 million in annual research and development expenditures to develop a plan to support increased public access to the results of federally funded research, with specific focus on access to scholarly publications and digital data resulting from such research.

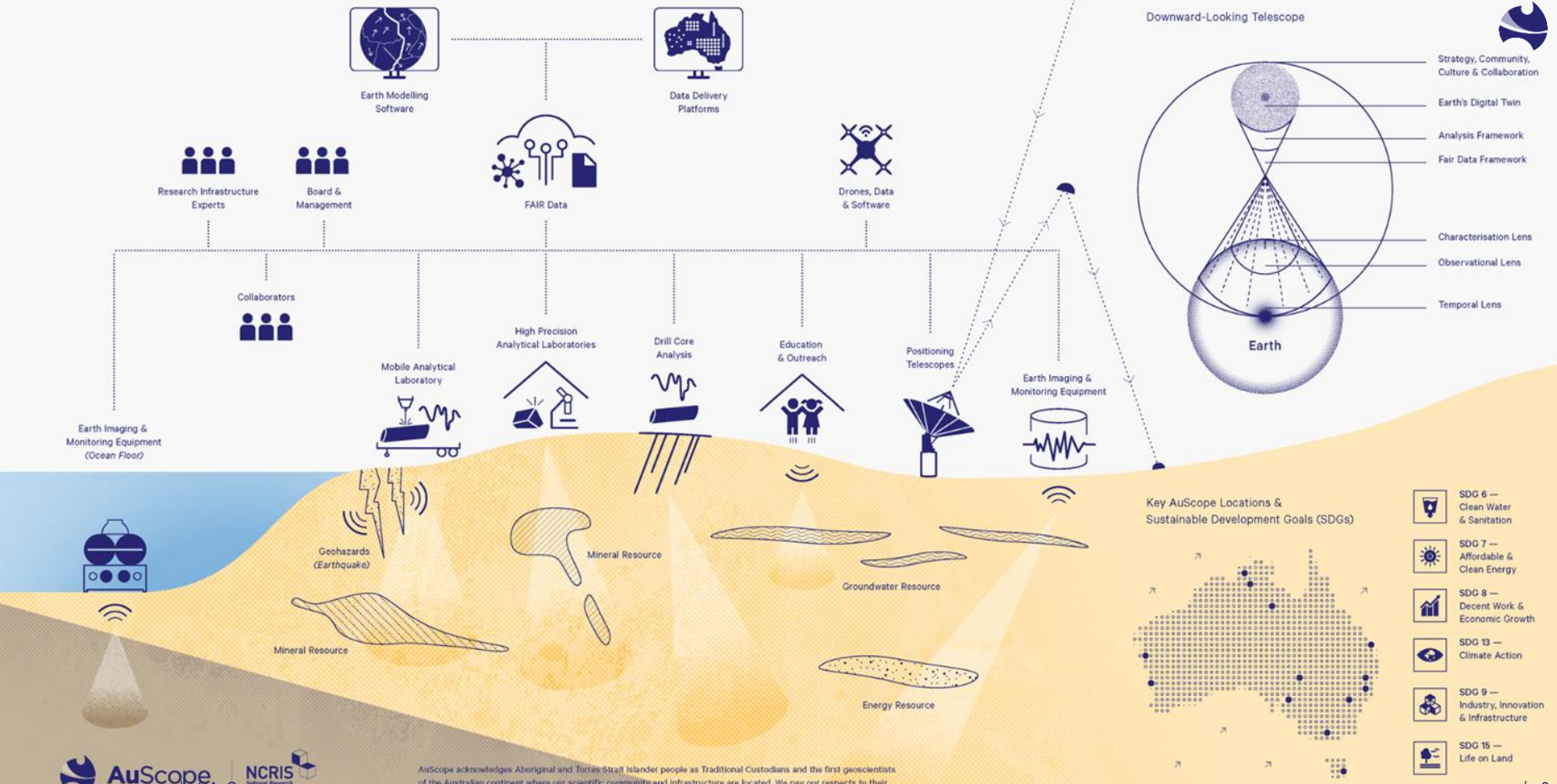
Nearly ten years later, every federal agency subject to the 2013 Memorandum has developed and implemented a public access policy in accordance with its guidance.[2] As a result, the American public has experienced great benefits: more than 8 million scholarly publications have become accessible to the public. Over 3 million people read these articles for free every day. The 2013 federal public access policy set the stage for a paradigm shift away from research silos and

[1] See the 2013 Memorandum: https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf
[2] See the 2021 OSTP Public Access Congressional Report: https://www.whitehouse.gov/wp-content/uploads/2022/02/2021-Public-Access-Congressional-Report_OSTP.pdf

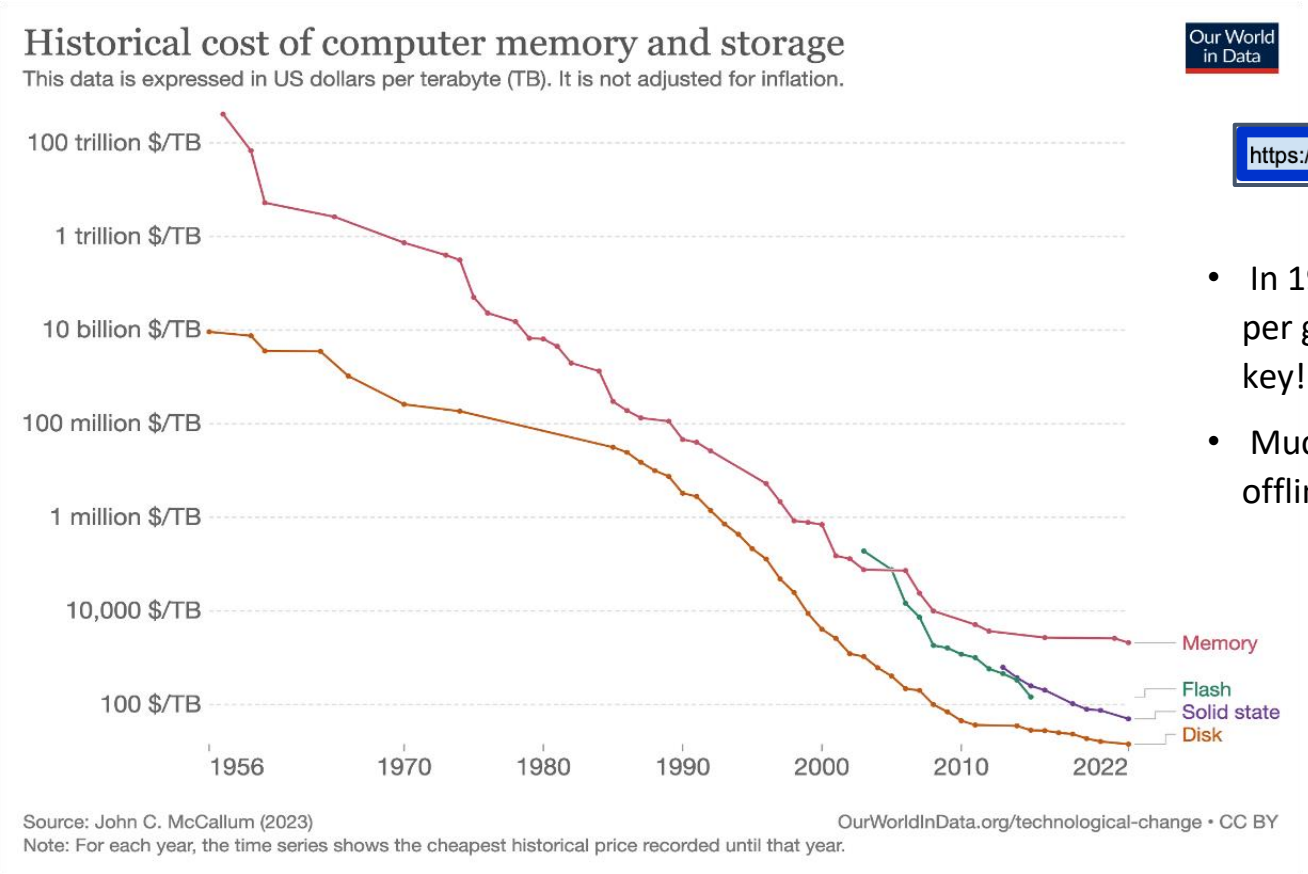
1

https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-access-Memo.pdf&ved=2ahUKEwjNpo-Rt5GJAxXXUGwGHVtzD9MQFnoECBkQAQ&usg=AOvVaw2g2QnCn_f_RkqyCexOOEVI

The Downward Looking Telescope Concept

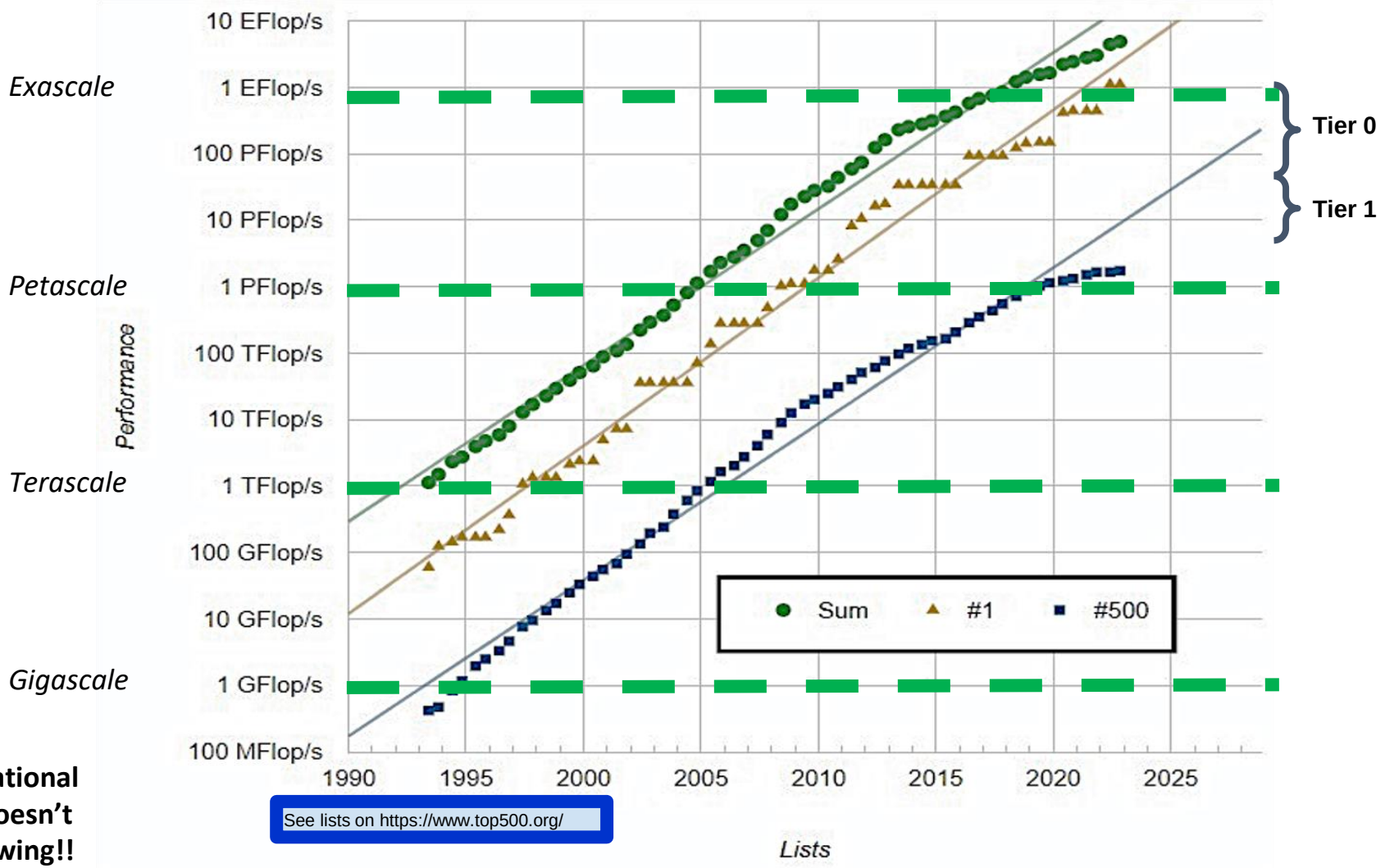


Up to ~2010 storage and memory were expensive



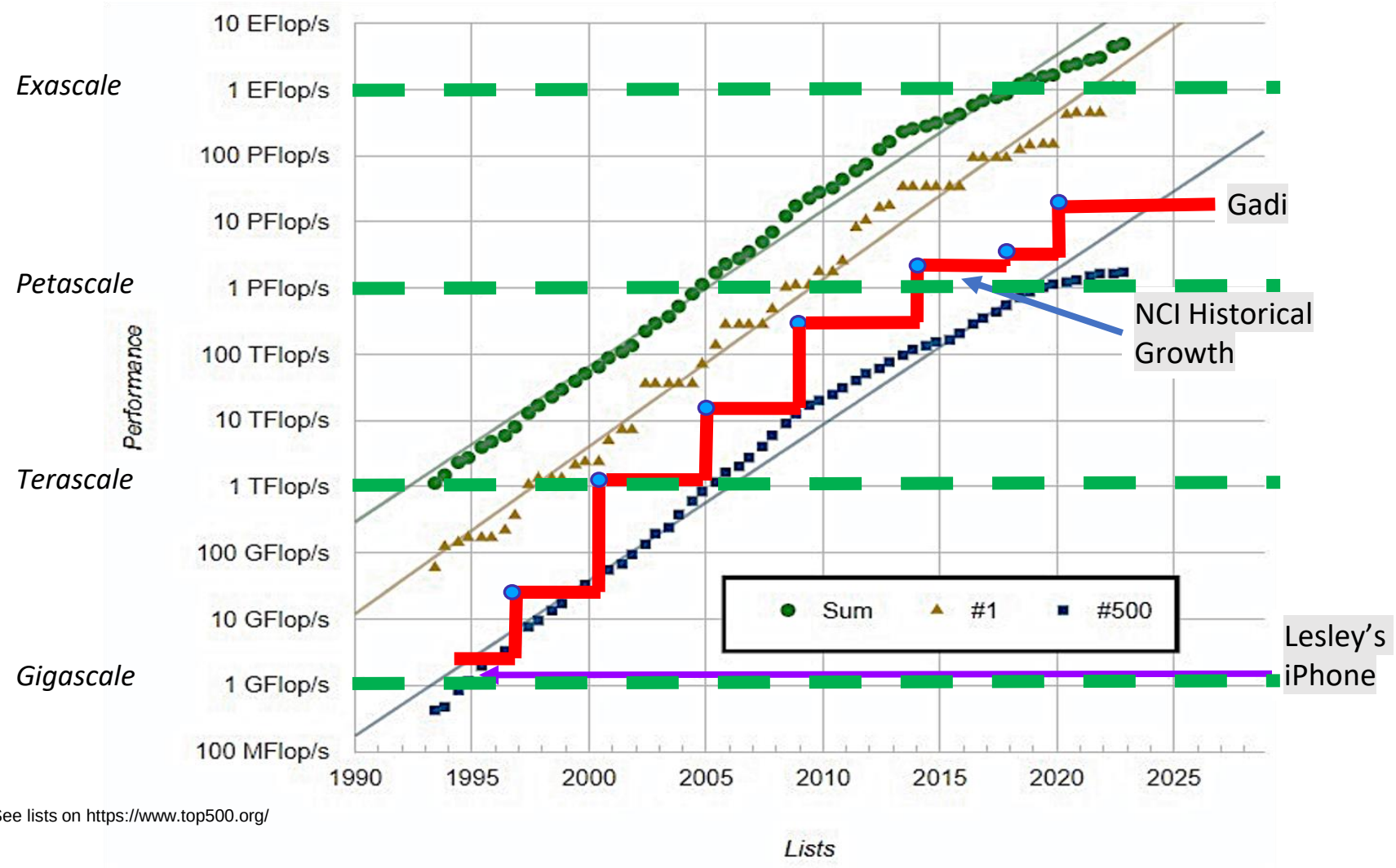
https://ourworldindata.org/technological-change

- In 1980 storage was ~$1M per gigabyte: efficiency was key!
- Much data was stored in offline tapes/hard drives



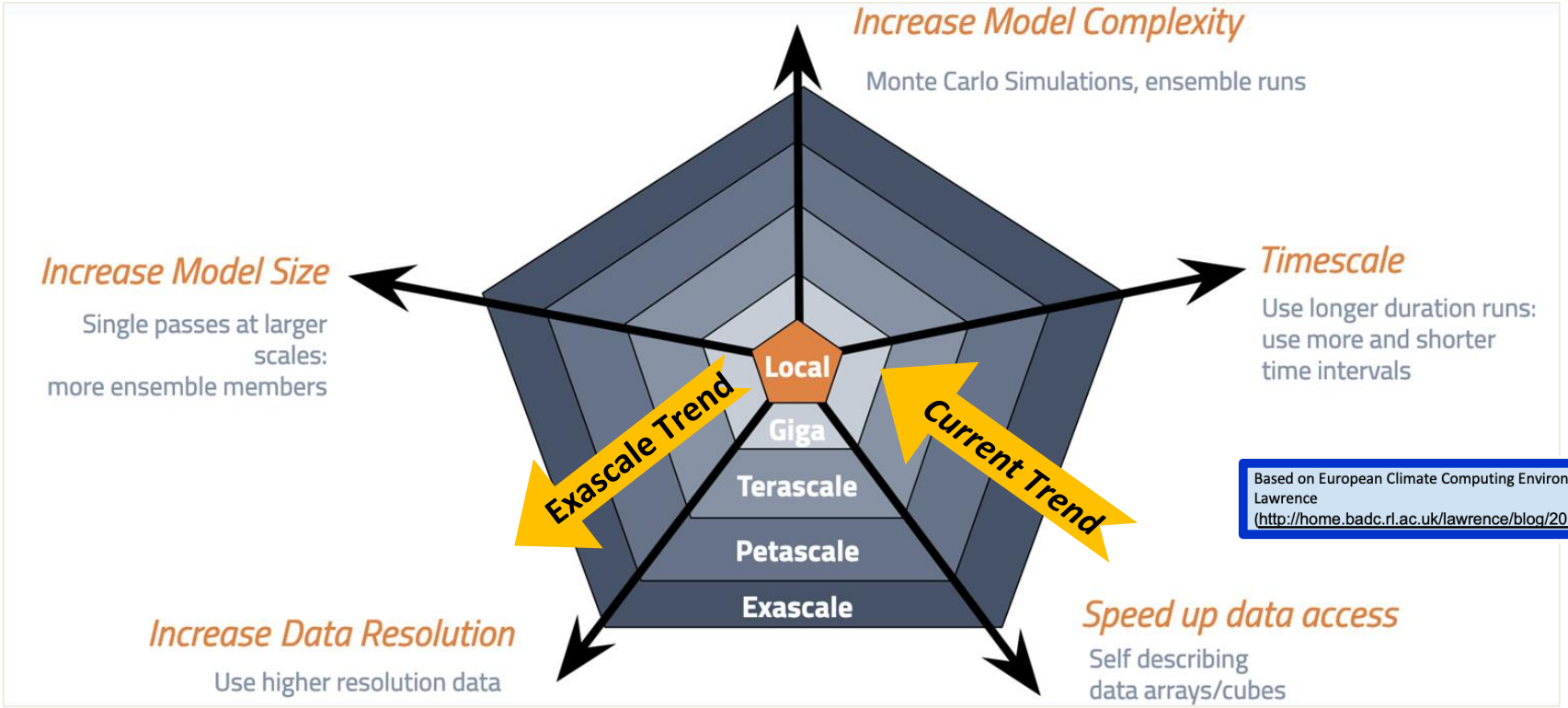
Computational Power doesn't stop growing!!

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



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

What Does More Computational Power Mean to our Science?



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

Exascale computing is already a reality in the USA: Frontier

OAK RIDGE National Laboratory

Scienc

News

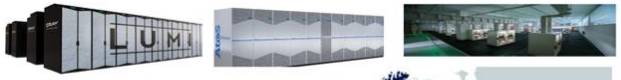
Frontier supercomputer debuts as world's fastest, breaking exascale barrier

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

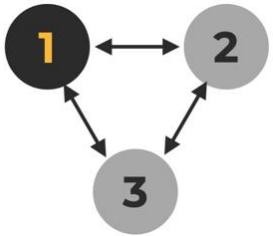
https://www.ornl.gov/news/ornl-celebrates-launch-frontier-worlds-fastest-supercomputer

European Preparation for Exascale

The European HPC ecosystem (pillar 1)



Infrastructure (EuroHPC) Applications (CoEs, NCCs)



Technology (EPI, FET, etc)

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

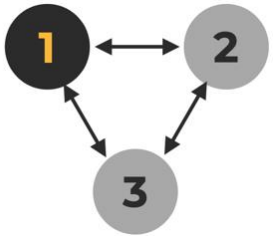
tier-1 (petascale)			T500
MeluXina	Luxemburg	15 PFlop/s	52
Karolina	Txequia	7 PFlop/s	85
Discover	Bulgaria	5 PFlop/s	123
Vega	Slovenia	4 PFlop/s	140
Deucalio	Portugal	10 PFlop/s	-

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

European Preparation for Exascale

The European HPC ecosystem (pillar 1)

Infrastructure (EuroHPC) Applications (CoEs, NCCs)



Technology (EPI, FET, etc)

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

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

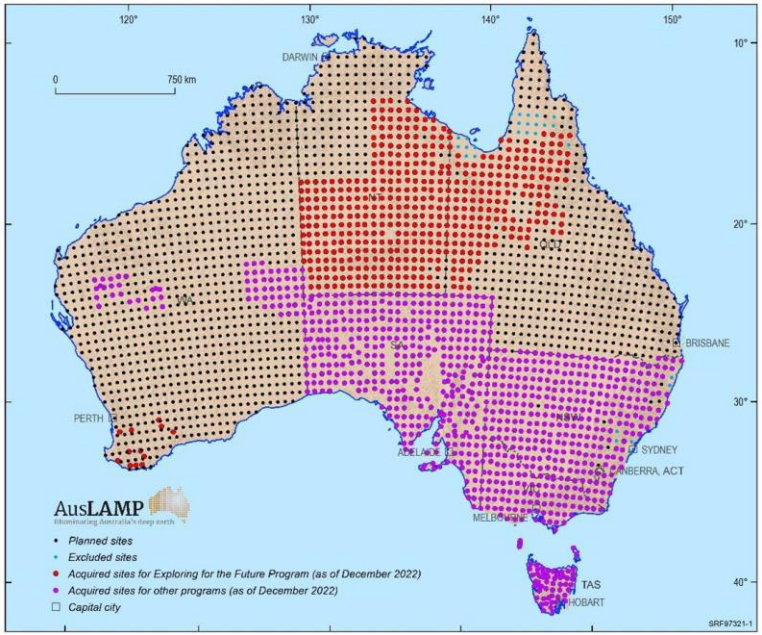
Slide from Arnau Folch, Centre of Excellence in Exascale computing for Solid Earth (ChEESE): https://cheese-coe.eu/

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 aimed 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/)

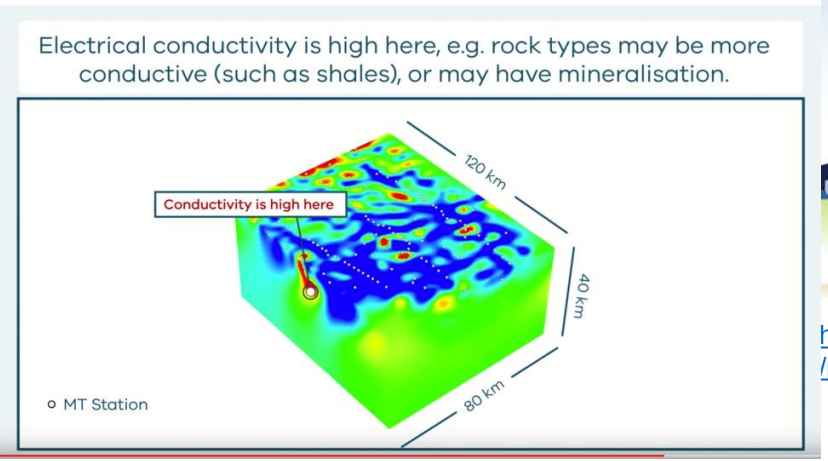
National Scale Analysis: The Australian Lithospheric Architecture Magnetotelluric Project (AusLAMP)

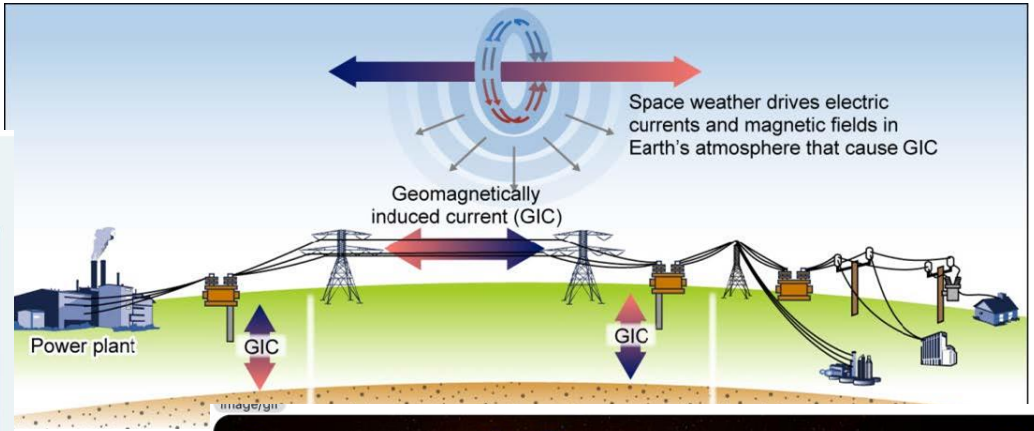


- A collaborative project between Geoscience Australia, the state and Northern Territory geological surveys, AuScope, universities and other research organisations.
- Aims to acquire long-period MT data at approximately 3000 sites across Australia.
- AuScope co-funded the Musgraves, South Australia and Tasmanian data collection: these were our targets
- Preparing for 2030 by providing access to the raw high-resolution observational data

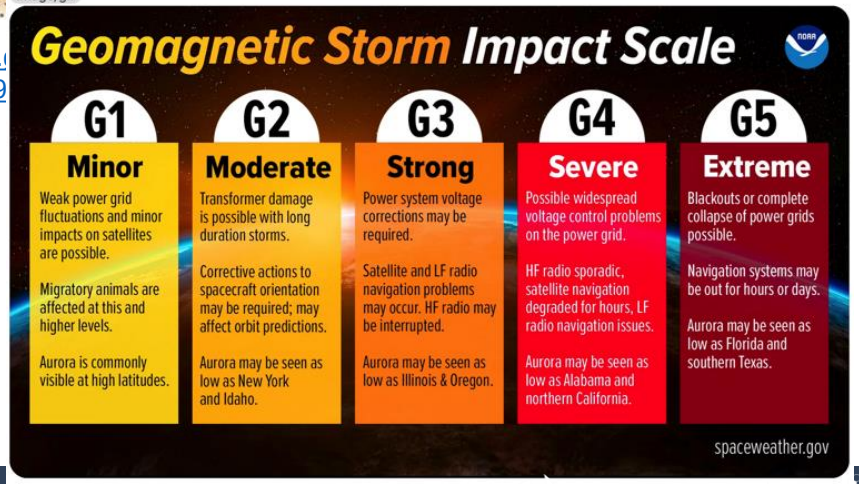
https://www.ga.gov.au/about/projects/resources/auslamp

Why is it important?: same data for different purposes





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



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

The basics of MT time series acquisition and processing

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

Processing Levels	Name	Typical Volumes	Description
Packed Raw Data	Raw Time Series	GBs to TBs	Telemetry data streamed from site loggers
Level 0	Edited Time Series	GBs to TBs	Time ordered instrument recorded data (e.g., raw voltages, counts) at full resolution
Level 1	Transformed Time Series	GBs to TBs	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	MBs	Geophysical parameters (e.g., impedance tensors) derived from frequency domain time series processing of Level 1 data
Level 3	Derived modelling inputs and outputs	MBs	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: 10.1080/22020586.2019.12073013

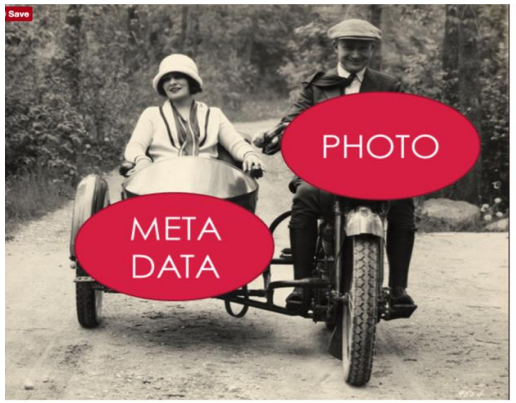
HPC reduced processing times significantly: more targeted research

Testing Parallel I/O via NCI-geophys environment on the AusLAMP Musgraves Province time series data: https://dx.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

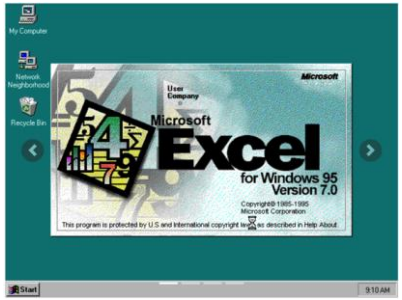
2030 Bringing the datasets together: The standards dilemma

- But in reality, the data we required for aggregated high resolution national coverages were distributed across geographically separate organisations
- Each organisation was acquiring and publishing primary and derivative datasets to meet the specific requirements and standards of their funders, partners, researchers, etc.
- The majority of these standards (ASCII-based, CSV, Excel, ASEG-GDF) were not suitable for HPC
- Many of these formats are for also data only: the metadata is separate and attached as a “side car”



‹ of your photo as the motorcycle. Metadata is digital information (date/location, descriptions, captions, keywords,

https://www.picturesandstories.com/news/2017/2/13/metadata-writing-on-the-back-of-a-digital-photo



Source: https://winworldpc.com/product/microsoft-excel/95

AusLAMP recording devices

PR6-24 Earth Data Logger

https://www.gfz-potsdam.de/en/section/geophysical-imaging/infrastructure/geophysical-instrument-pool-potsdam-gipp/pool-components/depas-pool/recorder-/-earthdata-pr6-24

LEMI 424

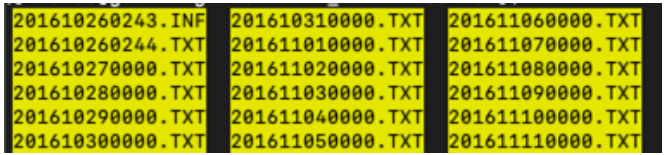
https://www.isr.lviv.ua/lemi424.htm

University of Adelaide Orange Boxes

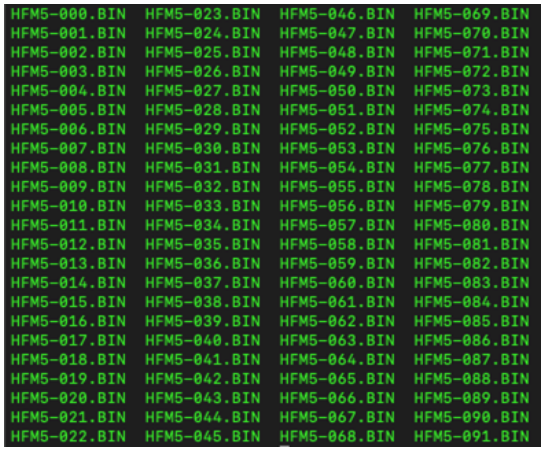
AusLAMP data formats

PR6-24 Earth Data Logger

LEMI 424

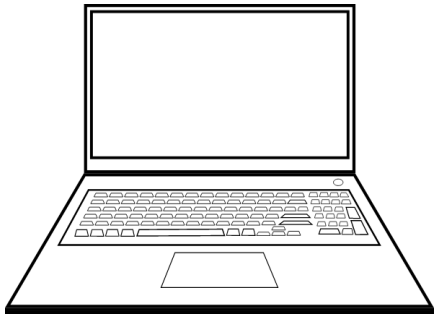


University of Adelaide Orange Boxes

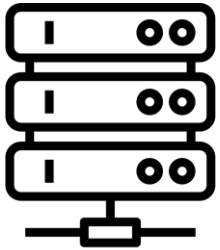


Each recording device captured data in a bespoke manner

AusLAMP data storage and delivery

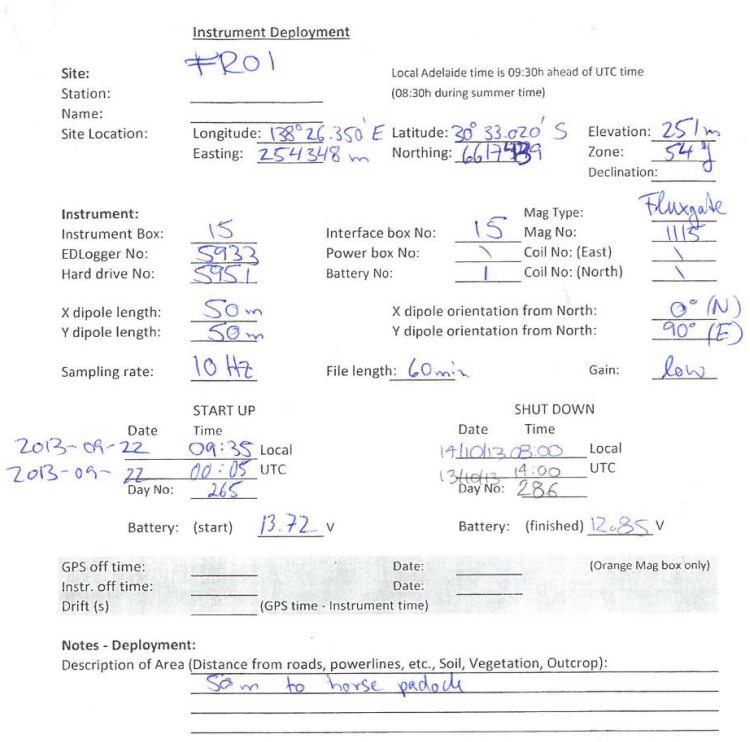


Dropbox



Local Server

AusLAMP Metadata



Instrument Deployment

Site: ~~R01~~

Station: _____

Name: _____

Site Location: Longitude: 138° 26.350' E Latitude: 30° 33.020' S Elevation: 251 m

Easting: 254348 m Northing: 6617~~9~~3~~9~~ Zone: 54 J

Declination: _____

Instrument:

Instrument Box: 15 Interface box No: 15 Mag Type: Fluxgate

EDLogger No: 5933 Power box No: \ Mag No: 1115

Hard drive No: 5951 Battery No: 1 Coil No: (East) \

Coil No: (North) \

X dipole length: 50 m X dipole orientation from North: 0° (N)

Y dipole length: 50 m Y dipole orientation from North: 90° (E)

Sampling rate: 10 Hz File length: 60 min Gain: low

START UP

Date Time

2013-09-22 09:35 Local

2013-09-22 00:05 UTC

Day No: 265

Battery: (start) 13.72 V

SHUT DOWN

Date Time

19/10/13 08:00 Local

13/10/13 19:00 UTC

Day No: 286

Battery: (finished) 12.85 V

GPS off time: _____ Date: _____ (Orange Mag box only)

Instr. off time: _____ Date: _____

Drift (s) _____ (GPS time - Instrument time)

Notes - Deployment:

Description of Area (Distance from roads, powerlines, etc., Soil, Vegetation, Outcrop):

50 m to horse paddock

AusLAMP Metadata

1	Not Used	Not Used	Not Used	Not Used	Compulsory	Optional	Optional	Optional	Compulsory	Compulsory	Compulsory	Compulsory	Compulsory	Compulsory	Compulsory	Optional
2	Site Number	Coords Available	Deployed Date Entered	Pickedup Date Entered	Site Number	UTC Start Time Minute	UTC Start Time Second	Deployment Julian Day	Recording Method	MT Recorder Type/ Model	Magnetometer Type/ Model	Electrode Type/ Model	Power Source Type/ Model	Data Confidentiality	North Reference	Drift calculation end time - Time
406	SA333	Coords	dep date	Retrival date	SA333	44	12	139	LP	Earth Data Recc	Bartington, Mo	SDEC (France), N	12Volt 72 Amp/Hr Battery, Power Supply Charging - Solar Panel, 60Watt	Open	Magnetic North	
407	SA333-2A	Coords	dep date	Retrival date	SA333-2A	23	0	170	LP	Earth Data Recc	Bartington, Mo	SDEC (France), N	12Volt 72 Amp/Hr Battery, Power Supply Charging - Solar Panel, 60Watt	Open	Magnetic North	
408	SA333-2B	Coords	dep date	Retrival date	SA333-2B	56	0	170	LP	Earth Data Recc	Bartington, Mo	SDEC (France), N	12Volt 72 Amp/Hr Battery, Power Supply Charging - Solar Panel, 60Watt	Open	Magnetic North	
409	SA334	Coords	dep date	Retrival date	SA334	55	0	139	LP	Earth Data Recc	Bartington, Mo	SDEC (France), N	12Volt 72 Amp/Hr Battery, Power Supply Charging - Solar Panel, 60Watt	Open	Magnetic North	
410	SA335	Coords	dep date	Retrival date	SA335	53	51	140	LP	Earth Data Recc	Bartington, Mo	SDEC (France), N	12Volt 72 Amp/Hr Battery, Power Supply Charging - Solar Panel, 60Watt	Open	Magnetic North	
411	SA336	Coords	dep date	Retrival date	SA336	17	24	140	LP	Earth Data Recc	Bartington, Mo	SDEC (France), N	12Volt 72 Amp/Hr Battery, Power Supply Charging - Solar Panel, 60Watt	Open	Magnetic North	
412	SA337	Coords	dep date	Retrival date	SA337	6	48	142	LP	Earth Data Recc	Bartington, Mo	SDEC (France), N	12Volt 72 Amp/Hr Battery, Power Supply Charging - Solar Panel, 60Watt	Open	Magnetic North	
413	SA338	Coords	dep date	Retrival date	SA338	3	17	142	LP	Earth Data Recc	Bartington, Mo	SDEC (France), N	12Volt 72 Amp/Hr Battery, Power Supply Charging - Solar Panel, 60Watt	Open	Magnetic North	
414	SA339	Coords	dep date	Retrival date	SA339	26	51	143	LP	Earth Data Recc	Bartington, Mo	SDEC (France), N	12Volt 72 Amp/Hr Battery, Power Supply Charging - Solar Panel, 60Watt	Open	Magnetic North	
415	SA340	Coords	dep date	Retrival date	SA340	35	0	140	LP	Earth Data Recc	Bartington, Mo	SDEC (France), N	12Volt 72 Amp/Hr Battery, Power Supply Charging - Solar Panel, 60Watt	Open	Magnetic North	
416	SA340-2	Coords	dep date	Retrival date	SA340-2	1	0	143	LP	Earth Data Recc	Bartington, Mo	SDEC (France), N	12Volt 72 Amp/Hr Battery, Power Supply Charging - Solar Panel, 60Watt	Open	Magnetic North	
417	SA341	Coords	dep date	Retrival date	SA341	10	0	324	LP	Earth Data Recc	Bartington, Mo	SDEC (France), N	12Volt 72 Amp/Hr Battery, Power Supply Charging - Solar Panel, 60Watt	Open	Magnetic North	
418	SA342	Coords	dep date	Retrival date	SA342	54	0	324	LP	Earth Data Recc	Bartington, Mo	SDEC (France), N	12Volt 72 Amp/Hr Battery, Power Supply Charging - Solar Panel, 60Watt	Open	Magnetic North	
419	SA344	Coords	dep date	Retrival date	SA344	35	0	135	LP	Earth Data Recc	Bartington, Mo	SDEC (France), N	12Volt 72 Amp/Hr Battery, Power Supply Charging - Solar Panel, 60Watt	Open	Magnetic North	
420	SA344-2	Coords	dep date	Retrival date	SA344-2	6	16	171	LP	Earth Data Recc	Bartington, Mo	SDEC (France), N	12Volt 72 Amp/Hr Battery, Power Supply Charging - Solar Panel, 60Watt	Open	Magnetic North	
421	SA345	Coords	dep date	Retrival date	SA345	59	15	132	LP	Earth Data Recc	Bartington, Mo	SDEC (France), N	12Volt 72 Amp/Hr Battery, Power Supply Charging - Solar Panel, 60Watt	Open	Magnetic North	
422	SA346	Coords	dep date	Retrival date	SA346	39	0	132	LP	Earth Data Recc	Bartington, Mo	SDEC (France), N	12Volt 72 Amp/Hr Battery, Power Supply Charging - Solar Panel, 60Watt	Open	Magnetic North	
423	SA347	Coords	dep date	Retrival date	SA347	7	0	234	LP	Earth Data Recc	Bartington, Mo	SDEC (France), N	12Volt 72 Amp/Hr Battery, Power Supply Charging - Solar Panel, 60Watt	Open	Magnetic North	
424	SA348	Coords	dep date	Retrival date	SA348	21	0	232	LP	Earth Data Recc	Bartington, Mo	SDEC (France), N	12Volt 72 Amp/Hr Battery, Power Supply Charging - Solar Panel, 60Watt	Open	Magnetic North	
425	SA349	Coords	dep date	Retrival date	SA349			231	LP	Earth Data Recc	Bartington, Mo	SDEC (France), N	12Volt 72 Amp/Hr Battery, Power Supply Charging - Solar Panel, 60Watt	Open	Magnetic North	
426	SA350	Coords	dep date	Retrival date	SA350	41	13	231	LP	Earth Data Recc	Bartington, Mo	SDEC (France), N	12Volt 72 Amp/Hr Battery, Power Supply Charging - Solar Panel, 60Watt	Open	Magnetic North	
427	SA351	Coords	dep date	Retrival date	SA351	52	20	231	LP	Earth Data Recc	Bartington, Mo	SDEC (France), N	12Volt 72 Amp/Hr Battery, Power Supply Charging - Solar Panel, 60Watt	Open	Magnetic North	
428	SA354	Coords	dep date	Retrival date	SA354	18	52	79	LP	Earth Data Recc	Bartington, Mo	SDEC (France), N	12Volt 72 Amp/Hr Battery, Power Supply Charging - Solar Panel, 60Watt	Open	Magnetic North	

The great South Australian AusLAMP MT Time Series

https://www.cleanpng.com/png-internet-scavenger-hunt-ocean-grove-camp-meeting-a-1397955/

- **417** AusLAMP survey sites were recorded in South Australia between 2014 and 2018.
- Four comprehensive data rescue passes were conducted in order to maximise the number of sites recovered and ensure their availability at NCI.
- The time series data was sourced from four distinct individuals, each with their own unique methods for organising the data and metadata. These methods included storage on local servers, hard drives and cloud platforms, with metadata managed in various formats such as spreadsheets, physical documents and the Evernote library software.

The great South Australian AusLAMP MT Time Series

https://www.cleanpng.com/png-internet-scavenger-hunt-ocean-grove-camp-meeting-a-1397955/

	Number of sites rescued	Percentage of total sites
Original ingest (2020):	205	49.1%
Second pass (December 2022):	270	64.7%
Third pass (June 2023):	386	92.5 %
Fourth pass (July 2023):	400	95.9%

Rescuing data is **not** enough - converting to modern international standards

ELSEVIER

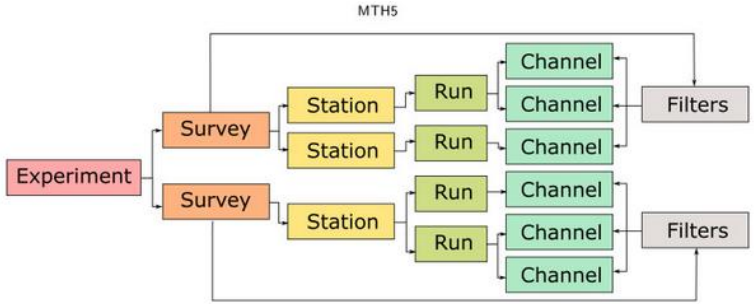
Computers & Geosciences

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]



- Current version of MT_metadata has 355 attributes
  - 152 required parameters
  - 203 non mandatory parameters (but useful to have)
- For legacy AusLAMP metadata, we created a Legacy AusLAMP MT_metadata profile that was a 72 element subset of the full MT_metadata standard
- Held regular meetings with the USGS based lead developer of MT_metadata and added in some revisions to the original standard (e.g., survey.funding_source.organization, survey.state, more options for station.release_license)
- Needed to go back to field notebooks and original collectors to get the additional metadata
- Changing to modern self-describing HPC/data intensive formats is NOT a simple mathematical transformation

Expanding our geophysical data collections on HPC

Geoscience Australia Geophysics Reference Data Collection

This collection has been compiled by Geoscience Australia from an extensive archive of over 2,200 geophysical surveys dating back to 1947. The datasets have been acquired by Geoscience Australia and

National ASTER Map of Australia

AuScope Australian National Virtual Core Library (NVCL) Collection

The Australian
world's largest
mineralogical

NVCL data ha
construction.
zone, playing
hyperspectral

NVCL facilities
developed Hyl
The Hylogger-
spectroscopy,
continuous do
CSIRO-develo

AusPass Passive Seismic Collection

The Passive Seismic Data contains mirrored datasets hosted on the Australian Passive Seismic Server (AusPass). AusPass is an initiative supported by funding from AuScope and the Australian National University to host passive seismic (i.e. continuous waveform) data via the standardised FDSN Web Service protocol. Data largely consists of distinct temporary seismic arrays consisting of tens to hundreds of individual stations, but also includes the permanent Australian Seismometers in Schools (AuSIS, S1)

AuScope Distributed Acoustic Sensing (DAS) Collection

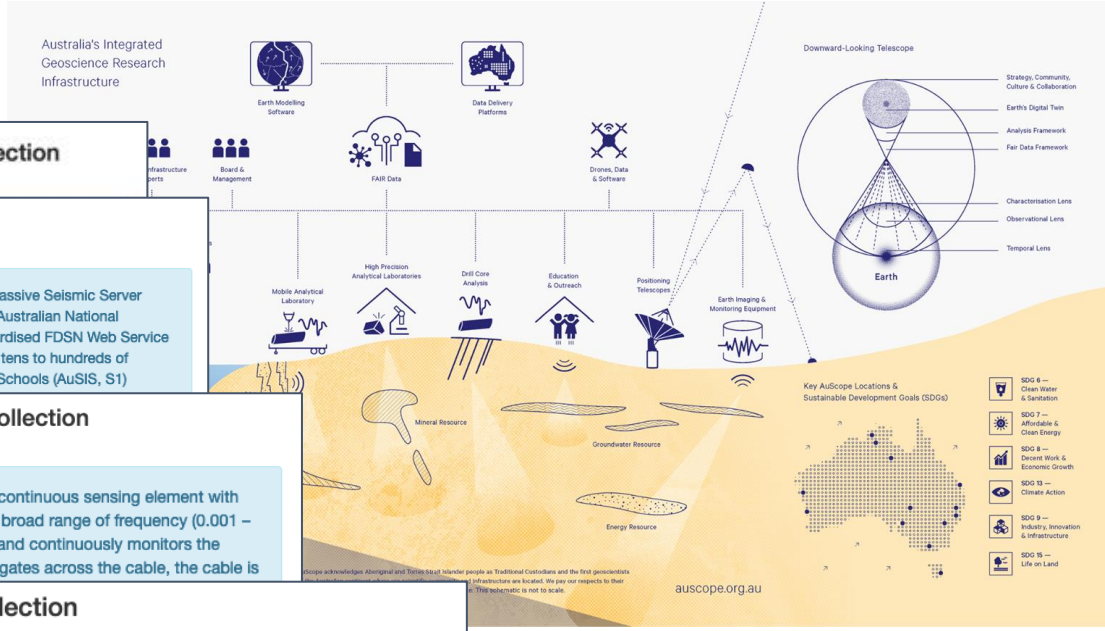
Distributed Acoustic Sensing (DAS) transforms fibre optic cable as one continuous sensing element with thousands of sensors at meter-spacing along the cable, measuring at a broad range of frequency (0.001 – 1000 Hz). A DAS interrogator sends laser signal pulses along the cable and continuously monitors the phase and amplitude of the backscattered light. As seismic wave propagates across the cable, the cable is slightly
phase
boreho
new op
source

This da
collect

AuScope Magnetotellurics (MT) Collection

This collection includes AuScope-funded geophysical data, as well as data that has been acquired by universities, industry, federal/state government agencies since the 1950s. Magnetotelluric (MT) survey data has been collected from thousands of sites across Australia.

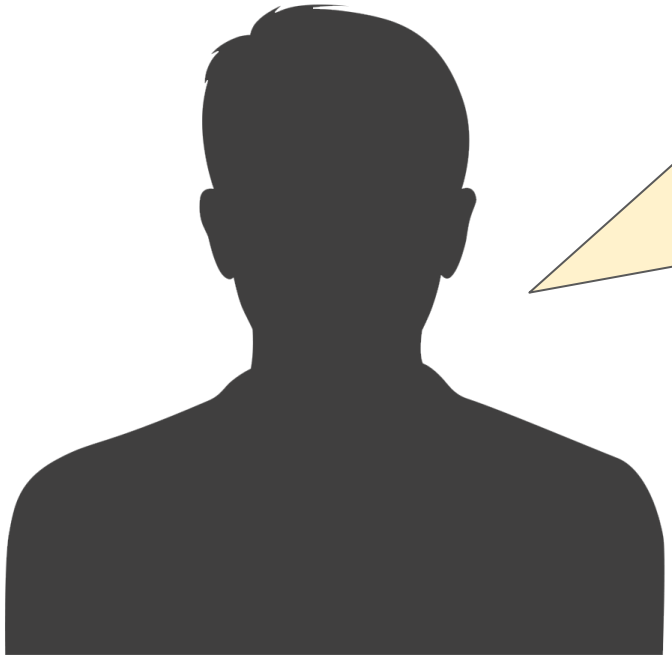
https://geonetwork.nci.org.au



Lessons learnt

- Data rescue can be a time consuming, costly and monotonous endeavor:
  - Requires extensive analysis and multiple attempts to successfully recover lost data
  - No guarantee of complete recovery
- Recovered data can be incompatible with current software or systems leading to additional challenges when trying to access or use the data
- Obtaining metadata is critical in order to be able to use the data as intended
- We could not have successfully undertaken our data rescue without identifying the original data collectors
  - Delaying our efforts any further might have jeopardised the rescue, as the collectors were in the process of transitioning to new institutions or had retired
- **Prioritise the development of a comprehensive data management strategy at the outset of a project/program in order to prevent the need for a data rescue later on!**

Long Term Investment in Data Curation



"digital information lasts forever –
or five years, whichever comes first"

Jeff Rothenberg

©iStock

NCI AUSTRALIA

NCI Contacts

General enquiries: +61 2 6125 9800

Support: help@nci.org.au

Email: hannes.hollmann@anu.edu.au

Address

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

License