Developing International partnerships for the harmonisation of solid Earth and environmental data infrastructures

Lesley Wyborn\textsuperscript{1}, Erin Robinson\textsuperscript{2}, Tim Rawling\textsuperscript{3}, Simon Cox\textsuperscript{4}, Ben Evans\textsuperscript{1}, Kerstin Lehnert\textsuperscript{5}, Jens Klump\textsuperscript{6}, Helen Glaves\textsuperscript{7}, Kirsten Elger\textsuperscript{8}, Shelley Stall\textsuperscript{9} and Mohan Ramamurthy\textsuperscript{10}

\textsuperscript{1}National Computational Infrastructure, ANU, Australia and Australian National Data Service, Australia
\textsuperscript{2}Earth Science Information Partners, USA
\textsuperscript{3}AuScope Ltd, Australia,
\textsuperscript{4}Land and Water, CSIRO, Australia
\textsuperscript{5}Lamont-Doherty Earth Observatory, Columbia University, USA
\textsuperscript{6}Mineral Resources, CSIRO, Australia
\textsuperscript{7}British Geological Survey, United Kingdom
\textsuperscript{8}GFZ German Research Centre for Geosciences, Germany
\textsuperscript{9}American Geophysical Union, USA
\textsuperscript{10}EarthCube, UCAR, USA
Evidence

1. Globally, solid Earth science data are collected by large numbers of organizations across the academic, government and industry sectors.
2. Spatially, the data collected covers multiple domains extending from the crust, through the lithosphere and mantle to the core.
3. Many observed phenomena cross national, if not continental, boundaries.

Question:
Why can’t we work together to develop international networks of Earth and environmental science researchers to contribute to growing global challenges such as:
   A. Scarce non-renewable resources
   B. Risk reduction for natural hazards
   C. Fundamental research on the nature of the planet
The last decade has seen a dramatic growth in online Earth science datasets, online tools, and computational power, particularly utilising Cloud or HPC hosted data and compute resources.

But

- There are inconsistent and incompatible data descriptions and formats.
- Software is developed locally around specific applications and data sources.
- There is a multiplicity of software providing similar and overlapping functions.

Quote from Industry supporter of a multi-client project in 2004:
“The Minerals Industry spends 80% of its time finding and reformatting data – what if that 80% could be used to develop better and smarter algorithms to process the data”
OneGeology's aim is to improve the WWW accessibility (including interoperability) and usefulness of global geoscience data needed to address many societal issues including mitigation of hazards, meeting resource requirements, and climate change.

- Started in Brighton (UK) in 2007

- 119 countries participating as at 21 May 2018

http://www.onegeology.org/
Need for community agreed standards for Interfacing and sharing

https://www.power-plugs-sockets.com/
But is it also about infrastructure

<table>
<thead>
<tr>
<th>Gauge</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,000 mm (3 ft 3(\frac{3}{8}) in)</td>
<td>Metre gauge</td>
</tr>
<tr>
<td>1,087 mm (3 ft 6 in)</td>
<td>Three foot six inch gauge</td>
</tr>
<tr>
<td>1,435 mm (4 ft 8\1(\frac{1}{2}) in)</td>
<td>Standard gauge</td>
</tr>
<tr>
<td>1,520 mm (4 ft 11(\frac{7}{8}) in)</td>
<td>Five foot and 1520 mm gauge</td>
</tr>
<tr>
<td>1,524 mm (5 ft)</td>
<td>Finnish gauge</td>
</tr>
<tr>
<td>1,600 mm (5 ft 3 in)</td>
<td>Five foot three inch gauge</td>
</tr>
<tr>
<td>1,668 mm (5 ft 6(\frac{1}{8}) in)</td>
<td>Iberian gauge</td>
</tr>
<tr>
<td>1,676 mm (5 ft 6 in)</td>
<td>Five foot six inch gauge</td>
</tr>
</tbody>
</table>

Source: [https://en.wikipedia.org/wiki/Track_gauge](https://en.wikipedia.org/wiki/Track_gauge)
• Data infrastructure is digital infrastructure promoting data sharing and consumption

• Data includes data software, samples and models: all are integral across government, academia and industry sectors in Earth, space, and environmental science research and are routinely ‘shared’ for recombination, reuse, to test reliability, etc

• To be shared and reused **effectively and efficiently**, information about data samples, methods, and tools need to be standardized, available, and linked across activities.

• Remember: the size of the community that you can interact with is the size of the community that uses the same standard – who determines what standard ‘wins’
Seminal publications about learning from the past in Infrastructures

Understanding Infrastructure: Dynamics, Tensions, and Design


Paul N. Edwards
Steven J. Jackson
Geoffrey C. Bowker
Cory P. knobel

January 2007

Common Patterns in Revolutionary Infrastructures and Data

| 2007 |

Common Patterns in Revolutionary Infrastructures and Data

| 2018 |

1. Summary

Societies have seen large infrastructures emerge when new technologies become available. From history we see that such infrastructures can have a huge influence on all aspects of societal life. Moreover, some patterns appear to recur in the evolution of such infrastructures. The adoption of the arrival of a new technology leads to a cycle of creation, use and transformation. A huge “infrastructure space” emerges and fragmentation results. Some solutions are more attractive than others, but in the end, there is a final phase where the system converges towards broader accepted principles and specifications that lead to exploitation and standardisation.

It appears that “the digital infrastructure” is moving into such a large infrastructure, with a potentially large influence on societies, individuals and commerce, in order to gain new insights into complex relationships in nature, societies and within. By integrating data from different fields we have seen an explosion of data-driven solutions for data management, access and processing. We have entered a phase of creation. Also, we have an increasing clear view of the current inefficiencies in working with data. These inefficiencies need innovation and broad participation, which will become even more important as billions of devices produce the data deluge of the internet of things. Stakeholders have begun looking for ways to deal with convergence that would increase efficiency without hampering innovation.

Compiling the evolution of the digital infrastructure with the evolution of infrastructures in electrification, computer networking and of interconnection (Internet), we can observe that, despite all initiatives already taken, we have not realised convergence on a set of universals that would lead developments and create a momentum towards an efficient and interoperable data infrastructure. We propose that such a set of universals could be based on the concepts of “Digital Objects” (DOI), persistent identifiers (PID), and metadata (including data cataloging). These concepts could greatly reduce current inefficiencies in data processing and chart the way towards automatic processing. In particular, the Core Data Model of the Research Data Alliance (RDA) provides a design for a universal Digital Object Access Protocol (DOIAP, comparable to IP for the Internet or HTTP for the Web) which can interconnect many organisations of data. In the future, such as cloud systems, file systems, SQL databases, NoSQL databases and so forth, the agreement on a fairly simple but powerful universal demonstration such as PID, DOIs, and DOIAP could make the infrastructure for many developers to invest in data infrastructure building. We believe that it is time to take this step towards convergence.

Acknowledgements

We would like to acknowledge the many contributions to this discussion from colleagues during the last months. In particular, we should mention here Robert Katz, Jonny Luserna, David Sang, and various colleagues from the Research Data Alliance (RDA) and the CORDAP initiatives.

1. Creation is a term used to describe the development of culture and languages. It describes a process in which communities generate and exchange ideas, empowers various stakeholders.
## Patterns of Infrastructure Development

<table>
<thead>
<tr>
<th>Edwards et al. 2007</th>
<th>Wittenburg &amp; Strawn 2018</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Deliberate and successful design of ‘local’ systems.</td>
<td>1. Inventions and development of start-up systems</td>
</tr>
<tr>
<td>2. Technology transfer across domains and locations</td>
<td>2. Technology transfer between regions and also society (creolization)</td>
</tr>
<tr>
<td>3. Infrastructure form via gateways that allow dissimilar systems to be linked into networks</td>
<td>3. Planning for system growth where &quot;reverse salients&quot; need to be tackled</td>
</tr>
<tr>
<td>4. Substantial momentum (mass, velocity, direction)</td>
<td></td>
</tr>
</tbody>
</table>

**System Building**

**Growth**

**Consolidation**
Patterns of Infrastructure Development

Kerstin Lehnert: EGU McHarg Lecture April 2018

Wittenburg & Strawn 2018

1. Inventions and development of start-up systems
2. Technology transfer between regions and also society (creolization)
3. Planning for system growth where "reverse salients" need to be tackled
4. Substantial momentum (mass, velocity, direction)

System Building
Growth
Consolidation

Wittenburg, P., and Strawn, G., 2018. Common Patterns in Revolutionary Infrastructures and Data
Creolization

- New components are continuously introduced trying to solve specific challenges
- Capabilities grow unevenly (e.g. big vs small data)
- Fragmentation

Leads to

- Inefficiencies in use and costs
- Winners & losers: some solutions are more promising and get more attraction
- Better understanding the underlying rules, principles and limitations.

After Wittenburg & Strawn, 2018)
Attraction via “Universals”

“Universals are … essential to create a momentum by overcoming fragmentation and achieving economies of scale.

- “Simple” principles, broadly supported
- Only influence directly a specific part of the overall infrastructure, enable efficiency at the top layers
- Form stable basis for new developments

After Wittenburg & Strawn, 2018)
The lessons of history: Industrial Revolution vs the Geoinformatics Revolution

- 1776 - Invention of steam engine
- 1829 - Invention of railways
- 1834 - First rail-networks
- 1880 - First Standards Association for individual components
- 1890 - Manufacturing age
- 1940 - Invention of the computer
- 1989 - First Generation Internet
- 1996 - First Grid networks
- 1996 - First Standards for components: W3C & XML appear
- 2007 - OneGeology & Geoinformatics Age

The Screw – the Minimum Component of the Industrial Revolution

The anatomy of the thread of a screw

Standards of the industrial age were developed at the level of the lowest common component
Universal Pattern in Science:
An observation is any action whose result is an *estimate* of a property value

**Scope**
- In situ observations
- Remote sensing
- Ex-situ (laboratory) observations
- Numerical models/simulations
- Forecasts
- Interpretations, classifications
The problem: all these are building relevant data infrastructures.
With all those organisations where do we Compete vs Collaborate?

We need collaborative informatics and computational platforms on which competitive research can be undertaken.

But where do we put the boundary? What are the universals that we can all build on?
• Systems can be built build in isolation, often competitively
• Where do we put the boundary?
  • When do local factors mean that international collaboration is not feasible
  • When do deadlines require building a one-off system?
  • Is funding an issue for collaboration?
  • Is there a trade off?
Varying boundaries between competition and collaboration
At the extreme end of competition: the special snowflakes

https://untamedhellcat.wordpress.com/2014/03/09/how-special-snowflakes-kill-a-raiding-guild/

European Plate Observing System (EPOS)

(Slides provided by Helen Glaves)
What is EPOS?

Overarching European Research infrastructure for the solid Earth

EU-funded H2020 project

Integrating diverse range of European Research Infrastructures for solid Earth science

On the European Strategy Forum on Research Infrastructures (ESFRI) Roadmap (Landmark?)
EPOS Functional Architecture

Data collection
Data generation

Data curation
Metadata
Registration
Community Services

Interoperability
Brokerage

Metadata registry
Processing
Aggregation
Integrated analyses
Visualization
Introduction

AuScope Model

- Geodetic data
- Earth Imaging
- Composition & Evolution
- Subsurface Observatory
- Materials & Properties

---

National Geospatial Reference Framework
AuScope National Geo Transect Program

- AuScope Grid
  - Storage Management
  - Access Interoperability

- AuScope Simulator
  - Modelling, Data, Inversion

- Earth Model

---

Industry Portal
Research Portal
Policy Portal
Education Portal
ENVRIplus
Participation

- 20 research infrastructures
- 7 associated RIs
- 4 domains (biosphere, marine, atmosphere and solid Earth)
GeoScience DWG position

• Main targets:
  – Harmonize geoscience data expression and facilitate usage
  – Stay connected and even propose enhancements to technologies / solutions
  – Connect people interested in the geoscience topic
  – Ensure proper connections with other groups / communities inside and outside OGC
Another IE driven by the GeoScience community

- Environmental Linked Features IE (ELFIE)

  ‘Demonstrate the use of existing and pending OGC standards for the encoding of environmental observation data in an integrated dataset of features linked according to ReSTful and Linked Data principles.’

- Initiators:
  - U.S. Geological Survey (US)
  - Land Information New Zealand (NZ)
  - BRGM (FR)

- Participants
  - Tumbling Walls and Dewberry (US)
  - Meta-linkage (AU)
  - INSPIRE (EU)
  - Natural Resources Canada (CA)
  - Manaaki Whenua and Horizons Regional Council (NZ)

From: https://github.com/opengeospatial/ELFIE
ESIP Vision: to be a leader in collection, stewardship and use of Earth science data, information and knowledge that is responsive to societal needs.
ESIP - Volunteer-driven, transdisciplinary, international

CROSS-DOMAIN
- Science Software
- Drupal
- Open Source
- Air Quality
- Geospatial
- Energy & Climate
- Agriculture & Climate

Products & Services
- Visioneers
- Earth Science Collaboratory

Discovery
- Information Quality
- Documentation
- Semantic Web

Decisions
- Information Stewardship
- Data Study

Info tech & Interoperability
- Cloud Computing
- Semantic Web

Collaboration
- Committees
- Working Groups
- Clusters

Education & Data Stewardship
- Education
- Disasters

Era Science Information Portal
ESIP GROUPS

Standing Committees
• Data Stewardship
• Education
• Information Technology and Interoperability
• Products and Services (Retired)
• Semantic Technologies

Administrative Committees
• Constitution and Bylaws (Renamed Governance)
• Finance and Appropriations
• Partnership
• Nominations

Working groups
• Visioneers
• Energy & Climate
  Data Management Training

Clusters
• Ag & Climate
• CLEAN Network
• Cloud
• Data Coordination
• Data Mgmt Training
• Data to Decisions*
• Disaster Lifecycle
• Discovery
• Documentation
• Drones
• Earth Science Data Analytics
• Energy & Climate
• EnviroSensing
• Information Quality

• Pre-prints/EarthArXiv*
• Science Communication
• Science Software
• Semantic Technologies
• Software and Services Citations**
• Sustainable Data Mgmt
• Toolmatch
• Usability
• VR/AR
• Web Services

*New Cluster
** Starting 2018

… And yours?

http://esipfed.org/collaboration-areas
Enabling FAIR Data Project

Align publishers and repositories in following best practices to enable FAIR and open data and to create workflows so that researchers will have a simplified, common experience when submitting their paper to Earth and space science journals

1. ESS publishers will follow consistent and rigorous policies and guidelines for sharing and citing data used in scholarly literature;

2. Open ESS repositories will enable those policies and other data applications by providing persistent identifiers, rich metadata, and related services for the data they hold; and

3. ESS researchers will understand how to consistently share, document, and reference the data they collect and use

So who is involved?

Leading ESS Publishers:
• Nature
• Science
• Proceedings of the National Academies (PNAS)
• PLOS
• Hindawi
• Elsevier
• Wiley
• AGU
• Copernicus
• Taylor & Francis
• American Meteorological Society
• American Astronomical Society

Leading Data Repositories and associations
• Re3Data
• NOAA NCEI
• USGS ScienceBase
• MaGIC
• Pangaea
• DANS
• IRIS
• GFZ
• Dryad
• Figshare
• Zenodo
• Center for Open Science
• DataOne
• IEDA
• NCI
• NCAR
• UNAVCO
• Unidata
• World Data Center for Climate (WDCC)
• World Data System members

Editorial Workflow Vendors:
• EJ Press
• ScholarOne
• Aries
• Coko
What constitutes a trustworthy repository for preservation of data?

Core Trust Seal certification of digital data repositories
There are numerous groups working on discovery, description and citation of software including:

1. Force 11 Software Citation Implementation Working Group
2. RDA Software Source Code Interest Group
3. ESIP Software and Services Citations cluster
4. WSSSPE (Working Towards Sustainable Software For Science)
5. DataCite (through minting DOI for software)
6. Software Heritage (a web archive to to collect, preserve, and share all software that is publicly available in source code form).

In Australia we have established a local Research Software Interest Group to tackle this issue and try to define best practice internationally

Trialling our decisions with a suite of open source Magnetotelluric Software
Evolution of the O&M Model

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


http://7-themes.com/6964893-fa-18e-jet-fighter.html
Universal Pattern in Science:
An observation is any action whose result is an estimate of a property value

Scope
- In situ observations
- Remote sensing
- Ex-situ (laboratory) observations
- Numerical models/simulations
- Forecasts
- Interpretations, classifications

https://www.w3.org/TR/vocab-ssn/
Motivation for a common model

- Integrated analysis and modelling
  - Discovery & data integration a significant challenge
  - Different disciplines use different words for the same things

- O&M provides a standard, domain-neutral terminology
  - Reduces ambiguity
  - Increases interoperability

- **X-disciplinary terminology**
  - Many private contracts one public agreement!

Slide courtesy of Simon Cox
Which leads to vocabularies and ontologies....
Universal Pattern in Science:
An observation is any action whose result is an *estimate* of a property value

**Scope**
- In situ observations
- Remote sensing
- Ex-situ (laboratory) observations
- Numerical models/simulations
- Forecasts
- Interpretations, classifications

https://www.w3.org/TR/vocab-ssn/
• IGSN builds upon an existing technical base and community.
  • IGSN is based on the Handle System.
  • The IGSN technical & organizational architecture is developed in close alignment with DataCite.

• IGSN e.V. has 24 members in 5 continents
• 6.5 million samples have been registered so far using IGSN
• Newest members in 2017 are ANDS (Australia), BGS (UK), USGS (USA), CNRS & Ifremer (France)
• Creeping away from Geo into plants, water, etc
• We are moving beyond geosciences and need to align with broader community norms
• Need to move from adapting the ‘System’ within one community
• Common Kernel with Community agreed outer shells
Current status of interactions between Government, Academic and Industry

**Government Agencies:**
Data Rich and in knowledge on that data

**Industry:**
Providing use cases to drive developments

**Academia:**
Expertise in cutting edge HPC/HPD research, and software development

? ? ?
Current status of interactions between Government, Academic and Industry

Government Agencies:
Data Rich and in knowledge on that data

Industry:
Providing use cases to drive developments

Academia:
Expertise in cutting edge HPC/HPD research, and software development

MT data is using ASEG standards

MT data is using International Association of Geomagnetism and Aeronomy standards of the IUGG

© NCI Australia 2018

C3DIS May 2018: Wyborn et al. (lesley.wyborn@anu.edu.au)
And so back to infrastructure ‘boundaries’
And spanning boundaries between the incompatibles...
What we need now are mechanisms to internationally link these major infrastructures to provide:

1. efficiencies in funding (stop reinventing the wheel!)
2. an environment where the research efforts can create globally interoperable networks of solid Earth science data, information systems, software and researchers

Source: https://medium.com/@bansalbhavik9/reinvent-the-wheel-or-not-5013f6d1ac2c
Ocean Data Interoperability Platform (ODIP/ODIP II)

ODIP: October 2012 – September 2015
ODIP II: April 2015 – March 2018

Collaborative project:

• Europe, USA, Australia, Canada
Europe: 19 EU-funded partners (9 countries)

- NERC-BGS/BODC, MARIS, OGS, IFREMER, HCMR, ENEA, ULG, CNR, RBINS, TNO, AWI, BSH, RIHMI-WDC, VLIZ, UniHB, CSIC, 52ONorth, IEEE, SOCIIB

USA: 11 Organisations

- Scripps Institution of Oceanography (SIO), Woods Hole Oceanographic Institute (WHOI), Lamont-Doherty Earth Observatory (LDEO), Florida State University (FSU): Center for Ocean-Atmospheric Prediction Studies, NOAA, US-IOOS, UNIDATA, MMI, ESRI

Australia: 5 organisations

- University of Tasmania (IMOS), CSIRO, Geoscience Australia (GA), NCI, ANDS

International: 4 organisations

- UNESCO IOC-IODE, GEO/GEOSS, POGO, ICSU – WDS
ODIP/ODIP II Objectives and Outputs

- Development of a series of prototype interoperability solutions demonstrating coordinated approach to marine data management on a global scale
- Promote development of a common global framework for marine data management
- Output 1: Metadata from regional data discovery systems accessible via global portals GEOSS portal and IODE Ocean Data Portal (ODP)
- Output 2: ISO Cruise Summary Reports harvested from regional nodes and exposed in the POGO portal
- Output 3: Establishing a global SWE community of practice and working towards OGC Sensor Web enablement standards for ocean sensors
First Steps

• Formation of the ESIP/RDA Earth, Space and Environmental Sciences Interest Group (https://rd-alliance.org/groups/esiprda-earth-space-and-environmental-sciences-ig)

• Objective: Focus on awareness, and coordination where applicable, of independent efforts across the international Earth, space, and environmental science communities.

• Key participating groups and their use cases include:
  1. The Earth Science Information Partners (ESIP)
  2. The Australian AuScope program
  3. The EU H2020 European Plate Observing System (EPOS)
  4. The American Geophysical Union (AGU) Enabling FAIR Data
  5. The US NSF Environmental Data Initiative (EDI)
  6. The US NSF EarthCube
  7. The Open Geospatial Consortium Domain Working Groups
  8. The European Network for Earth Systems Modelling (ENES)
Who are we?

The International Brigade of Earth Science Boundary Spanners

Lesley Wyborn: National Computational Infrastructure, ANU, Australia (lesley.wyborn@anu.edu.au)

Erin Robinson: Earth Science Information Partners, Boulder, CO, United States (erinrobinson@esipfed.org)

Tim Rawling: AuScope Ltd, Melbourne, Australia (tim.rawling@unimelb.edu.au)

Simon Cox: Land and Water, CSIRO, Clayton, Vic, Australia (simon.cox@csiro.au)

Ben Evans: National Computational Infrastructure, ANU, Australia (ben.evans@anu.edu.au)

Kerstin Lehnert: Lamont-Doherty Earth Observatory, Columbia University, New York, USA (lehnert@ldeo.columbia.edu)

Jens Klump: Mineral Resources, CSIRO, Kensington, WA, Australia (jens.klump@csiro.au)

Helen Glaves: British Geological Survey, Nottingham, UK (hmg@bgs.ac.uk)

Kirsten Elger: GFZ German Research Centre for Geosciences, Potsdam, Germany (kelger@gfz-potsdam.de)

Shelley Stall: American Geophysical Union, Washington, D.C., United States (sstall@agu.org)

Mohan Ramamurthy: EarthCube, UCAR, Boulder, USA (mohan@ucar.edu)