

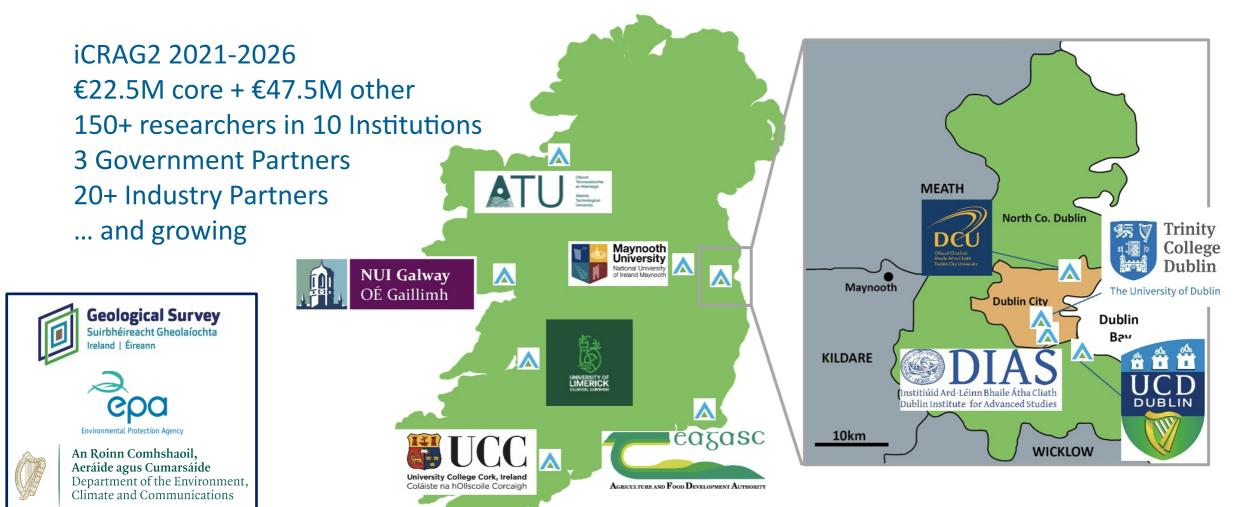
Perspective from a Irish-based data infrastructure

Chris Burbidge iCRAG, University College Dublin

EOSC NTE 2023-11-03

This presentation has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number R18462, and collaboration with partners.

A world-leading SFI research centre





iCRAG Geodata Platform

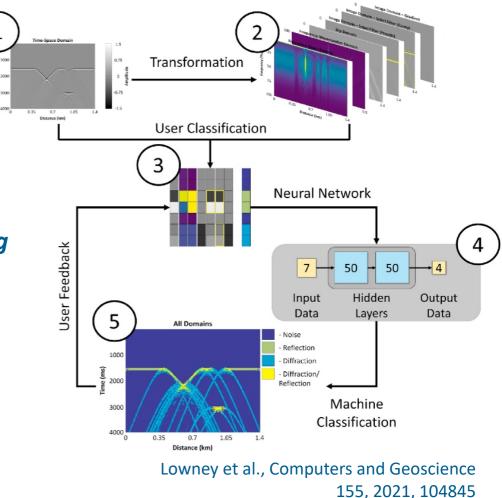
Geodata Platform underpins activities centre-wide

Patterns in data drive understanding of the data:

big data • *mathematical modelling* • *data analytics learning*

Improve on traditional 'model-driven' approach*

Increasingly now data infrastructure and management





Geodata Stakeholders

Centre Researchers, Students, Managers:

- specialist + training + inter-institutional + policy needs

National and International strategy and drivers

- Public, Industry, Funding Agencies, Government, Large Projects

Institutional strategy, security, liability - Large Projects & Networks, RII, IT, Library, GDPR, ... Legal

'Relevant', 'Practical', 'Inclusive', 'Non-Prescriptive'
 Data Organisation ● Data Sharing/Hosting ● Accessible Compute
 Actions ↔ Policies ↔ Infrastructure ↔ Resources
 Impact - Openness - Confidentiality - Intellectual Property

Who wishes to take responsibility + Who can take responsibility







Business and Architecture

Complex, diverse, needs around the centre *Community is the basis for the system*

Larger 'simpler' projects can self-organise *Make the system scalable for the future*

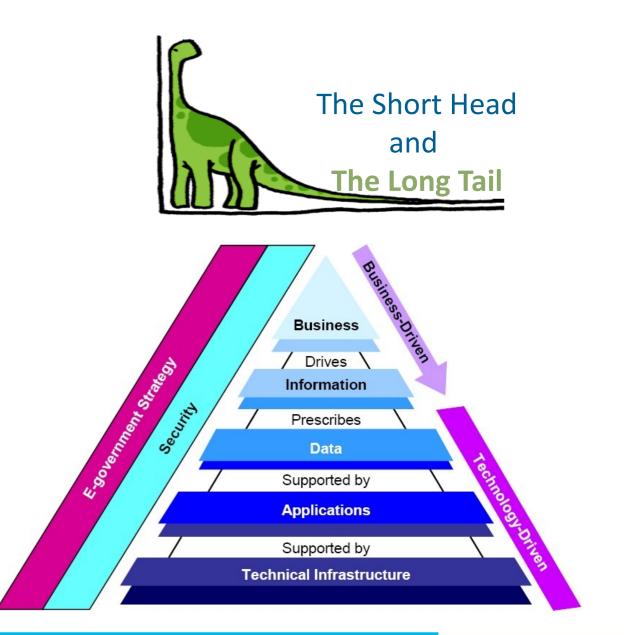
Largest not presently well catered for *Align the system with larger initiatives*

How to organise support for the Stakeholders?

Business defines needs

Technology defines possibilities

Data provides linkage for Enterprise





Enterprise Architecture

Shells of linked Interests

- individuals
- centre
- institutional
- national
- international

Making datasets citable

Heterogeneous Information Network

Three infrastructure foci

- accessible compute and GPU
- datawarehousing
- object storage & file sharing

| | | Central High Performance Com National | putation resources | | |
|--|---------------------|--|--|----------------------------------|---------------------------------|
| | | Caspir @ ICHEC - high performance distributed c | computation on Linux | | |
| | | Institutional | | | |
| | | Sonic @ UCD RIT Daedalus - high performance distributed computation on Linux | | | |
| | ···· iCRAG developr | | amont | | |
| | | iCRAG compute: | | | |
| | | GPU Server | 3/4D plotting | | |
| | | - modern high performance GPU geoscience modelling and | · · · · · · · · · · · · · · · · · · · | | iCRAG @ UCD Coordination |
| | | visualisation for Windows users | | | UCD |
| | | Compute Server | machine learning | | iCRAG |
| | | - high performance geoscience | 3 | | visualising/interpretation |
| | | cloud computation for Windows users and packages | statistical inference | | digitised historic records |
| | | Compute Server - IaaS+ high performance | regridding | $\land \land \vdash \mathrel{>}$ | commercial partners |
| | | geoscience computation on Linux | geophysical inversion | | model data production |
| | | Storage – fast cloud-based storage system | | | project team records |
| rving | | iCRAG datawarehouse: | | | indiv. research records |
| Data | | Storage – linked cloud storage s | space | | national databases |
| Ireland | | Active Datawarehousing Solutio - gathering, harmonising and automating diverse iCRAG data | transform | | Partner Umbrella Project |
| Data Centre | | origins, flows, 3/4D long tail data | | | |
| < | | iCRAG develo | | | Other Institutes |
| | | iCRAG active storage and shar | ing: IRLDAT | | ICRAG |
| alogues | | Active File Management and Sh Datalake Server Development - develop, build, test and prototype deployment | haring Solution file management and distribution | | big data synthesis |
| | | | | | lab data production |
| | | | | | |
| | | iCRAG develo | opment | | live monitoring data |
| Central Research Data Management Resources | | | | | |
| | | Institutional | | (Ch) | Other Projects |
| | | UCD Research IT - personal project tools and workspaces - department and project storage UCD Library - personal data management planning - department and project management methods - global data management policy | | ergagement | Other Coordinator Organisations |
| | | | | | Other Users |
| | | National | | | |
| | | HEAnet - National Research and Educa | tion Network | | |
| | | | | | |

European Plate Observ

European Geological E Infrastructure

Geological Survey of I National Geoscience D

Community Digital Repositories and Cata

- Deep Digital Earth

System

- Archive - Distribution

- GEOSS - ICHEC - Pangea - IRIS

....



Accessible Compute and GPU

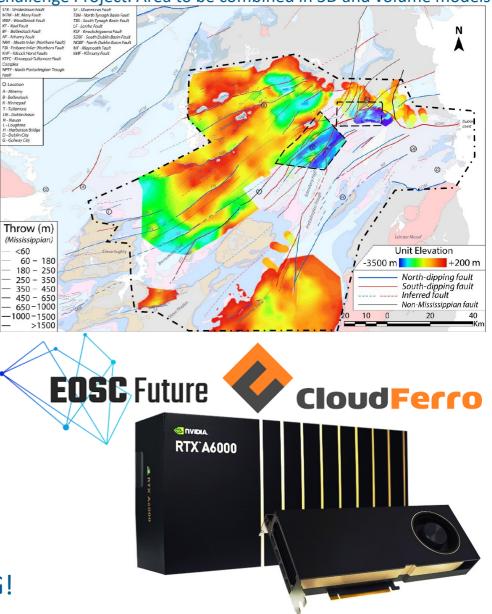
ML, AI, Modelling: Compute & general GPU
3/4D visualisation: gaming server GPU
Community familiarity: Windows GUI Linux HPC
Specialist geoscience apps: Licensed Windows

Remote in - Install - Upload - Run - Download

laaS+ proposals:

- EOSC FUTURE & national with OCRE supplier
- Institutional approvals yes; not yet datawarehousing
- Data supply contract yes; not yet laaS+

Dev Ops partnering with associated infrastructure:- for many "Short Head" projects on the go around iCRAG!





Metadata and Architecture

Automate to efficiently access The Long Tail

FAIR ≈ Machine Readable

Datawarehousing

- Challenging to resource: in-house work needed anyway
- Build, harmonise, verify the metadata foundation

EARTH:CHEC

ICHEC Euro CC2 Digital Innovation Flagship project

- iCRAG and ICHEC staff
- Mapping and Cataloguing Data Resources
- Formalised planning and cross checks
- Prepare for institutional Enterprise Architecture





Irish Sea Food Web Model. ICES. 2020. WKIrish6. http://doi.org/10.17895/ices.pub.5551

Datawarehousing

Need to FAIR The Long Tail

- align diverse interests and language @ metadata
- 'Geoscience' in the data ecosystem

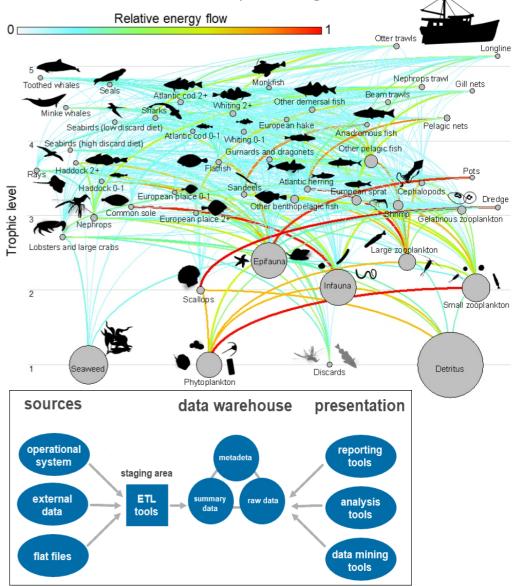
Active Research Data, not Archiving

- just an Extract-Transform-Load layer for Archives?
- open source vs. Commercial/isable?

Prospects

- in-house build/scale/amalgamate grass roots DBs
- build 'iCRAG dataset' into a larger DB or CRM
- integrate with (inter)national initiatives

All of the above, not one or another



https://phoenixnap.com/kb/data-warehouse-architecture-explained



Object Storage and File Sharing

IRLDAT - Piloting Irish Active Data Storage: B2Drop (+ Share...)

An all-iCRAG federated cloud based approach?

Active file management production need est ~300TB

IRLDAT iCRAG examples

- National Petrophysics DB, Aline Melo
- National Geothermal DB, Koen Torremans
- Lidar & Radiometric Surveys, Burbidge/Crowley
- Level 0 Geophysical Modelling, Ivan Lokmer
- Marine Blue Carbon modelling, Mark Chatting
- iCRAG pXRF Database, Evie Burton







https://www.cultureslate.com/lists/u7a57oxsczugt8bgwxhkrk6uhi1up2



Object Storage and File Sharing

IRLDAT Phase 1 - iCRAG on EUDAT

- 1. Central folder structure: administration
- 2. Federated accounts: registration, setup
- 3. Larger projects: EUDAT quotes + DMPs

IRLDAT Phase 2 - iCRAG on ICHEC AWS S3

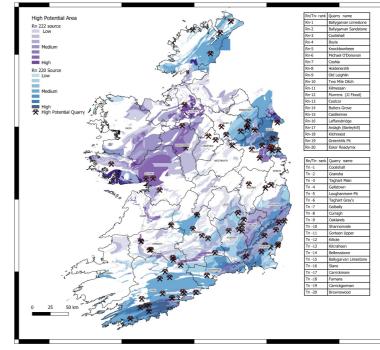
1. Folder structure recreated + B2Access groups

2. Key use case now set up:

- 27 GB geodata file mix sync, including >GB images and zips
- Windows B2Drop Ubuntu via Nextcloud client with Single Sign On
- 1st synthetic Geoscience dataset provided through EOSC Future contract

EOSC driving the iCRAG use case in IRLDAT

Radon and Thoron potential and Irish quarries for EOSC data supply prioritisation. Geochron Ltd., with permission





Conclusions

Geodata stakeholders - many engaged and being engaged: 'data' is broad

Storage and Sharing - being actively developed through EOSC and EUDAT

Metadata - linking stakeholder interests and architecture through EuroCC

Datawarehousing - the main challenge for FAIR: ideas laid out and evolving

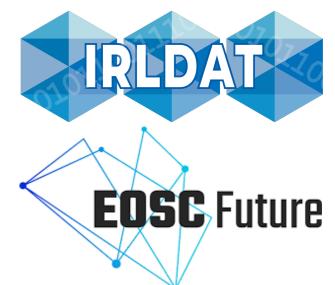
Overall:

- key is that we know enough to know what we need
- working closely with UCD RIT, ICHEC our national HPC, and HEAnet our NREN
- seeking progress through focussed & strategic research & infrastructure proposals

We need all of these...







Perspective from a Irish-based data infrastructure

Chris Burbidge iCRAG, University College Dublin

EOSC NTE 2023-11-03

This presentation has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number R18462, and collaboration with partners.