Contributions to the development of the next-generation NERC Environmental Data Service: Building Interoperability - a NERC Data Commons RoadMap

Report from the EDS Enhancement Project, Phase 1b

Gordon S Blair[1] with contributions from many members of the EDS and its collaborators including:
Garry Baker[2], Emma Bee[2], Jo Bowie[2], James Byrne[3], Louise Darroch[4], Alice Fremand[3], Wendy Garland[5], Kathryn Harrison[1], Rachel Heaven[2], Petra ten Hoopen[3], Martin Juckes[5], Andy Kingdon[2], Richard Kingston[6], Alexandra Kokkinaki[4], Edd Lewis[2], Molly MacRae[5], Stephen Mobbs[7], Gwen Moncoiffe[4], Charlotte Pascoe[5], James Passmore[2], Helen Peat[3], Sam Pepler[5], Andy Shepherd[6], John Siddorn[4], Helen Snaith[4], Alex Tate[3], Poppy Townsend[5], Colm Walsh[4], John Watkins[1]

[1]UK Centre for Ecology and Hydrology, [2]British Geological Survey, [3]British Antarctic Survey, [4]National Oceanography Centre, [5]Science and Technology Facilities Council, [6]University of Manchester, [7]National Centre for Atmospheric Sciences

Background

The Environmental Data Service (EDS) Enhancement project (Phase 1b) builds on the developments of the EDS which have been undertaken during 2018-22 and will also be undertaken during 2023-28. The EDS developments aim to create a data service for NERC which, from the users' perspective, operates as a transparently single service, thus providing users of NERC managed data with coordinated data access, management and exploitation tools. A large part of the work required to build the future EDS involves creating the infrastructure to make data interoperable, both between NERC managed datasets and with other externally provided data.

Our resultant study on interoperability was organised as 4 complementary work packages as discussed below:

WP1: Building Interoperability - a NERC Data Commons RoadMap

The overall aim of this WP was to derive a roadmap for the implementation of a data/asset commons building on the foundation of EDS. This broke down into the following objectives:

i) to understand the requirements and structure for a data/asset commons for environmental science;
ii) to investigate what a commons architecture needs from the underlying EDS in terms of the required interface between the two;
iii) to appreciate the role of commons technology in building and extending a Community of Practice around environmental data.

The underlying hypothesis was that such an approach would enable integrative science across the NERC community and to other communities, building on an approach that intrinsically addresses FAIR Principles.

WP2: EDS Integration Experiments

Whilst WP1 explored the ideas needed to create a NERC data commons, leading to the delivery of an architecture for a commons together with a roadmap for its construction, this WP complemented this architectural work by focussing on a series of experiments intended to deepen our understanding of key areas of the commons architecture. There were two key reasons for doing this:

i) to understand the capabilities of candidate technologies that can support a commons approach;
ii) to inform the conceptual work in WP1 through practical experimentation, thus rooting WP1 in the practical challenges facing the EDS.

Specific experiments were carried out in five areas:

i) prototyping a narrow middle for a data commons;
ii) PID registries (specifically for instrumentation);
iii) catalogues, with a focus on SpatioTemporal Asset Catalogs (STAC), a community led common structure for describing and cataloguing spatiotemporal assets;
iv) semantic interoperability;
v) governance.

WP3: Scoping an EDS TRE

The ability to handle sensitive data, including social, economic and health data, is a key objective for the EDS community, including for the Digital Solutions Programme. This will permit an analysis of this data in conjunction with environmental data to understand the underlying influences and relationships, and better support decision making and positive outcomes across a range of sectors.

To underpin these objectives, it is necessary to understand how the TRE concept can be realised in practice, based on the JASMIN platform. This WP focussed on the scoping, designing and building of a pilot TRE on JASMIN to enable health data to be stored and analysed alongside environmental data.

WP4: Communications, Use Cases and Project Management

This WP implemented three EDS use cases:

i) FAAM (Facility for Airborne Atmospheric Measurements – https://www.faam.ac.uk;
ii) imagery and derived data from autonomous and remotely piloted vehicles;
iii) CMIP7 ontology development driven by community engagement.

The aim of these use cases was either to explore how the ideas and developments delivered in WP1 & 2 can be applied to real data or to trial user engagement strategies that will help the EDS broaden its user base and enhance accessibility.

The WP also provided overall project coordination, risk management, reporting and communications.

About this report
This report presents results of the project in the form of a series of recommendations and a roadmap emanating from the work. The report represents the key deliverable from WP1 in terms of focussing on the commons roadmap, drawing also on the integration experiments from WP2. In practice, we also drew on the work in the other two workpackages, in particular the results of the TRE developments (WP3) and also the three use cases from WP4.

The overall approach adopted in the project is shown in Figure 1.

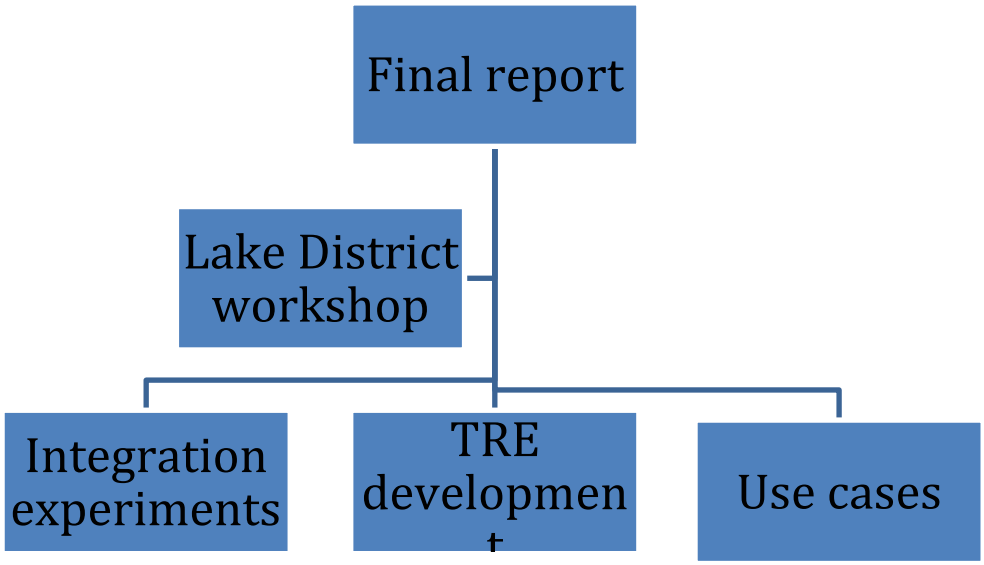


Figure 1. Structure of the research feeding into this report.

The overall report is a synthesis of the results of the various tasks from WPs 2, 3 and 4 also combined with insights gained from the workshop organised by WP1. Each of the tasks were asked to report their findings using a common template which captured:

1. The initial objectives of the work and the extent to which they have been achieved;
2. The internal and external collaborations that helped to produce the results;
3. A summary of the approach taken;
4. The key findings including reflections of what worked well, what did not work so well and key lessons learned;
5. A series of recommendations emanating from the work both for EDS developments and for the commons roadmap.

In some cases, multiple templates were filled out for a given task and the set of templates are included in Appendix A. A report on the workshop is also included in Appendix B.
A summary of all of these reports is given in Table 1.

Workpackage	Task	Template produced	Appendix
WP2: Integration experiments	Prototyping a narrow middle	Prototyping a narrow middle	A-1
		Object stores	A-2
	PID registries		A-3
	Catalogues (STAC)		A-4
	Semantic interoperability	Harmonised metadata	A-5
		iAdopt	A-6
	Governance		A-7
WP3: Scoping a TRE	TRE development		A-8
WP4: Use cases	FAAM		A-9
	Autonomous vehicles		A-10
	CMIP7 ontology development		A-11
WP1: Commons roadmap	Workshop summary		B

Table 1. Tasks per workpackage and associated templates produced.

Recommendations

The following recommendations were made from WP1 with a summary of recommendations contained in Table 2.

Recommendation number	Summary
1	Evolve towards a commons architecture, building on the strong legacy of the existing data centres
2	Plan a federated approach from the outset
3	Focus on a commons for a range of digital assets (not just data)
4	Strong emphasis on governance and associated mechanisms (e.g. communities of practice)

Table 2. Summary of recommendations.

The central recommendation (***recommendation 1***) emerging from this work is that the EDS should evolve towards a commons architecture, building on the strong legacy of the existing data centres and the steps already taken around alignment and integration. This is important to ensure we deliver against each of the FAIR Principles as we embrace the Fourth Paradigm of science (that is data intensive scientific endeavour [1] powered by advanced techniques in data science and AI). This is also important as we respond to the

complexities of dealing with a changing climate, requiring more integrative and systemic science. Following this Enhancement project, we are confident that this is the right step for the future of the EDS. We are also confident we understand the key building blocks following the work carried out in the integration experiments on the narrow middle approach coupled single sign on for authentication and subsequent authorisation, cataloguing systems, standardised persistent identifiers and APIs, all underpinned by common semantic frameworks.

We further recommend from the outset we plan for a federated approach (***recommendation 2***), allowing us to bridge between different domains. We suggest federation should not be an afterthought but intrinsically built into the commons approach (cf. federation by design). With this, we should strive for FAIRness across federated systems, that is supporting federated search, enabling access and interoperability across federated systems, and enabling reuse across domains. We also suggest federation should apply at multiple levels including;UKRI, s(across Digital Research Infrastructure initiatives), EDS (across the different data centres), and within data centres across different scientific disciplines. This approach recognises and embraces the inevitable heterogeneity of approaches and standards across the scientific community, including the different levels of maturity in terms of FAIR. It puts in place bridges between them in what is effectively a system of systems approach.
Building on this, we recommend that we do not just focus on a data commons but take the opportunity to develop a commons for a range of digital assets (***recommendation 3***) including methods, models, pipelines and workflows, notebooks, portals, etc. We strongly suggest an extensible approach starting off with a core set of digital assets with this set being extended over time, noting the strong benefits reported for having data and methods together in a commons (Appendix A-7). Further research will be required to support the full range of asset types, e.g., to support streaming data, multi-dimensional model output, drone imagery and so on, but also noting the inroads we are already making into many of these areas.

Finally, we recommend that a future commons initiative should place strong emphasis on governance and associated mechanisms such as developing and maintaining associated communities of practice (***recommendation 4***). In keeping with commons philosophies, this should not be a top-down form of governance. This would not work given the diversity in underlying communities. This governance should include reaching agreement on underlying standards, interfaces, ontologies and vocabularies and mechanisms to ensure compliance with these community agreements. The EDS through te Data Operations Group already has groups working on most of these areas and we can build on these collaborations to build community agreements.

This would require investment (see below) and it is not possible to deliver this vision within existing EDS budgets (recommissioning budgets and additional enhancement opportunities). We note the investments in other areas, e.g., BioFAIR (£34 million over 5 years), Smart Data Research UK (SDR UK) (£59m over the next 7 years), FDRI (£13m over 3 years, £38m total). We also note the opportunities to work across such initiatives to achieve federation across domains and to share best practices. Further opportunities include bringing together environmental data with health, socio-economic systems, omics, and indeed a range of other exciting permutations.

The roadmap

We are committed to deliver against this vision of an asset commons, offering a step change in EDS capabilities. We view this as a 3-phase process as illustrated in Figure 2.

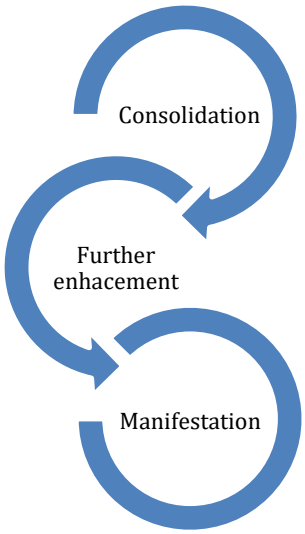


Figure 2. Planned phases in delivering an EDS commons.

The first phase involves **consolidation**, building on the results of the EDS Enhancement Phase 1b Project, as reported in this document. There are many dimensions to this work, and it is important to enter into a period of reflection and discern the key outcomes and disseminate results to the broader EDS community. We note that this process is already underway with the completion of the templates in Appendix A and the launch of the seminar series on NERC Environmental Data Service (EDS) Futures. We note there is already an EDS Roadmap produced as part of the 2023 recommissioning process and one concrete step is to ensure the key results of the enhancement project are reflected in this roadmap, producing an update accordingly.

We particularly need to take on board findings on:

1. The findings of the cross-EDS working group on harmonised meta-data and the associated investigations of the iAdopt framework;
2. The work on TREs and in particular how we can achieve federation between sensitive data and other data sources;

3. The insights and experiences around catalogue systems and indeed the role of semantics in enhancing semantic discovery;
4. The insights and experiences around the role of object stores and associated cloud-based services to support a range of complex data types, including multi-dimensional model output;
5. The knowledge gained from our studies of instrumentation PIDs;
6. The experiences across the consortium of other new data types, including from autonomous vehicles.

This phase has started and will continue to March 2024, overlapping with the start of the next phase, focussing on **further enhancements**, to include research into:

1. A service and tools for generating and managing Persistent IDentifiers (PIDs) for instruments and data with graphs of provenance, aligning with Research Data Alliance (RDA) working group activity in complex citations and instrument identifiers.
2. The development of standardised TRE services (building on DRI Phase 1b) project to enable integration of sensitive data with environmental datasets within JASMIN (including working with DARE UK).
3. Working across UKRI to enable federation across different domains with a particular focus on bringing together genomic and environmental datasets as a proof of concept;
4. Support for streaming data including real-time or near real-time data, from data acquisition, through quality assurance to ingestion;
5. The bringing together of DataLabs with commons concepts to realise the vision of a collaborative commons;
6. Further development around authentication and authorisation including support for single sign on across EDS services.

It is also important to carry forward key elements from Phase 1b, most notably the cross-EDS working group on harmonised meta-data given the importance of semantic web principles in underpinning our drive to FAIRness.

Note that at the time of writing, we are currently seeking funding to support these activities. This second phase will ramp up in January 2024 and will continue for 18 months. This will then lead to the third phase, the **manifestation** of our vision of an asset commons for the Environmental Sciences, supporting the federated model we seek. This will involve discussions with NERC and the broader UKRI community and the development of a proposal to take this work forward, including detailed costings of our plan. Again, there will be an overall between phases, with this work starting in January 2025 with a planned delivery by the end of that year (subject to discussions with NERC/UKRI).

[1] Hey, T., Tansley, T., Tolle, K., Gray, J. (2009). The Fourth Paradigm: Data-Intensive Scientific Discovery, Published by Microsoft Research, ISBN: 978-0-9825442-0-4.

Appendices

Work package	Task	Template produced	Appendix
WP2: Integration experiments	Prototyping a narrow middle	Prototyping a narrow middle	A-1
		Object stores	A-2
	PID registries		A-3
	Catalogues (STAC)		A-4
	Semantic interoperability	Harmonised metadata	A-5
		iAdopt	A-6
	Governance		A-7
WP3: Scoping a TRE	TRE development		A-8
WP4: Use cases	FAAM		A-9
	Autonomous vehicles		A-10
	CMIP7 ontology development		A-11
WP1: Commons roadmap	Workshop summary		B

A-1

Prototyping a ‘narrow middle’ for a data commons

List of initial objectives:

1. To carry out a ‘deep dive’ into the ‘narrow middle’ philosophy advocated by Grossman [1] to determine the applicability of the approach for the Environmental Data Service (EDS);
2. To experiment with the core services underpinning a narrow middle approach, namely registration and management of permanent IDs, authentication and authorisation, cataloguing of assets, and supporting access to services through open APIs.
3. To reflect on how these elements work together at the heart of a commons architecture to deliver FAIR digital assets (data and beyond).

To what extent have the objectives been realised:

We have been able to meet all our objectives, making strong progress on understanding the potential of an asset commons as a key EDS Enhancement, and the benefits of a narrow middle approach. We also have made strong inroads into understanding the elements of a narrow middle and technologies that can support this approach. We have made less progress of authentication and authorisation and recognise this as an area of importance going forward, for example around single sign-on approaches and associated authorisation schemes. To balance this, we have managed to do extra work on the broader DRI landscape, including the benefits of combining commons approaches with the previous work on DataLabs to form a collaborative commons, and also on the broader context around systemic approaches to environmental science and the need for skills development [2].

Collaborations:

Internal to the project:

- With individual tasks across the project, most notably around PIDs (Appendix A-3), cataloguing (Appendix A-4), governance (Appendix A-7), semantic interoperability (Appendices A-5, A-6).
- With the whole project, most notably through the EDS Enhancement Project Workshop: Building a RoadMap for a NERC Data Commons, 25th-26th May, 2023, Lake District, UK.

Externally:

- UKRI Digital Research Infrastructure community (https://www.ukri.org/what-we-do/creating-world-class-research-and-innovation-infrastructure/digital-research-infrastructure/),
- Elixir (https://elixir-europe.org/), Elixir-UK (https://elixiruknode.org/) and BioFAIR (https://biofair.uk/),

- Australian Research Data Commons (ARDC) (https://ardc.edu.au/), incl. EcoCommons (https://www.ecocommons.org.au/),
- With the IMFe and P-IMFe projects looking at an Information Management Framework for environmental digital twins, esp. around commons architectures and cataloguing,
- With the UKCEH-led Floods and Droughts Research Infrastructure project (FDRI) (https://www.ceh.ac.uk/our-science/projects/floods-and-droughts-research-infrastructure-fdri), feeding into their underlying commons-based DRI.

Summary of approach

(summarise how you have gone about the research, methods used, etc,):
The approach has been multi-faceted involving the following key elements:

1. We have taken input from literature on commons approaches, including the literature around narrow middle approaches, as well as taking input from projects in the UK and internationally that are developing commons-based technologies (see list of key external collaborations above).
2. We have carried out specific experiments and developments with underlying technologies, most notably around cataloguing and Open APIs, also taking input from other tasks within WP2 (on PIDs, semantic interoperability, the additional cataloguing work on STAC, and governance).
3. We arranged a key 2-day workshop involving the whole project team on Building a RoadMap for a NERC Data Commons, to bring together the various elements of the project and inform a common vision and associated roadmap to deliver a NERC data commons within the framework of the NERC EDS. The report on this workshop can be found in Appendix B.
4. We aligned our work with other activities in UKCEH around Digital Research Infrastructure (DRI) Futures. This involves a broader look at DRI requirements going forward to support more systemic, integrative and collaborative science involving a survey, interviews and workshops around the four UKCEH sites, with early insights from this feeding into the EDS Enhancement project (noting that DRI Futures continues beyond the timescale of the Phase 1b timescale and hence analyses are not yet completed).

The focus on the work has been on understanding the role of commons technologies in enhancing the EDS Service. Commons technologies are being widely deployed in a variety of domains as a response to delivery of the FAIR principles for scientific data management, and to encourage and enable open science. For the purpose of this project, we define a 'commons' as follows:

"A commons is a common place supporting asset discovery, access, interoperability and asset re-use (cf. the FAIR Principles), tailored for a community and managed by that community for the common good.[2]"

This definition extends data commons concepts to a broader asset commons, ensuring the FAIR use of scientific advances and discovery across all aspects of environmental research. It also emphasises the importance of community in commons-based approaches, and the resultant community-led approach is important to avoid the imposition of inappropriate standards and technologies across the commons. Rather, communities are responsible for determining the right approaches for that community and for implementing an appropriate homegrown governance model (see report on 'Governance' <link>. As mentioned above, we have particularly focussed on Grossman's 'narrow middle' architecture pattern, whereby the fewest possible core services are identified and applying standardisation to these core services is prioritised [1]. Grossman identifies four services as part of this core, namely the identification of assets through permanent digital identifiers (PIDs), a service associating metadata with such assets, authentication and authorisation services, and data model services for defining data models and for querying data with respect to this data model. This architectural pattern offers a pragmatic and minimalist approach to standardisation, focusing on this core and enabling innovation and diversity at the end points including in the development of tools for data input, curation, analysis and visualisation (Figure 1). Using this as a commons implementation offers simplicity of standardisation across assets, potentially enabling a federated commons to be possible for cross-community collaborations and systemic science. We return to this important point below.

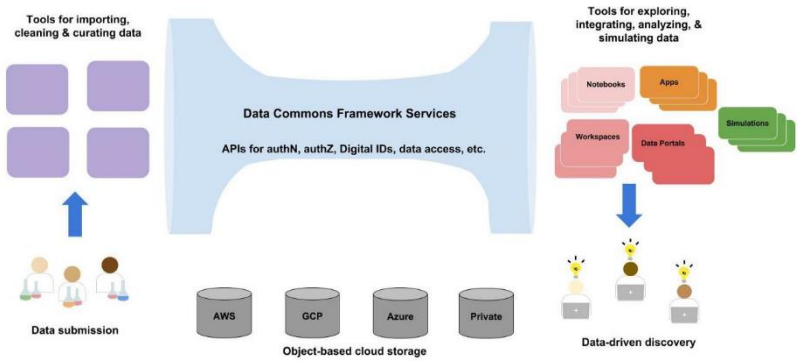


Figure 1. Grossman's narrow middle architectural pattern (from [1]).

What outputs have been produced (prototypes, reports, papers):

As well as contributions to this report, we have the following reports, presentations and events:

1. A report on the 'Building a RoadMap for a NERC Data Commons' available as an appendix to this report (Appendix B).
2. A webinar on 'NERC Environmental Data Service (EDS) Futures: The Role of Commons Technology in enhancing FAIRness' delivered by Prof. Gordon Blair (UKCEH) on 6th

November, 2023 as part of a series of talks highlighting the results of this project to the broader NERC EDS community (organised by Helen Peat, BAS).

3. A draft paper highlighting our broader DRI Futures agenda: Kelly Widdicks, K., Samreen, F., Blair, G.S., Rennie, S., Watkins, J. (2023). Digital Research Infrastructure (DRI) Futures for Environmental Science: Principles and approaches for realising the opportunities of DRI for systemic research, in preparation.
4. Running a 'World Café' session on 'Federation across domains' at the Elixir-UK All-Hands meeting, Norwich, 7th-8th November 2023, looking at how we can achieve interoperability across different domains including environmental and life sciences and data.

In addition, we have developed our software base in key areas in preparation for developing commons capabilities around the EDS:

1. Extensions to the EDS cataloguing system to enhance discoverability and the range of assets supported, specifically: i) authoring and implementing best practice guidance on discovery metadata, in close collaboration with other NERC data centres emanating from this project (Appendix A-5); ii) extending the range of digital assets included in the cataloguing system to include data, models; iii) developing a research data graph that will enable us to link and explore relationships between datasets, models, publications, and other research output. Note that this was made easier by the software already having a flexible and extensible underlying software architecture due to its use of the Java-based Spring Framework (https://spring.io/projects/spring-framework).
2. The implementation of a specific, tailored instance of the catalogue to underpin the P-IMFe platform, hence promoting FAIRness in digital twin construction.
3. Key extensions to DataLabs to support better access to key EIDC datasets for within the collaborative environment offered by DataLabs, this enabling integrative and collaborative science;
4. Experimentation with Open API technologies, most notably the OGC standard in this area (<link?>), a particular focus on large and heterogeneous datasets and multi-dimensional model output;
5. Experimentation with underlying cloud native architectural principles and associated cloud native tools such as object storage and tools to access or process potentially multi-dimensional datasets (e.g. Xarray).

We are also working on additional developments under the auspices of our DRI Futures project including i) extensions to DataLabs to support the collaborative commons approach and supporting different user journeys through the DRI building on UX design perspectives; ii) consideration of additional asset types, including streaming data, data science and AI methods and software pipelines/workflows.

Findings:

What are your overall reflections from the work (what worked, what did not work, overall lessons learned):

What worked well:

1. The focus on commons technologies and in particular narrow middle based approaches proved to be a success, with the narrow middle pattern having a number of key benefits offering, i) a pragmatic approach to dealing with the plethora of standards related to data management; ii) a flexible approach in dealing with a diverse range of domains and their inevitable differences in standards used; iii) an evolutionary approach that can grow over time and support innovation at the edges.
2. There is an elegant evolutionary pathway for the NERC EDS, with the intrinsic separation of concerns between the publishing and sustainable stewardship of data and other digital assets, from enhanced support offered by a broader commons - for discovering these assets, supporting more standardised access to the assets, interoperability between assets and re-use (that is enhanced FAIRness).
3. The generalisability inherent in an asset commons is important (cf. a data commons), with the combination of data with methods that can carry out a range of functions of that data (data wrangling, quality assurance, analyses, visualisation, etc) being particular powerful – as reported by some of our collaborators (ARDC, BioFAIR).
4. Governance is crucial to the realisation of a commons approach and we focussed extensively on this topic as the project developed (see associated report on our governance work).

What did not work:

1. Federation emerged as an important topic latterly in in the project and in retrospect we should have focussed on this at the outset to ensure that support for federation is intrinsic to the technological approach and the associated governance.
2. We should also have looked more carefully at authentication and authorisation and the identification of approaches that are robust and also compliant with a narrow middle approach.
3. Both federation and security hence remain important areas for future work.

Overall lessons learnt:

The most important lesson learned from this work is that collaboration is key, bringing together: i) all the data centres to develop an approach that embraces our similarities and differences; ii) all the skillsets and perspectives from across the centre as this is a cross-disciplinary challenge ; iii) all the sub-tasks from the EDS Enhancement project as a comprehensive approach requires input on PIDs, cataloguing, open APIs, semantic interoperability, trusted research environments and governance as well as the insights from other domains from emerging areas such as incorporating standards and approaches for drones.

What recommendations would you make for the next phase of EDS commissioning:
Our overarching recommendations from this task are as follows:

1. We strongly recommend that the NERC EDS embraces a commons approach to ensure enhanced delivery against the FAIR Principles;
2. This cannot be achieved from within existing EDS budgets but instead required additional investment (cf. the experiences in the life sciences community through BioFAIR);
3. A commons approach should be developed in close collaboration with other UKRI DRI initiatives to step towards a federated approach to FAIR data and other assets;
4. Federation is important at different levels – including within given NERC centres, across individual NERC data centres and across UKRI initiatives in this area;
5. Governance is also key and the technological approach should be developed in close collaboration with associated governance principles and approaches;
6. An EDS commons should be based on a narrow middle approach to achieve the benefits of this approach as reported above (supporting a pragmatic, flexible and evolutionary approach);
7. An EDS commons should also embrace a range of digital assets and not just be limited to data, for example to support emerging opportunities around digital twins;
8. The combination of data and methods is particularly powerful and should be prioritised in such a development;
9. A commons approach should build on the excellent support from existing EDS services in offering sustainable and trustworthy publishing of important environmental data set (cf. the CoreTrustSeal) ;
10. Overall, there is a real opportunity to achieve a step change in our support for FAIR digital assets and this requires close collaboration right across the community.

What recommendations would you make for the wider objective of developing a commons approach for environmental assets:

- As above

References

1. Grossman, R.L., 2018. A Proposed End-To-End Principle for Data Commons. *Medium.* https://medium.com/@rgrossman1/a-proposed-end-to-end-principle-for-data-commons-5872f2fa8a47, accessed August 2023.
2. Kelly Widdicks, K., Samreen, F., Blair, G.S., Rennie, S., Watkins, J. (2023). Digital Research Infrastructure (DRI) Futures for Environmental Science: Principles and approaches for realising the opportunities of DRI for systemic research, in preparation.

A-2

Enabling technologies for data centres in an EDS commons

List of initial objectives:

(1) Demonstrate the use of modern and scalable storage technology to effectively serve a variety of environmental data types from data centres to an EDS commons.
(2) Demonstrate the integration of this technology with NERC business tools (CEH DataLabs) to facilitate analysis, visualisation and understanding of environmental data.
(3) Demonstrate the use of NERC business tools to drive internal workflows at data centres.
(4) Understand the use of these enabling technologies across all EDS centres.

To what extent have the objectives been realised:

Our investigations showed that cloud-like technology, data object storage, can be an effective tool to serve a mixture of file formats to a commons and to integrate with NERC business tools (CEH DataLabs) that support the analysis of environmental data. Our investigations showed that this externally based interactive computing platform may not be suitable to drive internal workflows at a data centre due to software licensing and data security issues. Software tests were carried out at BODC, an EDS data centre with a traditional data management architecture. We were unable to fully understand the potential of such software at other EDS data centres due to the short timeframe of the project.

Collaborations:

Internal to the project: Work was done in collaboration with CEH to enable the use of CEH DataLabs.

Externally: As this work was software based, we did not work with anyone external to the project.

Summary of approach (summarise how you have gone about the research, methods used, etc,):

A commons is a digital platform that allows a community to effectively manage and share its assets, including its data assets, from a unified point-of-view. Thus, the technology that serves data from EDS data centres and facilitates the analysis of that data may play a significant role in the success of an EDS commons. Currently, gaining access to data from the NERC Data Catalogue is varied. Currently, NERC business tools that facilitate analysis of data, do not have direct access to the data held at EDS data centres. Therefore, this work looked to investigate more modern, scalable data storage technologies to serve a commons and its integration into a potential tool (CEH DataLabs[1]) of the commons, that will enable users to analyse, visualise and understand NERC data. The CEH DataLabs is a shared NERC interactive computing platform that can drive workflows through digital notebooks. Thus, we went on to investigate if it can be used to drive internal data management workflows at EDS data centres as a tool of an EDS commons. Our investigations were primarily based around cloud-like, data object storage technology as a means to efficiently serve data to a commons. Data object storage uses a flat file architecture, commonly stored in binary repositories, as opposed to traditional hierarchical file or block storage systems. This means files can be more efficiently accessed and the architecture can be easily scaled across distributed systems. As files are stored as addressable objects, this technology can serve unstructured data (e.g. PDFs, maps, images, plain text) as well as traditional structured scientific data files (e.g. NetCDF, csv, excel) through a unified end-point, allowing for both synchronous uploads and downloads as well as bidirectional low-latency streaming. Object stores are layered by an Application Programming Interface (API) allowing machines to easily access files over the web by any user or service for download or upload, depending on permissions. Open-source, self-hostable solutions are available and we used the S3-compatible MinIO[2] object store for our investigations. Using existing infrastructure, a temporary MinIO object store was deployed at BODC, one of the EDS data centres that has multiple repositories of data with hierarchical file storage delivering varied file types. To check suitability for other domains, we also included geological sediment images and extensible markup files from the British Ocean Sediment Core Research Facility (BOSCORF) situated at the National Oceanography Centre. MinIO was deployed in a Docker container to a general purpose host using Docker Compose. Network configurations were also modified to enable external access from DataLabs users with the correct authentication credentials.

What outputs have been produced (prototypes, reports, papers):

A temporary MinIO data object store was deployed at BODC for testing with a small amount of data. It is not considered production-ready or to be maintained in the long-term.

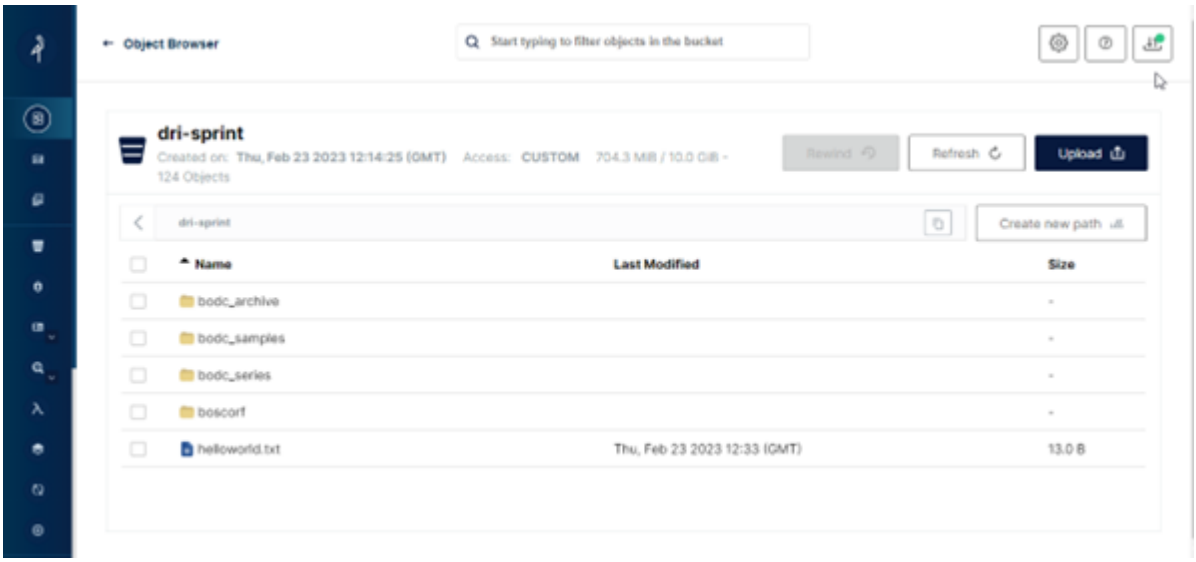
Findings:

Our investigations showed that cloud-like technology, data object storage, can be an effective tool to serve a mixture of file formats to a commons. As seen in fig. 1, we created a 'bucket' (in this case 'dri-sprint') of different data formats that included:

- Source data (.xlsx) provided in a data submission and added to the BODC Archive
- Discrete water sample data (.csv) ingested through the BODC Samples Schema
- In-situ sensor data (NetCDF-based) ingested through the BODC Series Schema
- Unstructured images (.tif, .jpg) of scanned sediment cores from BOSCORF
- Metadata files (.xml) of scanned sediment cores from BOSCORF

In S3-compatible object storage, a 'bucket' refers to an arbitrary collection of data. In our example in fig. 1(a), the data objects are organised by prefixes (e.g. bodc_samples). This mimics the folder-like structure of traditional file systems even though objects are actually organised in a more efficient flat architecture - this approach has the advantage of presenting a familiar experience to end-users already comfortable with the use of common file explorers as provided by Windows, MacOS, and most Unix-like operating systems.

(a)



(b)

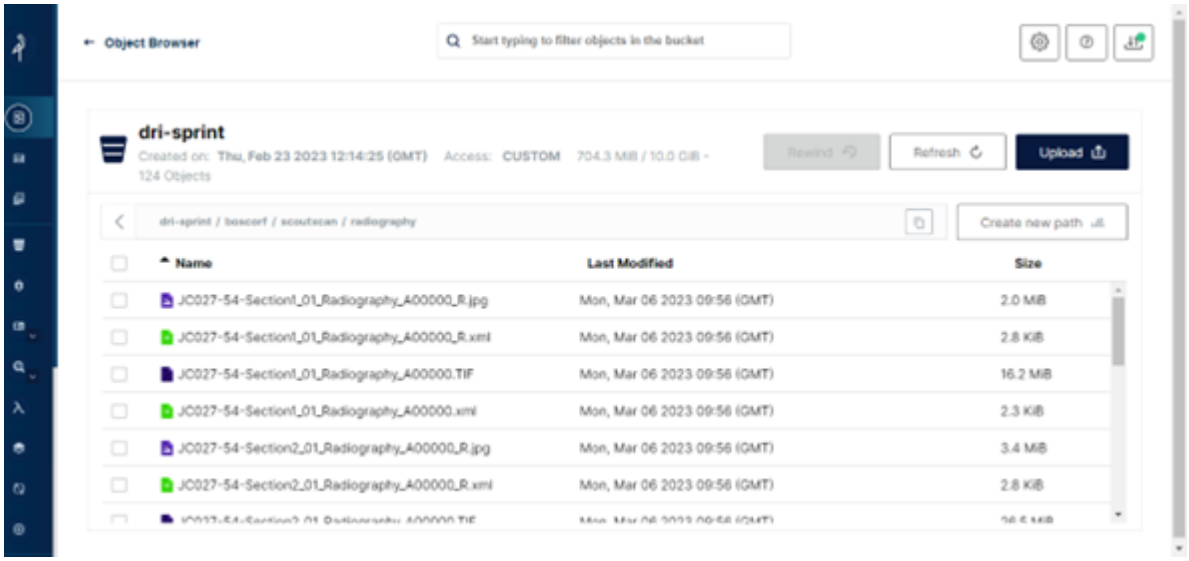


Figure 1. Examples of the GUI provided by the MinIO object store deployed at BODC. (a) Example of displaying a 'bucket' (in this case 'dri-sprint') containing different data formats from different data repositories and domains. (b) Example of listing structured (.xml) and unstructured (.TIF and .jpg) data from geological sediment cores in MinIO object store.

Each object, whether in a structured or unstructured format can be accessed through API end-points (table 1) depending on granted permissions. As shown, object storage offers the ability to access any arbitrary format of data, from different repositories as well as different EDS domains in a common and consistent way. A more production-ready deployment would also allow for automatic replication, the resolution of corrupted files and load-balancing across multiple geographic sites. This would potentially support effective delivery of data from data centres to an EDS commons. In addition, the use of the now industry-standard S3 API specification could act as standard interoperable across both EDS data centres as well as many other service and cloud providers.

Table 1. Example of object store API end-points for structured and unstructured data. Note that these URLs are not accessible to the general internet at this time.

Format	End-point url
Structured data	https://minio.bodc.ac.uk/dri-sprint/bodc_samples/1024981.csv

Unstructured data	https://minio.bodc.ac.uk/dri-sprint/boscorf/scoutxcan/laminography-slices/jc027-54-section1-Depth000_-45.mm.tif

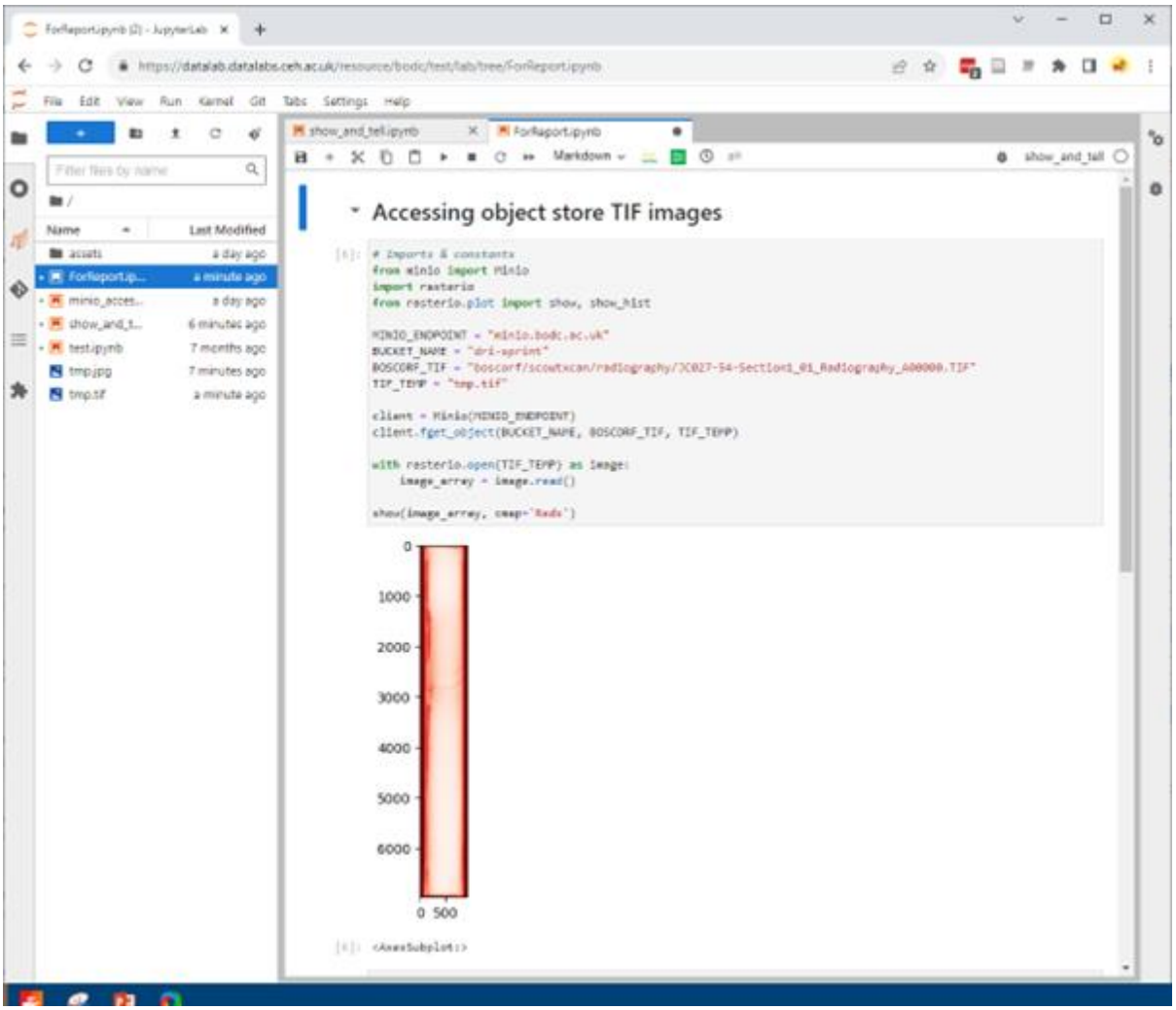


Figure 2. Visualising a sediment core scan (TIF) image in a CEH DataLabs JupyterLab notebook that was directly accessed from the BODC MinIO object store, without requiring an intermediate download step from the user.

Our investigations showed it was possible to integrate data objects held at an EDS data centre with NERC business tools (CEH DataLabs) that could support the analysis of environmental data in a commons. This could be achieved by enabling our object store with an IP range for the DataLabs application and the officially supported first-party open source *minio*[3] Python client. This could also be achieved using the *s3fs*[4] python library, which mimics local file operations to better enable streaming data for use by existing tooling. Through the interactive digital

notebooks hosted by DataLabs, we were able to visualise (fig. 2) and analyse (fig. 3) Tagged Image Files (TIF) of sediment cores directly from the MinIO object store. We were also able to render one of BODC's internal data storage formats, QXF, a NetCDF-based format specific to BODC (fig. 4). Thus, object storage may offer a way to serve the assets of EDS data centre to a potential tool of the commons, that will enable users to analyse, visualise and understand environmental data in a NERC digital ecosystem without the additional complexity and overhead of manually synchronising data between sites.

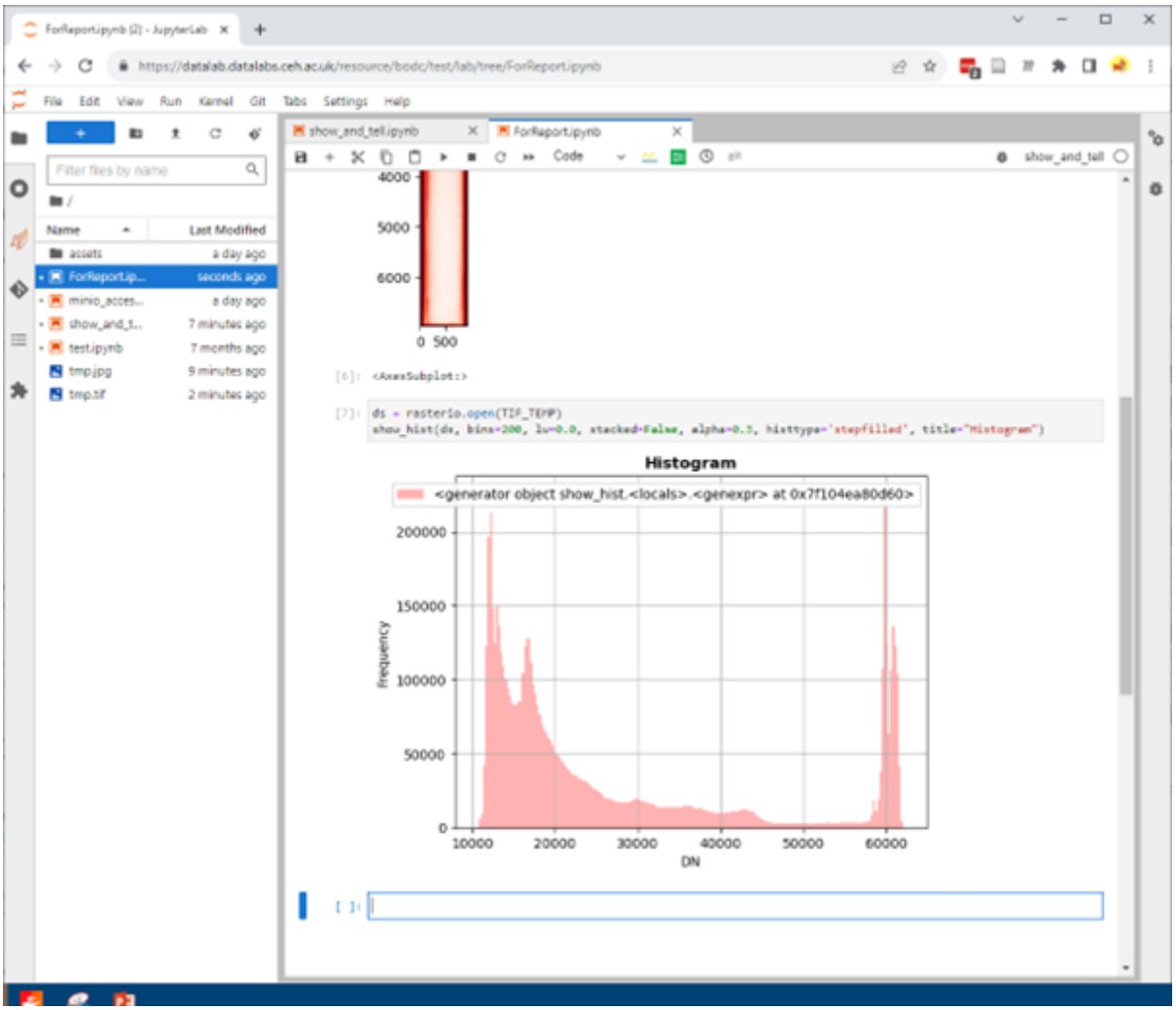


Figure 3. A histogram of pixels in a sediment core scan (TIF) image in a CEH DataLabs JupyterLab notebook that was directly accessed from the BODC MinIO object store.

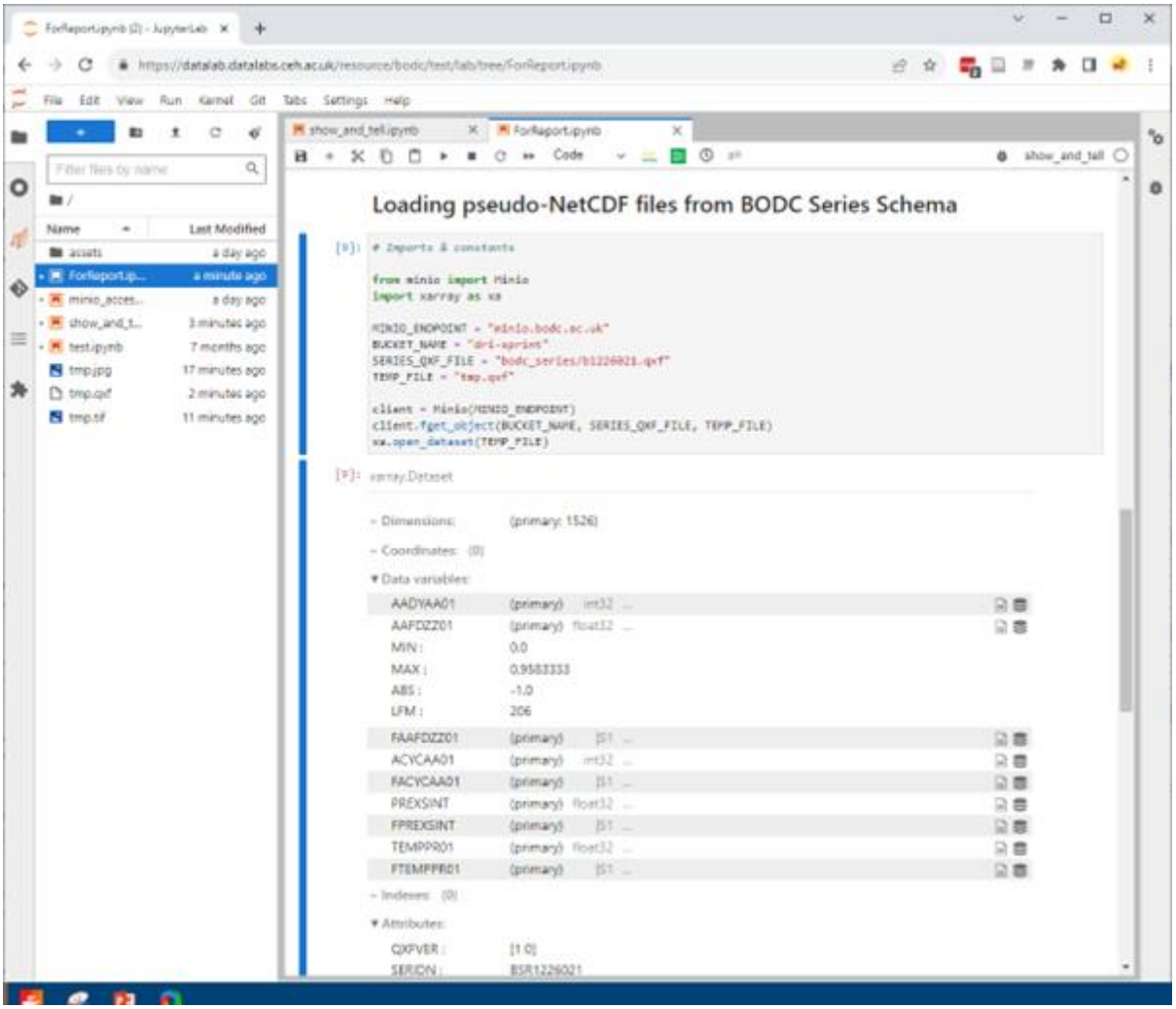


Figure 4. Rendering data from a BODC internal storage format file (pseudo-NetCDF) in a CEH DataLabs JupyterLab notebook that was directly accessed from the BODC MinIO object store.

Digital notebooks like those hosted by the CEH DataLabs can be used to describe complex workflows that can be executed in an ordered and repeatable fashion, making these workflows more efficient. For example, the routine steps undertaken by data managers to ingest and expose data at EDS data centres. The central CEH DataLabs application may offer a low maintenance way for data centres to distribute and execute workflows for an EDS commons. However, secure access to internal systems and licensing of third-party software may present an issue. BODC, for example, while possessing a significant amount of bespoke generic, platform-agnostic tooling, also makes heavy use of in-house MATLAB-based software that depends on both data and metadata served by an internal database, something that is currently unsupported by DataLabs. Alternatives, such as pre-compiling and copying software to the DataLabs environment, or developing numerous web APIs around said software are possible,

but would present a major investment in developer time. Additionally, any such software (regardless of its underlying technology) that is made available on DataLabs must then be able to access filesystems, databases, and other potentially sensitive facilities internal to the data centre in question. This obstacle would be alleviated somewhat if the data centre were to employ S3-compatible storage (such as MinIO) as the standard mechanism for storing and internally serving data, as any internal software written would be designed with support for this technology by default.

What are your overall reflections from the work (what worked, what did not work, overall lessons learned):

What worked well:

The deployment of the MinIO object store in a Docker container was relatively straightforward. However, this investigation was carried out with a small subset of data. Fully deploying an object store will require careful planning around enabling legacy data, scalability and server orchestration, as well as considerations for support from pre-existing software, much of which would require some (generally routine) retrofitting.

What did not work:

Due to the short time frame of the project we were unable to fully understand the potential, limitations or current use of object stores in other EDS data centres, for example, CEDA who utilise the high-performance computing facility, JASMIN, which hosts a S3-compatible object store. Furthermore, MinIO's metadata capabilities are somewhat limited in comparison to what can be provided by purpose-built schemas, which while sufficient for administrative purposes such as access permissions and versioning, would not be appropriate for discovery or descriptive metadata and further investigation is needed.

Overall lessons learnt:

Our initial investigation into object storage suggests this architecture is a tangible enabling technology for serving data directly from data centres to an EDS commons and the potential analytical tools that may be used. In addition, it may also offer advantages for purely internal usage. However, this was only a small investigation. Our tests with sediment cores and a

mixture of structured and unstructured data suggests data object storage may be flexible enough for a variety of data held at other EDS data centres. MinIO can theoretically perform with high volume objects (up to 50 TiB) without issue but this will depend on the capabilities of backend hardware. Some investigation with the CEDA JASMIN High-Performance Object Store will also be needed where high data volumes are limiting. It is also important to consider that for internal processes, bandwidth requirements for using an external or cloud-based object storage service could become prohibitive as well as the costs of data egress (download) from these stores.

What recommendations would you make for the next phase of EDS commissioning:

There is a desire to support automated data ingestion through the development of new and common software, infrastructure and tools within the Acquisition and Ingestion (A&I) theme of the EDS roadmap. This not only applies to near real-time data but to delayed-mode data from research scientists. Data object storage offers a common foundation to automatically manage and access data across the EDS data centres, giving greater flexibility to those who provide and use NERC environmental data. Furthermore, the integration of these modern storage technologies with new, powerful solutions (such as Pachyderm[5]), that automate complex data transformation pipelines and integrate machine learning may transform EDS data management. However, fully deploying a comprehensive object storage solution will require careful planning around enabling/upgrading legacy data and systems, scalability and server orchestration which may bear costs. A cost analysis against legacy data management infrastructure may also be beneficial.

What recommendations would you make for the wider objective of developing a commons approach for environmental assets:

A commons is a digital platform that allows a community to effectively manage and share its assets, including its data assets, from a unified point-of-view. Thus, the technology that serves data from EDS data centres and facilitates the analysis of that data may play a significant role in the success of an EDS commons. Currently, gaining access to data from the NERC Data Catalogue via *Gemini* records is varied. Modern, scalable technology like data object storage has the potential to effectively and consistently deliver data from EDS data centres to a commons in a variety of formats and volumes. Indeed, a more production-ready deployment would also

allow for automatic replication, the resolution of corrupted files and load-balancing across multiple geographic sites. However, further investigation will be needed to understand the potential, limitations or current use of object stores in other EDS data centres. Furthermore, we were not able to consider the impact on accompanying metadata which is limited with each object in a MinIO store. As we enter a more digital world, sensing techniques and technologies that generate more unstructured data are expected to rise. In parallel, we might expect a proliferation of discovery and descriptive metadata. Thus consideration to flexible, scalable metadata solutions (such as NoSQL document stores, etc.) might also be needed.

A-3

Developing a framework for instrument identification and use across the NERC Environmental Data Service (EDS) and digital data ecosystems

List of initial objectives:

(1) Deliver a recommendation for a new instrument persistent identifiers (PIDs) registration service building on the Research Data Alliance (RDA) Persistent Identifiers for Instruments (PIDINST) framework[1,2] to enable integration of instrument provenance across the Environmental Data Service (EDS) and other domains, broadening the EDS user base.

To what extent have the objectives been realised: We have achieved this aim through research, the development and collaboration of two online surveys and a recommendation for a new EDS instrument PID service.

Collaborations:

Internal to the project: An online survey to gather requirements for a new instrument PID service was developed with representatives from different data centres in the EDS.

Externally: A landscaping survey of UK sensor networks and sensor data (and in turn, the landscape of instrument PIDs) was developed in a collaboration between this project and a project supported by the NERC Constructing a Digital Environment (CDE) involving senior expert members from the CDE network. The survey to gather requirements for a new instrument PID service was also developed with support in kind from the RDA PIDINST community.

Summary of approach (summarise how you have gone about the research, methods used, etc,): Addressing some of the environment's most challenging issues, requires the assimilation of data from multiple sources, over a variety of scales, resolutions and

frequencies. However, little is often known about the devices and operational settings used to generate these data, key information to effectively use the data. As we enter a more digitally-enabled world, the number of sensing and measuring devices used to observe our environment will rise, generating more data with increasing levels of sophistication and automation. Accurately analysing, modelling and simulating these data to gain new environmental insights and understanding will become more challenging as the volume, complexity and automation of environmental data grows. Thus, common ways to reliably identify, link and access information about the devices used to generate environmental data are required.

The PIDINST working group developed an international strategy to identify instruments through globally unique persistent identifiers (PIDs) that was endorsed by the RDA[3]. PIDs are particularly suitable for this purpose because they are long-lasting references to a digital resource, or a digital resource that represents a physical thing such as an environmental instrument. They are considered domain-agnostic and are particularly significant in an advancing digital world, identifying and connecting entities across different systems without ambiguity. They can also resolve descriptive information (metadata) about a resource, enabling systems to gather information efficiently. The idea of using identifiers to connect systems to the key information to effectively use the data from a device is not new. Between 1997 and 2010, the Institute of Electrical and Electronics Engineers (IEEE) proposed a suite of standards for developing SMART transducers in networked industrial environments, known as the IEEE 1451 family of standards[4]. A key feature is Transducer Electronic Data Sheet (TEDS), that allows sensors to act in a truly 'plug and play' fashion. TEDS contain key information for the transducer, such as manufacturer identifiers, capabilities, calibrations and interfacing requirements, the key information needed by connected networks to effectively use the signal from a transducer. It can exist as an external file, accessible to the network through a MAC identifier sent from the transducer. This makes it possible to decode data from legacy devices without affecting costs. In the EU SenseOCEAN project, BODC used universally unique identifiers (UUIDs) to resolve marine-specific profiles[5] of sensor metadata in a machine-readable format (Open Geospatial Consortium Sensor Web Enablement SensorML)[6]. This aimed to reduce transmission costs from remotely deployed sensors in the marine environment. However, these services have not made use of globally unique and persistent identifiers that can be used across broad sets of users, especially in environmental

settings where sensors might be quite distributed. Thus, the aim of this work was to investigate the feasibility of a new service that identifies instruments using PIDs, enabling the provenance, interoperability and sharing of contextual information across the EDS, the NERC Digital Ecosystem[7] and broader user communities.

The work was delivered in 3 phases. The first phase involved research into the capabilities of candidate technologies that can support an EDS instrument PID service. In particular; existing applications that either register or use instrument PIDs; schemes that use instrument identifiers (not necessarily instrument PIDs) to automatically access comprehensive instrument metadata; and common standards that improve the sharing of this information, especially by machines. The second phase involved the development of two anonymous online surveys (delivered through SurveyMonkey). An online survey specifically dedicated to gathering requirements for an EDS hosted instrument PID registration service[1] was developed primarily as part of this project and released in August 2023. Prior to this, questions aimed at understanding the prevalence of instrument identifiers and instrument PIDs were developed in collaboration with another online survey. This survey was designed to landscape current sensor networks and sensor data (and thus instrument PIDs) in the UK[2]. The survey was developed in collaboration with a NERC CDE mini demonstrator project investigating a new specialised EDS service to integrate and archive sensor data from NERC sensor networks for long-term use. Device management will be pivotal to this service. The survey was released in July 2023. Phases 1 and 2 were then used to inform a recommendation for an instrument PID registration service hosted by the EDS.

What outputs have been produced (prototypes, reports, papers):

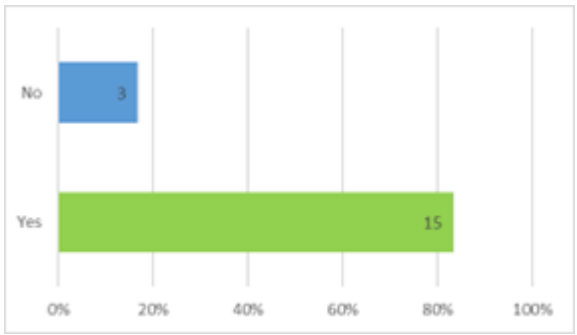
- An online survey to landscape instrument PIDs over a wide range of UK sensor networks, managers and users of sensor data as part of a collaboration with the CDE[8] (https://www.surveymonkey.co.uk/r/eds-sensors-survey)
- An online survey to gather requirements for an EDS instrument PID service (https://www.surveymonkey.co.uk/r/EDS-instrument-pid-survey)
- An online presentation of results (including the prevalence of instrument PIDs) from the landscaping survey (https://digitalenvironment.org/digital-environment-projects/#nerc_sensor_network_service).
- A recommendation for a new EDS instrument PID registry service (this report).

Findings:

To our understanding, this will be the first time a formal service for identifying instruments with the PIDINST framework has been proposed in the UK. This seems timely, as the adoption of instrument PIDs is continuing to gain momentum internationally. The European ACTRIS network is already ascribing PIDs for instruments. DataCite, a well-established international facility that connects research outputs and resources through PIDs, is providing support for instrument resources in their PID metadata schema 4.5[9,10]. The European collaborative data infrastructure, EUDAT, has launched the B2INST instrument PID service[11] that enables users to create instrument records and link documentation, though not necessarily in machine actionable ways. This report summarises the main findings from our surveys and conceptual design for a new instrument PID service to support provenance in an EDS commons, enhance analysis and understanding in the NERC Digital Ecosystem and evolve NERC services to wider communities.

Q. Do you manage or use information about the sensing devices used to generate the sensor data? (n=18)

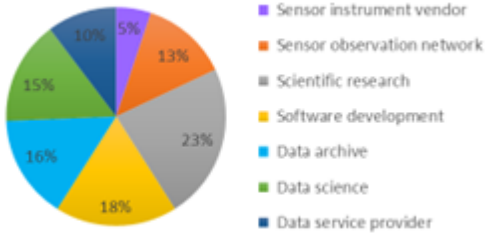
(a)



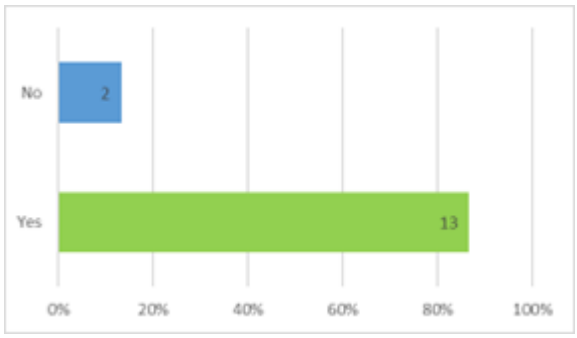
Q. Are unique identifiers used to identify any sensing devices? (n=15)

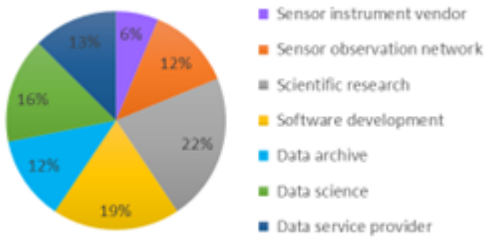
(b)

(d)



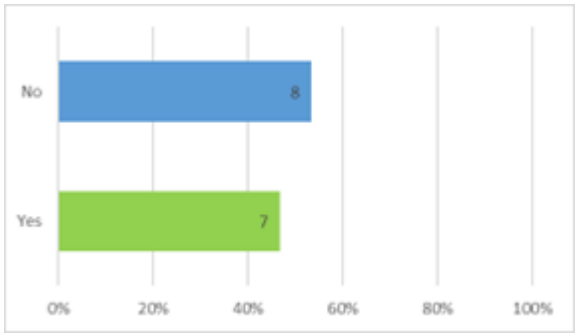
(e)





Q. Are devices assigned globally unique and persistent identifiers? (n=15)

(c)



(f)

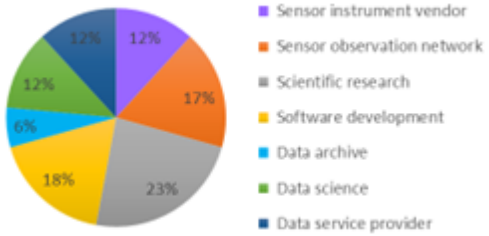


Figure 1. Results from the landscaping survey. Charts (a - c) show percentage responses to questions (number of responses inside bar). Charts (d - f) show demographics of 'yes' responses.

Survey results:

The landscaping survey was targeted at all actors in sensor lifecycle, in communities both internal and external to NERC. The landscaping survey had a total of 36 respondents, of which 15-18 answered the questions related to instrument identification and device management. A total of 9 respondents answered our PID requirements gathering survey. Given the smaller number of respondents, our PID gathering survey was more qualitative than quantitative.

The key findings were:

- **A majority (13 of 15) of respondents were using unique identifiers of some sort for sensing devices.**
- **Around half (7 of 15) were using sensors assigned with persistent identifiers.**
- **A majority (13 of 15) of respondents were capturing or using operational information about sensing devices.**

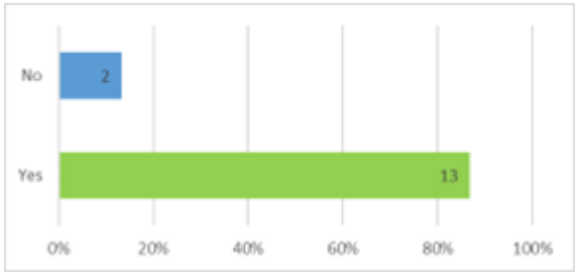
- **53-73% (n=15) of respondents are capturing or using dynamic operation information about sensing devices, such as the sensor configuration during an installation of the device.**

- There is indication that primary applications (n=9) for instrument PIDs include: the unique identification of instruments; linkage to research outputs; traceability of data; and enabling quality control (and/or assurance) of data.
- There is an indication that an instrument PID should be created early in the instrument lifecycle (n=7) potentially during procurement or at first installation of an instrument.

Figure 1 shows a breakdown of results from respondents who answered the questions related to instrument identification and device management in the landscaping survey. Just under half (15 of 36) of the total survey respondents were found to use or manage in general identification and even persistent identification of sensors were in operation (fig. 1b and 1c). Furthermore, this was practised by a wide demographic of respondents (fig. 1e and 2f) associated with the sensor data lifecycle.

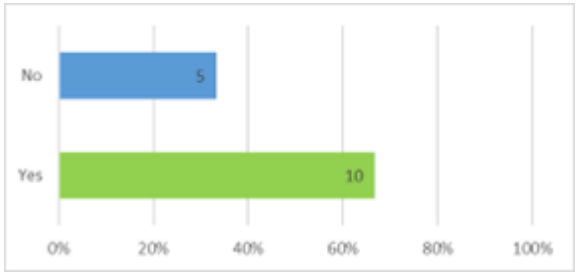
(a)

Q. Do you capture or use operational information about sensing devices? e.g. technical specifications, operating modes, calibrations, data sheets etc (n=15)



(b)

Q. Do you capture or use the purpose for the sensor installation? (n=15)



c)

Q. Do you capture or use the configuration of the sensor during installation? (n=15)

(d)

Q. Do you capture or use any information about the mounting platform used to install the sensor? (n=15)

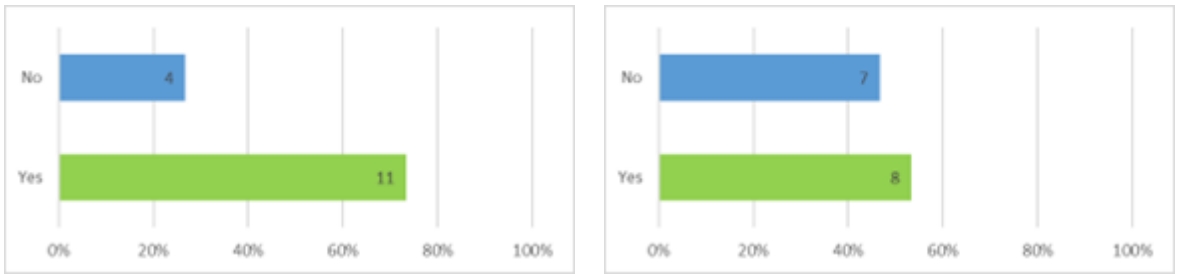
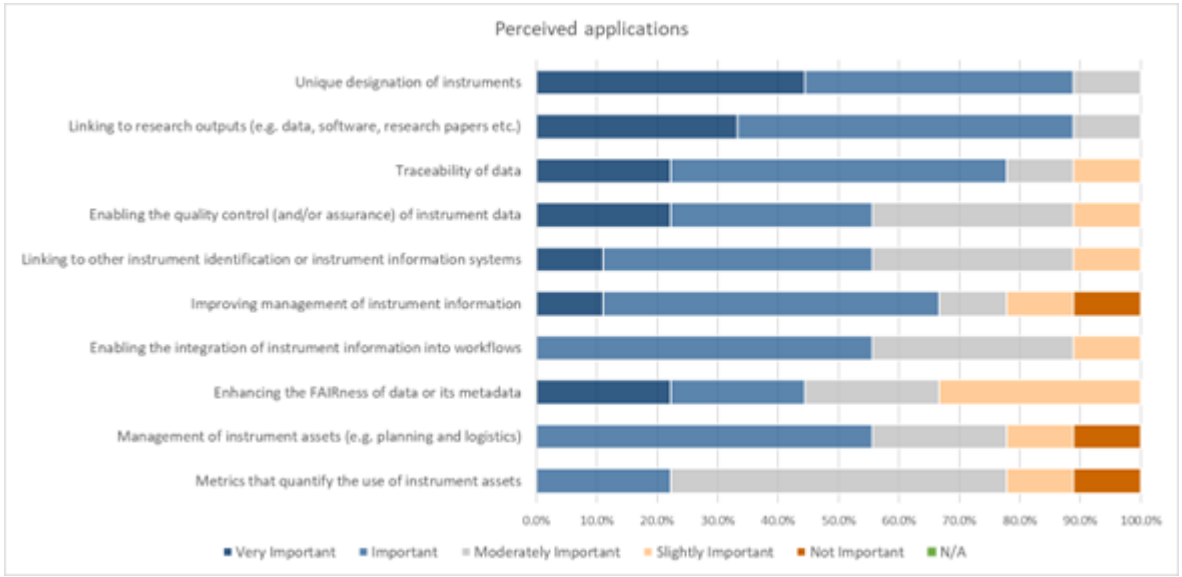


Figure 2. Shows responses to questions about the operational information of sensors from the landscaping survey.

Figure 2 shows the responses to questions related to more operational information about sensors (such as technical specifications) in the same section as instrument identifiers on the landscaping questionnaire. Like with the use of instrument identifiers, these results suggest that there is active management and end-use of operational instrument information (fig. 2a). This also includes more dynamic information, in other words, information which changes over time. For example, with each installation of a sensor (fig. 2b, 2c and 2d).

Q. What are the most important applications that you perceive instrument PIDs will be used for? What are the most important problems it should solve in your opinion?

(a)



(b)

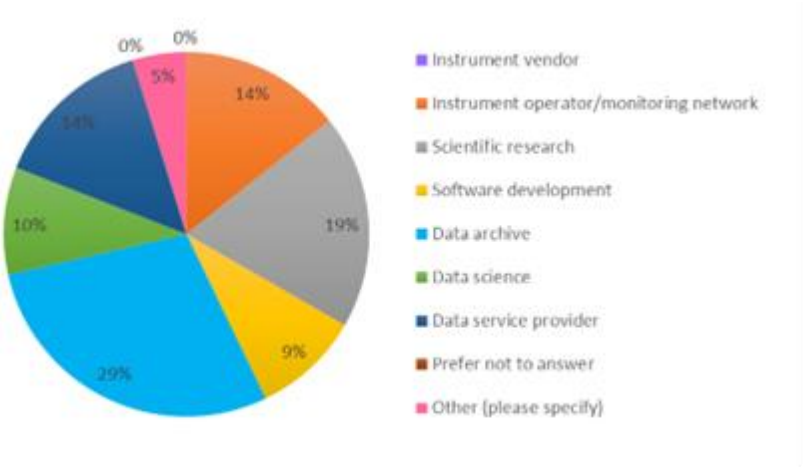


Figure 3. Applications perceived for instrument PIDs by respondents in the PID requirements gathering survey (n=9) where (a) visualises the importance of applications by respondents via Likert Scale and (b) visualises the demographic of respondents. Note results are considered indicative.

Using a Likert scale, our PID requirements gathering survey indicated that important applications for instrument PIDs were weighted towards; the unique designation of instruments; linking to research outputs; traceability of data and enabling the quality control and/or assurance of instrument data. The respondents to this question were from a broad demographic of actors in the lifecycle of an instrument (fig. 3b). However, given the number of respondents to this survey, this information should be considered more qualitative than quantitative.

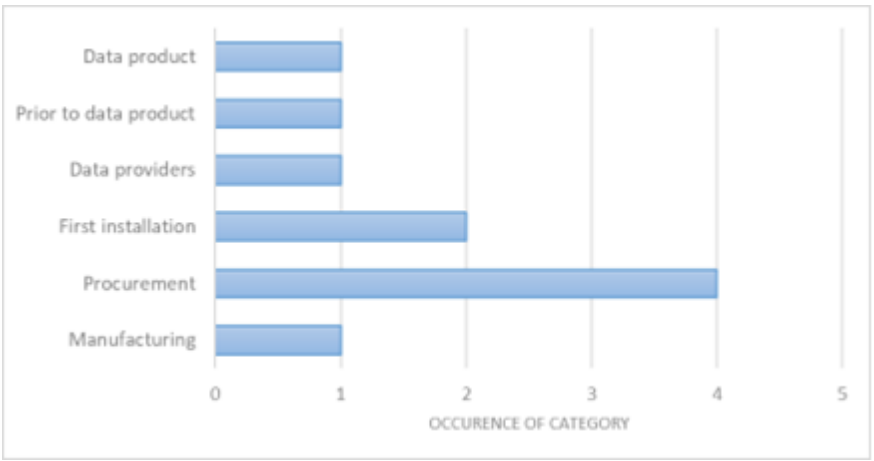


Figure 4. Occurrence of categories observed in the open-ended question 'Which event in the instrument lifetime would justify the creation of an identifier?' (n=7). Note results are considered indicative.

In our PID requirements gathering survey, we categorised comments to the open-ended question, 'Which event in the instrument lifetime would justify the creation of an identifier?'. Figure 4 shows the number of comments that fall into each category. Most of the comments related to the creation of an instrument identifier early in the instrument lifecycle, primarily during procurement or the first installation of the instrument. However, given the number of respondents to this survey, this information should be considered more qualitative.

Conceptual framework summary

Figure 5 outlines a conceptual design for a new EDS instrument PID registration service. It recommends a new instrument PID service that fulfils two key functions: a PID minting and resolution service for new instrument devices and an inventory of instrument PID assets, supporting PIDs issued by any authority in a commons approach.

Key recommendations:

- *A PID minting service for new instrument PIDs.*
- A resolution service that resolves to the key information needed by connecting systems to effectively use device data in digital ecosystems.
- An inventory of instruments that enables the provenance and interoperability of new PIDs minted by the EDS and existing PIDs from other issuing authorities in support of a commons approach.
- Value the expertise and knowledge of those providing instrument information to maximise the value of the system to enable it to evolve with different communities
- Limit barriers-to-entry through registration requirements that are maintained as low as possible, with providers being encouraged to gradually increase the information (and in turn, FAIRness of the information) they provide through training and support.
- Adopt user-defined Common Data Elements, building blocks that can be chosen or adapted to supply the most appropriate instrument information properties or profiles of properties

- Enable access to user-defined processing scripts to maximise the value of data when necessary
- Facilitate links to research outputs and enhance research metrics through the OpenAIRE Graph by registering our new service through the JISC OpenDOAR directory of repositories
- Deliver a series of pilot studies with representative EDS communities

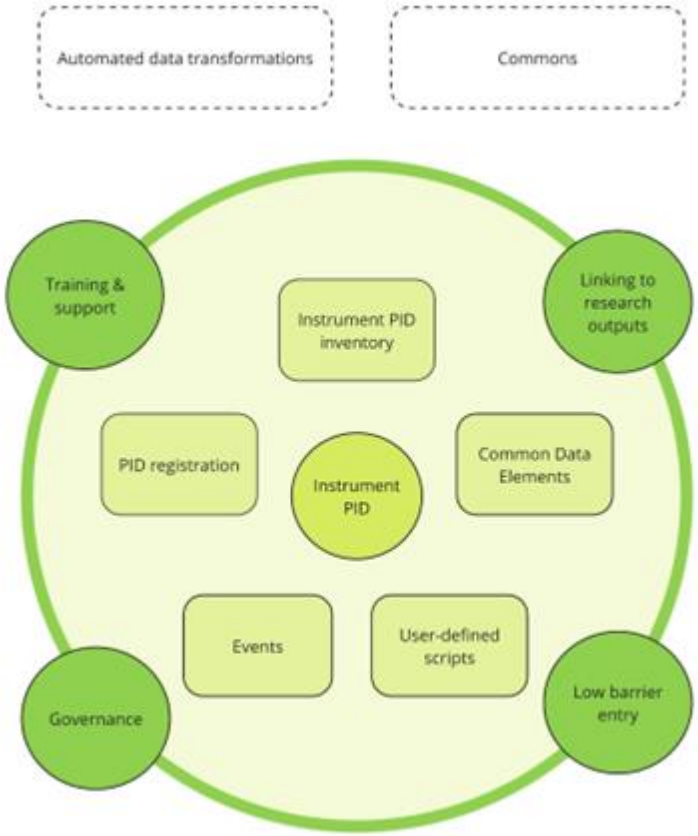


Figure 5. Overview of a conceptual framework for an instrument PID service.

Conceptual framework full recommendation

Figure 5 outlines a conceptual framework for a new instrument PID service as part of the EDS. Our results from the landscaping survey showed that the unique identification of sensors is actively used in the sensor landscape (fig. 1b) and over a wide demographic of actors related to the sensor data lifecycle (fig. 1e). Many of the same actors are involved in maximising the

value and use of environmental data in the NERC Digital Ecosystem, indicating there is potential to unambiguously identify instruments in this system using globally unique persistent identifiers, enabling system-wide connectivity. However, our results also show that existing PIDs for instruments are already in circulation (fig, 1c) by similar sets of actors (fig. 1f). This suggests that environmental devices in general may not always be registered (and in turn governed) by the EDS authority in the first instance, presenting a challenge should their identifiers be used within the EDS and an EDS commons. Thus we recommend a new instrument PID service that fulfils two key functions. Firstly, to act as a PID minting service for new instrument devices, including a resolution service to the key information needed by connecting systems to effectively use device data in digital ecosystems. Secondly, to provide an inventory of instruments, essentially an indexed catalogue of new PIDs minted by the EDS and existing PIDs from other authorities. These records will contain the essential metadata needed to support provenance and interoperability in a commons approach.

General approach

A key to the success of any EDS service will be its ability to work effectively with the differing requirements of each of its domains. In addition, it will need to evolve and adapt to new requirements outside of the EDS in order to attract broader and more specialised communities. Our results from the landscaping survey showed that just under half (15 of 36) of the total survey respondents were found to use or manage general information about sensors (fig. 1a). Furthermore, a majority (13 of 15) of these respondents were capturing or using more complex, operational information about sensing devices (fig. 2a). These results suggest there is a level of capability and expertise for managing and working with device information that already exists in the sensor landscape and potentially devices in general by extension. Some of this expertise will be highly specialised to certain devices and communities. Therefore, we recommend any new service values the expertise and knowledge of those providing instrument information to maximise the value of the system. Essentially, the service is flexible enough to allow providers (the experts) to select the most appropriate information to facilitate sharing with connected systems, which may not be the same for every instrument, application or community practice. To achieve this, we recommend a PID framework that is a hybrid of 'bottom-up' principles in addition to 'top-down' ones. It consists of 5 key components, two of which enable flexibility to differing

requirements. This includes user-defined metadata building blocks that can be chosen or adapted to supply the most appropriate instrument information. Furthermore, a library of custom-defined processing scripts to enable the uplifting of complex instrument data. Each of the 5 components are standalone services that are able to interact with each other as well as independently. This will enable the PID service to be used with existing external services or existing instrument PIDs.

In addition to flexibility, it must be easy for new users to engage with the system. Thus, we recommend limiting barriers-to-entry through registration requirements that are maintained as low as possible, with providers being encouraged to gradually increase the information (and in turn, FAIRness of the information) they provide through training and support. This strategy has also been adopted by the National Library of Medicine towards FAIR data submissions[12]. The six mandatory properties of the PIDINST schema for metadata registered with instrument PIDs[13] is considered appropriate to unambiguously identify instruments. This would formulate a low number of registration requirements. Furthermore, a mechanism to unambiguously identify instruments would support important applications indicated by our surveys highlighted by a wide set of actors in the instrument lifecycle (fig. 3). These include unique designation in external systems, linking between entities (e.g research outputs) and as well as providing provenance for data. Altogether, we believe these general approach recommendations will enable the service to work in harmony with existing information management systems and different community requirements.

Key components

PID registration: This component will involve the minting of new instrument PIDs by permissible users of the system. The service should allow humans as well as machines to create and update records as well as deduplicating[14] records, a process recommended by the PIDINST working group. Each PID should resolve to a useful state of information about its instrument via a resolving service as standard for resolving PIDs.

Common Data Elements: The National Cancer Institute Genomic Data Commons has successfully used Common Data Elements (CDE) to increase the interoperability of medical metadata and data across the service. These common elements are advantageous because they can be reused by others, or grouped into complex sets to form questionnaires making

them quite flexible across different users. They are defined unambiguously in both human and machine-computable terms, contributing to FAIR principles.

This building-block approach has the potential to also standardise instrument information which is shared with the EDS instrument PID service and subsequently resolved by a PID. In particular, operational instrument information (such as technical specifications) which are needed to quality control/assure instrument data, one of the important applications indicated by our survey results (fig. 3a). Instrument information can vary from simplistic properties, such as an instrument's serial no., to complex sets of properties, for example, a profile of a sensor's configuration settings, a process such as a calibration or a collection of attributes describing an observed property. Indeed, our survey results showed the active use of operational sensor information in the current sensor landscape.

In using such an approach, we suggest the following aspects are considered. Elements represent 'semantic snippets' to support FAIR principles and machine readability. Elements can be assigned persistent identifiers themselves to enable findability and versioning. Elements can be qualified by multiple controlled (machine-readable) terms of any authority, extending their use to different communities with specialised practices. Issuing authorities are defined as elements and are used alongside controlled terms in snippets. Elements can include links to external documents, e.g. machine-readable Digital Calibration Certificates[15], a global standard for instrument calibration that is currently in development. All elements for an instrument should resolve to a useful state of information about it via a resolving service. This information will need to be efficiently consumed by applications or workflows in a digital ecosystem if this information is to be used for automated analysis of data. Ideally the state of an instrument should resolve to a formal, accessible and broadly applicable language for knowledge representation following FAIR data principles[16].

Similar principles can be applied to some values of properties within CDEs that are shared with the instrument PID service, In some cases, property values themselves may be highly standardised to certain communities. We suggest standardised values are also accompanied by controlled terms (where applicable) as well as the issuing authority. Frameworks like the RDA recommended iADOPT[17] may also help constrain and harmonise observed properties by their component semantic properties. A similar strategy has been used for statistical variables in the Google Data Commons[18].

User-defined scripts: As technology advances, the number of sensing and measuring devices used to observe our environment will rise, with increasing levels of sophistication and automation. It is expected this will generate more complex and unstructured data in larger volumes presenting analytical challenges in digital ecosystems. To facilitate complex or automated analysis of such data, we recommend a code-agnostic repository of scripts that are accessible through resolved instrument PIDs, enabling end-users or connected systems to transform instrument data. This seems timely with the continued evolution of such applications such as Pachyderm[19], that orchestrate data transformations and drive machine learning models through automated pipelines using code-agnostic scripts.

Events: Operational instrument information can change over time particularly with different installations and monitoring applications. Our survey results showed that respondents were capturing or using dynamic instrument information. In particular, the purpose, mounting platform and sensor settings of device installations (fig. 2b-d). Thus we recommend a component that can register changes in metadata over time for an instrument PID. We recommend that persistent identifiers redirect to URI endpoints that can then be queried by time to enable systems to extract the most relevant information required. Accessing all the instrument information at once may be extensive and may affect computational performance in digital ecosystems.

Instrument PID inventory: An indexed inventory of instruments that are used in the EDS will support a commons approach, as it is anticipated that not all instruments in the EDS will be assigned with PIDs issued by the EDS minting service. For example, where datasets arise from a project with international partners. This will enable all instrument PIDs that are used in the EDS to be available to the commons, no matter where they were issued. The primary function of the inventory will be to support provenance and interoperability in the commons, particularly in dataset assets.

The principle behind the RDA iADOPT interoperability framework for observed properties may offer a way to enhance interoperability between instrument PIDs, particularly when they are registered by different authorities. This framework harmonises observed properties by constraining a set of properties (components) that are common to each of them (e.g. using common categories for a variable's property and object of interest). The properties of the

PIDINST metadata schema are conceivable common properties of instruments. Some of these could be constrained against formal semantic resources to enable interoperability across instrument PIDs registered in different places. For example, manufacturer and instrumentType may be constrained against the NERC Vocabulary Server controlled vocabulary collections for manufacturer[20] and device categories[21] in the inventory.

Hierarchical recommendations

Linking to research outputs: One of the important applications indicated by our PID requirements gathering survey was linking instrument PIDs to research outputs. Instrument PIDs could conceivably be linked directly to, for example, datasets if they are minted using the Digital Object Identifier (DOI) type of persistent identifier through the 'relatedIdentifier' property and 'relationType' of 'IsCompiledBy'. However, gathering metrics from direct links is not as straightforward. Knowledge graphs like the FREYA PID Graph[22] or the OpenAIRE Graph[23] are emerging to support this purpose. It is possible to contribute to the OpenAIRE Graph by registering our new service through the JISC OpenDOAR directory of repositories[24]. The RDA PIDINST group is currently working to integrate instruments into the Graph. Furthermore, there is scope to integrate instruments into an emerging concept around complex citations[25], an efficient way to cite multiple PIDs in journals.

Governance: Establishing effective governance will be a critical factor for the success of this service. Governance will likely be two-fold. Community governance is a primary driver in the commons approach, that sets the common rules for participating and sharing digital information. As a tool of an EDS commons, the overarching governance will be driven by the needs of the commons. There will also need to be some level of operational capacity that will potentially need to be governed below this level to ensure business as usual. Our long-term aim is to develop this framework incrementally through delivering function and value frequently. Principles similar to the maturity mapping levels proposed by Eaves et al. (2022)[26] regarding governance of digital public goods may be of value.

Training and support: Training and support tools will be made available to encourage and assist users to publish instruments through the service. They will also train users on disciplinary metadata, controlled (machine readable) terms and tools to incrementally enrich

information about their instruments, maximising their value in connecting digital systems. Our results indicated that instrument PIDs are most likely to be created at instrument procurement or first installation (fig. 4), essentially at the data collection stage of the NERC Digital Ecosystem. Thus, training and support could be prioritised towards the actors involved in this stage.

What are your overall reflections from the work (what worked, what did not work, overall lessons learned):

What worked well: The online surveys (especially our contribution to the landscaping survey) enabled us to get a snapshot around current capability and expertise for managing and working with device information.

What did not work: The online survey to gather requirements for instrument PIDs was developed late in the project as a replacement for anticipated project-wide stakeholder workshops. As a result, we only gathered responses from a small number of respondents and these results were considered qualitative only. Feedback also suggested this survey was long and required knowledge of PIDs in general. We intend to revise, shorten and republish this survey to a more simplistic set of questions to fully formalise our initial interpretation of these results prior to the next phase of any development,

Overall lessons learnt: Online surveys should be developed early in a project as they may require several iterations before they are able to provide good results. Consideration must also be given to UK GDPR and contacting respondents.

What recommendations would you make for the next phase of EDS commissioning:
There is a drive to develop digital infrastructure, standards and protocols towards the use of sensor networks and automated data ingestion in the Acquisition and Ingestion (A&I) theme of the next phase of EDS recommissioning. This will be critical as the number of sensing and measuring devices used to observe our environment is expected to rise. Accurately analysing, modelling and simulating this data to gain new environmental insights and understanding will become more challenging as the volume, complexity and automation of environmental data

grows. Thus, common ways to reliably identify, link and access information about the devices used to generate environmental data are required, especially if we want to effectively use the data in connected digital systems. Our findings showed there is a potential for a new identification service for instruments that generate environmental data using globally unique and persistent identifiers following international strategy endorsed by the RDA. Using flexible components such as Common Data Elements, the service has the potential to evolve to new and specialised communities with broad use beyond the EDS (for example the JISC equipment.data.ac.uk). We recommend completing a Business Model Canvas in addition to feasibility studies.

What recommendations would you make for the wider objective of developing a commons approach for environmental assets: Our report recommends a new instrument PID service as part of an EDS asset commons that fulfils two key functions: a PID minting service for new instrument and an inventory of instrument PID assets to support the provenance and interoperability in the commons approach. In the next phase of development, we recommend pilot studies with representatives from the EDS and wider communities, potentially addressing environmental themes that impact society and the economy (e.g. flooding, coastal hazards, climate etc.):

(1) Test the feasibility of the Common Data Elements graph through a pilot service using established instrument information systems. This will involve validating structured data markup (e.g. SHACL) for building Common Data Elements and subsequent knowledge representation following resolution (e.g. World Wide Web Consortium (W3C) Semantic Sensor Network (SSN) ontology, OGC SWE, W3C Web of Things, W3C JSON-LD).

(2) Pilot an inventory of instrument PIDs to support provenance and interoperability in a commons using business information management systems from the EDS and NERC (e.g. National Marine Facilities Inventory Management System, JISC equipment.data.ac.uk, NCAS Data Project).

A-4

The STAC Experiments

Background

This document is a summary of the "STAC experiments" done by the work package 2 of the EDS Phase 1b project.

What is STAC

SpatioTemporal Asset Catalog (STAC) is an open standard designed to share geospatial assets, such as satellite images and other Earth observation data files (https://stacspec.org/en). STAC provides a common framework for describing the location, time, and other key attributes of these assets. It allows organisations and developers to create catalogues of geospatial data that are easily discoverable, shareable, and interoperable across different platforms and tools.

STAC is particularly valuable in the field of Earth science and remote sensing, as it simplifies the process of discovering and accessing geospatial data from various sources, however it is more general than that as anything with temporal or spatial information can be presented as STAC records. It is fast becoming a community standard for Earth Observation as it neatly connects records for assets like Cloud Optimised GeoTiff files, and higher level discovery catalogue records.

Why try STAC

STAC is a good example of a commons approach in the Earth Observation Domain. It is quite minimal, but has a lot of the elements needed for a commons - consistent identifiers, connection between assets, etc.

Objectives:

List of initial objectives:

- Experimentation with SpatioTemporal Asset Catalogs (STAC[1]) to see how generalizable lower level cataloguing of data centre holdings can be to facilitate use.
- Determine whether feasibility of offering a consistent common search across all data centres holdings which is more fine grained than dataset discovery records. Produce some demonstrators to prove this.
- Assess capability of STAC as unifying common technology across EDS disciplines.

To what extent have the objectives been realised:

- Experimentation - A number of experiments were done across datasets with a variety of types and shapes. Even within the CEDA data centre the variety makes it very difficult to make a single generalised workflow for making a STAC representation of the archive. Specific realisations of STAC have been used to access datasets, for example, BODC used STAC the Haug-Fras use case (See pIMFe project). These experiments were labour intensive. It is clear that STAC is a useful tool in this space but generalising is labour intensive.
- Feasibility of scaling to the whole EDS - we have shown that production of a low level STAC catalogue for the complete CEDA archive was possible (ATOD and FBI experiments below). However, to produce the complete coverage we must sacrifice utility. We could pick out datasets where we cover large swaths of the data holdings and give useful search functionality.
- We have assessed the capability of STAC as unifying common technology across EDS disciplines. We can see that STAC is a useful addition to unifying the data across the data centres, but it is not something that should be used universally used.

Collaborations:

Internal to the project:

CEDA and BODC had an interest in STAC.

Externally:

The Met Office is considering using STAC for its datasets. They are interested because the STAC approach could act as a framework for discovery level record. We have had several meetings principally to convey some lessons learned.

The EO data hub project is carrying on the STAC work at CEDA. This project is aimed at producing an Earth Observation data platform so the STAC as the preferred underlying technology fits very well. This project has many space sector industry partners.

The pIMFe project lead by BODC was already using STAC in its pilot for making a digital twin of the Haig Fras marine conservation area.

Summary of approach

(summarise how you have gone about the research, methods used, etc,):

The general approach to this work was to try and achieve the objectives in an agile way. Expanding scope to take in more data as we thought it could be dealt with. We also used real data to highlight real practical problems with any implementation.

We started with a list of experiments and adapted them to see if

ATOD - “All The Other Data”:
We scaled up existing trials with STAC to cover all data files by using the directory structure to provide search facets in a single “Other” collection. This did work but does not present a very useful search interface.

FBI - STAC: Can we put a thin layer on our existing File Based Index to present current catalogues as STAC? We did manage to do this and connect as a hierarchical STAC collection view of the archive based on directories in the CEDA archive. Its utility is debatable.

CMIP as STAC - Can you present climate model output as STAC records that enable use? Several attempts have been at this. The highly structured CMIP5 and CMIP6 files make it easy to extract detailed information for the STAC records. However, STAC items in this case are generally global in extent and cover decades or hundreds of years. SpatioTemporal discovery search is not very useful. SpatioTemporal information is needed in this case when using the data, but this is not facilitated by STAC. We turned to Kerchunk as a technology to help us with the use side. (We are looking a STAC with Kerchunk assets for the EO data hub project because of the work here).

PID’s in STAC. BODC looked at including instrument PIDs in STAC records. There are several ways to do this. This is an example of where the flexibility of STAC leaves you with too much choice. A specific profile would be needed to make it generally useful.

Discovery records in STAC. While STAC provides a good connection to records of low level assets such as files or API endpoints, it also has a collection concept which perhaps maps to the EDS discovery records. With this in mind we tried to map CEDA discovery records to something we could put in STAC. It was relatively easy to flatten the relational database into records that we can envision being encoded as collection. There are a lot of ancillary records that are not suited to STAC.

What outputs have been produced (prototypes, reports, papers):

- STAC Browser View of CEDA Archive via radiantearth
- CEDA STAC Collection
- CMIP6 with Web Front End
- PV presentation https://cds.cern.ch/record/2861133

Findings:

What are your overall reflections from the work (what worked, what did not work, overall lessons learned):

What worked well:

- For the EO community, STAC looks like the way forward, but even this is not trivial.
- If spatiotemporal search is the principal axis for discovery and access, it's a good fit for a technology that can accommodate many datasets. This is not always the case for the data in the EDS.
- Mixing in access technologies like Kerchunk looks like a great way to progress a more unified workflow from discovery to use.
- The concept of a data granule, the STAC "item", the smallest discoverable unit of data. This works in a lot of cases and clarifies how we think of datasets.

What did not work:

- STAC has a lot of extensions. These may tie you to a community and pull you away from more flexible and generic use…
- But, STAC is very general; you have to superimpose community standards if it's going to be useful.

Overall lessons learnt:

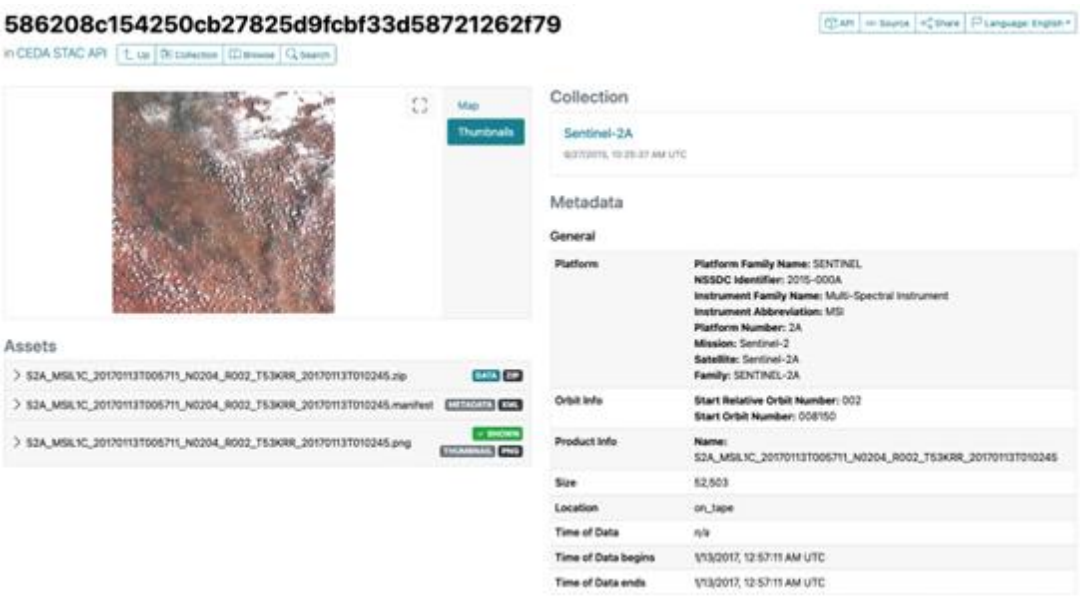
- Trials on small discrete datasets are not enough to work out how much effort you need to expend to cover the entire archive. Scale matters.

What recommendations would you make for the next phase of EDS commissioning:

- STAC will be important for EO. We will have to offer STAC interfaces for EO and EO related datasets.
- It is too hard/expensive to offer STAC for everything. We should not attempt this.
- There is a need to look at the whole workflow discovery to use. People discover a list of assets and then have to have a way to use them. STAC is helping here but there is a need to combine this with data access API's like Kerchunk to complete the use. We should investigate aggregation interfaces like Kerchunk.

What recommendations would you make for the wider objective of developing a commons approach for environmental assets:

The commons should define an "item" for the EDS, the smallest discoverable unit of data. If it did this we stand a chance of being able to access/process/use the list of items that a discovery service may produce.



View of EO data in external STAC client.

A-5

Discovery metadata harmonisation and enhancement across the NERC Environmental Data Service (EDS)

Objectives:

List of initial objectives:

Harmonise and enhance underpinning metadata across the EDS data centres, supporting semantically enriched discovery and interoperability of digital assets in a Commons approach by:

- Drafting EDS specific guidance for populating common GEMINI compliant discovery metadata elements for publication of digital assets in the NERC Data Catalogue Service.
- Elevating the requirement levels for certain metadata elements/sub elements to improve data accessibility.
- Enhancing interoperability through mandated use of common controlled vocabularies wherever possible and domain specific ones when required.

Implementing EDS specific discovery metadata guidance with improved accessibility and more widespread use of controlled vocabularies will make the EDS more integrated and improve the Findability, Accessibility, Interoperability, and Reuse (FAIR) of digital assets.

To what extent have the objectives been realised:

An EDS discovery metadata working group with representation from each EDS data centre has been established to bring together expertise for underpinning Commons approaches. The group has been primarily focusing on discovery metadata records for datasets, which constitute the principal common digital asset in the NERC Data Catalogue Service. The group has been collaborating to create new EDS specific guidance for creating GEMINI compliant discovery metadata records. This has involved:

- Adapting element semantics for greater consistency.
- Expanding the use of controlled vocabularies to increase interoperability.
- Elevating the requirement levels for some elements to enhance data accessibility.

The group is currently finalising the newly proposed guidance.

Collaborations:

Three meetings have been held with representatives from each EDS data centre (UK PDC (BAS), NGDC (BGS), BODC (NOC), CEDA (NCAS), EIDC (CEH)) to discuss and agree on proposed guidance adjustments, suitability of proposed controlled vocabularies, and the viability of elevated requirement levels.

Feedback received from Ettore Murabito from the NERC Digital Solutions (DSH) team highlighting inconsistencies and shortcomings that DSH observed in EDS discovery metadata records. DSH also highlighted a list of metadata elements that are relevant to their operations.

Summary of approach (summarise how you have gone about the research, methods used, etc,): One notable commonality within the EDS is that all data centres generate GEMINI compliant discovery metadata for publication in the NERC Data Catalogue Service. While GEMINI provides a comprehensive standard, some elements, particularly those involving free text and those where controlled vocabularies could potentially be used, can be interpreted loosely. Additionally, the requirement levels for important elements, such as the Resource Locator, can be somewhat relaxed. Indeed, a URL and information on data access is crucial for the value of the metadata record, particularly in a Commons, but this information is not always present . As a consequence, there are inconsistencies in how each data centre populates their metadata, as highlighted by both the DSH team and the EDS October 2021 DCS Reviews: Summary Report[1].

To enhance EDS discovery metadata records for greater commonality, interoperability, and accessibility in a Commons, an EDS discovery metadata working group was formed with representation from each data centre. The group adopted the following strategy:

1. Establish new EDS specific discovery metadata guidance by adapting GEMINI element semantics to enhance consistency and interoperability within the EDS.
2. Each data centre identifies a representative metadata record based on its usage and a separate record with observable properties applicable to most data centres. Apply the new guidance to the representative and interoperable (datasets with common

observable properties) metadata records, to test the new guidance on the interoperability of datasets across EDS domains and its compatibility with datasets that have the highest impact in each data centre.

3. Add enhanced metadata records to a Catalogue Service for the Web (CSW) specifically for the EDS discovery metadata working group for live testing in the NERC Data Catalogue Service, serving as a proof of concept.

As of now, the EDS discovery metadata working group is finalising the EDS specific guidance for discovery metadata.

What outputs have been produced (prototypes, reports, papers):

- EDS discovery metadata guidance document (currently being finalised).
- Live 'proof of concept' on the NERC Data Catalogue (currently being finished).

Findings:

While this activity has not yet reached its final proof of concept stage (live testing in the NERC Data Catalogue), the EDS discovery metadata working group has made successful strides in adapting GEMINI element semantics for improved consistency, interoperability, and data accessibility. Group collaboration revealed that there is a good level of overlap in how metadata element semantics could be restructured to increase consistency within the EDS while also meeting the needs/use cases of each data centre. The group also agreed to elevate the requirement levels for certain elements and sub elements, notably the Resource Locator element. Shifting its status from conditional to mandatory will:

- Satisfy NERC Digital Solutions (DSH) request for Machine-to-Machine (M2M) access to EDS data.
- Enable general M2M data accessibility.
- Enhance the connection between the data products and the organisations that provide them.
- Increase the FAIRness level of EDS datasets.

To increase interoperability, the group were able to reach consensus on mandating the use of specific controlled vocabularies for many elements. However, this was not possible for the

elements that indicate the general subject area of the data resource using keywords, as the EDS covers a broad range of scientific disciplines. However, this issue can be addressed through the use of mappings and mapping frameworks like the RDA IADOPT[2] which provides a standardisation framework for the description of observable properties.

This activity has shown that element semantics and requirements need to remain broad to a certain degree because a complete 'one size fits all' approach for harmonising and enhancing discovery metadata across the EDS is not entirely feasible, due to different community practices. Despite this, there are many commonalities among data centres which offer opportunities for improvements. Their implementation will increase consistency, interoperability, data accessibility, and overall FAIRness. The enhanced GEMINI EDS profile will also become a component of the narrow middle in the commons architecture.

What are your overall reflections from the work (what worked, what did not work, overall lessons learned):

What worked well:

Strong enthusiasm for this project was evident, with numerous EDS colleagues with expertise in the GEMINI standard and controlled vocabularies eager to participate. The level of experience and enthusiasm within the group eased the process of identifying commonalities between the data centres and reaching consensus on semantic element adjustments. As a result, the group successfully identified enhancement opportunities that would make EDS discovery metadata more consistent, interoperable, and make data assets more accessible. Tightening the requirements of some metadata elements will improve consistency across the EDS. Indeed, making it possible to consistently access data in metadata records will maximise the value of an EDS Commons and benefit connecting systems such as the NERC Digital Solutions. However, this will depend on developing easily accessible data access points at each data centre.

What did not work:

Due to the wide range of scientific disciplines covered by the EDS, limiting certain elements to specific controlled vocabularies was challenging. Nevertheless, the group agreed to use controlled vocabularies whenever feasible in these cases.

Overall lessons learnt:

This activity revealed that element semantics and requirements should maintain a degree of broadness because a complete 'one size fits all' approach for harmonising and enhancing discovery metadata across the EDS is not entirely possible. However, there are many commonalities among data centres that present opportunities for improvement.

What recommendations would you make for the next phase of EDS commissioning:

This work has produced guidance towards improving the overall quality, consistency and interoperability of discovery metadata for datasets in the EDS. Following our proof of concept, we recommend this policy is applied as a common metadata standard to support general federated EDS systems under the Access and Delivery (A&D1) theme of the EDS commissioning.

What recommendations would you make for the wider objective of developing a commons approach for environmental assets:

If the new EDS metadata guidance proves to enhance consistency, interoperability, and data accessibility, the recommendation would be:

- For the new guidance to be applied to all existing and future metadata records. This is likely to significantly improve the EDS Commons narrow middle architecture, benefiting future technologies for greater discoverability of and access to environmental data assets.
- Modify the GEMINI Schematron validation tool for the EDS to assess compliance with the new EDS Metadata guidelines, allowing for automatic testing of XML from EDS data centres prior to publication in the NERC Data Catalogue Service.
- Continue EDS discovery metadata working group meetings on a quarterly basis to refine EDS metadata guidance (where necessary) and maintain a collective Commons approach to future FAIR enhancements to EDS discovery metadata. For example, Data Catalogue Vocabulary (DCAT) and Schema.org mappings to GEMINI elements. This would allow EDS data centres to augment current discovery metadata records with additional metadata properties to further describe datasets outside the constraints of the GEMINI standard. It would also make EDS discovery metadata interoperable with other data portals that use DCAT and Schema.org which would significantly enhance their FAIRness.

A-6

T2.2 EDS Ontology Architecture: Trialling of the RDA I-ADOPT Interoperability Framework and its ontology to connect vocabularies for descriptions of observations across the EDS

Objectives:

List of initial objectives:

- Deploy a prototype of the I-ADOPT framework for the interoperability of observations in the NERC Vocabulary Server (NVS), a key EDS service for vocabularies[1]
- Use the prototype to demonstrate the interoperability of observation descriptions between EDS marine (BODC Parameter Usage Vocabulary), atmospheric and modelling (Climate and Forecast Standard Names) domains using a subset of Essential Climate (ECVs) and Ocean Variables (EOVs)
- Explore how I-ADOPT representation of environmental variables could benefit and be implemented across the EDS

To what extent have the objectives been realised:

Our objectives were achieved in so far as all the activities planned took place; we created an I-ADOPT profile in the NVS and applied I-ADOPT to a subset of terms from two large parameter naming schemes used in EDS datasets. We are now able to demonstrate the usefulness of the I-ADOPT framework in giving easier access to deep information held in EDS data files and facilitate cross-domain interoperability within the EDS and beyond. We also identified possible pilot implementations across EDS data centres.

Collaborations:

Internal to the project:

- Meetings with representatives from EDS data centres (NGDC, PDC, CEDA) to discuss possible implementation of I-ADOPT and future strategy related to I-ADOPT

- Meeting with EIDC (CEH) to present the implementation of I-ADOPT in an EDS commons

Externally:

- eLTER, Research Data Alliance (RDA) I-ADOPT WG, and EnvThes[2] in collaboration with Barbara Magagna, manager of the EnvThes vocabulary hosted at CEH and Co-Chair of the RDA I-ADOPT WG alongside Gwen Moncoiffé
- Ontoportal technology[3]: discussions with Clement Jonquet and Data Terra Earth Portal team (via FAIR-EASE and FAIR-IMPACT) - exploring usefulness of the ontoportal technology as a solution for a central access point to terminologies from the NVS and other resources.
- NERC Digital Solutions
- Climate and Forecast (CF) Community
- Polar Community
- European marine and environmental science community in the frame of European projects like ENVRI-FAIR, EOSC-Future, FAIR-EASE, Blue-Cloud2026, and eLTER

Summary of approach (summarise how you have gone about the research, methods used, etc,):

To demonstrate the I-ADOPT prototype, we targeted a small subset of observations from two large and complex vocabularies of international importance hosted on the NVS that are used globally by marine and atmospheric observation and modelling scientists and data managers: the CF Standard Names managed by CEDA/STFC (identified as P07 collection in the NVS) and the BODC Parameter Usage Vocabulary or BODC PUV (identified as P01 in the NVS). These two vocabularies overlap (i.e. some concepts in each vocabulary are either exact or close matches) but are difficult to align due to their complexity and high level of description. This creates issues when trying to search across, or combine data from files annotated with either of these two vocabularies.

The I-ADOPT framework and ontology[4] is an international recommendation endorsed by the RDA (Fig. 1). It offers a way to decompose these complex observational terms into their essential common components such as the property kind (e.g. temperature, wave height, speed, (bio)mass, practical salinity, (directional) velocity), the object type targeted (this could be a chemical substance, a physical process, a biological organism and its components), and

the matrix in which the object is embedded (e.g. atmosphere, water body). So interoperability between dataset with variables named according to different conventions can be enabled by matching the I-ADOPT atomic components associated with these terms.

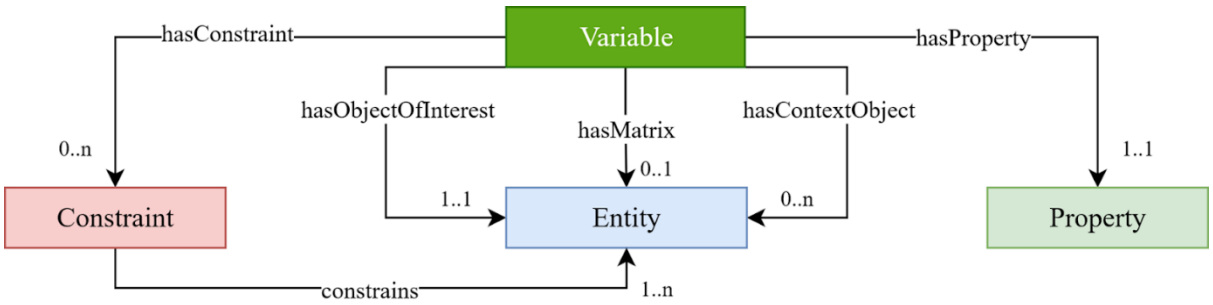


Figure 1. A schematic view of the basic I-ADOPT (InteroperAble Descriptions of Observation Properties Terminologies) ontology

To test this theory, we added the I-ADOPT ontological object properties (hasProperty, hasObjectOfInterest, hasMatrix, hasConstraint) in the NVS database and created a new profile for I-ADOPT (Fig. 2); we selected a subset of observational terms from both the CF Standard Name and the BODC PUV that critically assess the state of the Earth's climate and oceans (ECVs and EOVs). We then identified semantic resources to be used as common vocabularies. Here we had a choice: to only use NVS vocabularies as a primary source of atomic concepts, or to prioritise the direct use of terms from external resources that are increasingly used as references for core environmental properties and concepts. These included QUDT[5] for quantity kinds, ChEBI[6] for chemical entities, EnvO[7] for generic environmental terms. The decision was to take the second approach whenever an exact match was possible, and use terms from the NVS whenever more specific terms were required. While both approaches would have worked (since the NVS terms are also often mapped to these external resources), the reasons to select the latter was three-fold: 1) avoid creating duplicate concepts if these already existed elsewhere; 2) demonstrate the feasibility of adopting a common set of external reference vocabularies across domains; 3) foster collaboration between terminology managers and infrastructures. This approach will be discussed and evaluated by the I-ADOPT WG and other adopting organisations.
In cases when the NVS was used as a source of atomic concepts, the following vocabulary collections were used:

http://vocab.nerc.ac.uk/collection/S06 for property measured

http://vocab.nerc.ac.uk/collection/S27 for chemical substances

http://vocab.nerc.ac.uk/collection/S21 for environmental matrices

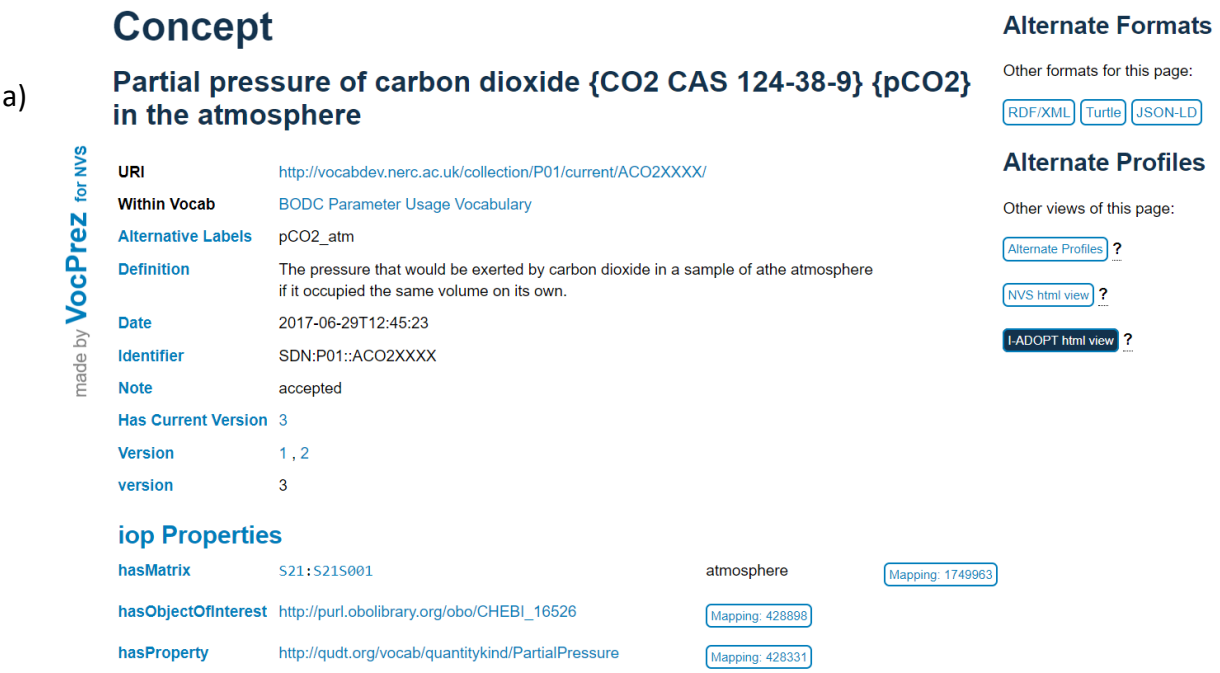
http://vocab.nerc.ac.uk/collection/S18 or http://vocab.nerc.ac.uk/search_nvs/S29 for physical objects and processes

The fact that both vocabularies are published on the NVS gave us the freedom to test the I-ADOPT implementation on our vocabulary development server without affecting the published versions of these vocabularies.

What outputs have been produced (prototypes, reports, papers):

- Creation of an extension to the NVS covering new ontological properties and a new profile that will enable us to express any observable property terminologies managed via the NVS according to the I-ADOPT ontology, enabling semantic bridging of variable names across multiple domains
- Presentations of the I-ADOPT framework, its implementation in the NVS and the proposed strategy at conferences and workshops including RDA Plenary P20 and P21 sessions in March and October 2023, Polar Data Forum in October 2023
- 166 external mappings and 137 internal mappings decomposing 40 CF Standard Name (i.e. P07) terms and 73 BODC PUV (i.e. P01) terms into their atomic elements. The mappings are currently kept on our development server (Fig. 2) awaiting review by external colleagues prior to publication onto the NVS

a)



b)

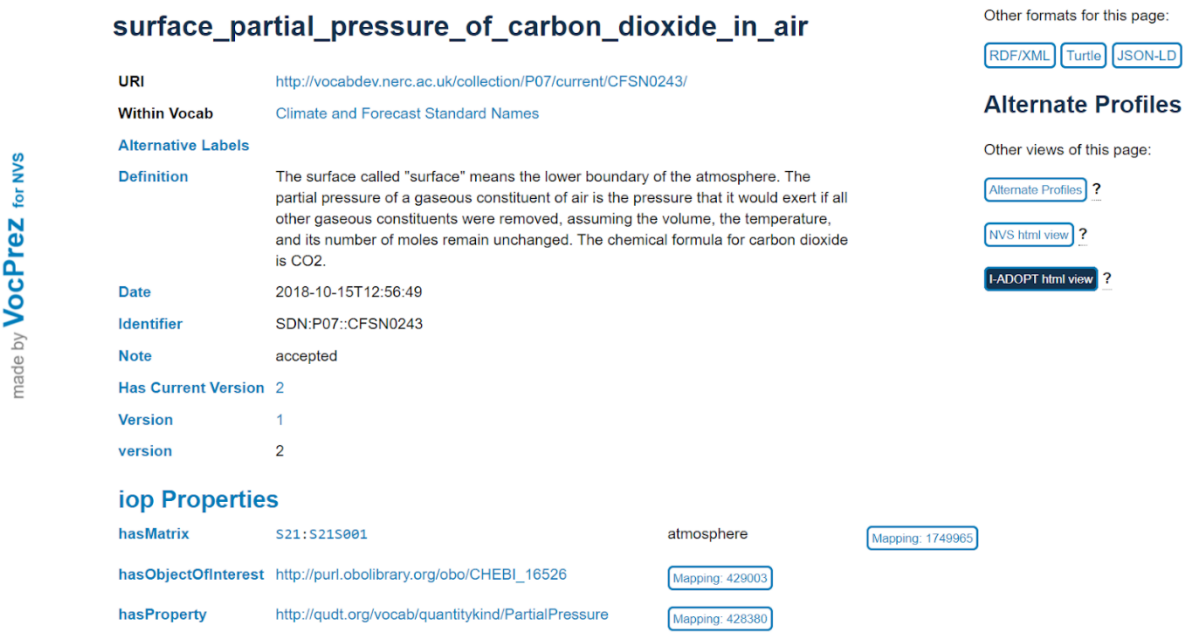


Figure 2. Example of an implementation of I-ADOPT on the NVS development server, against a BODC PUV concept (a) and against a CF Standard Name concept (b) using external terminologies like QUDT and ChEBI. Access to terms mapped using the I-ADOPT object properties is accessible via the NVS I-ADOPT profile (dark blue button on the right shows activation)

Findings:

What are your overall reflections from the work (what worked, what did not work, overall lessons learned):

What worked well:

- The NVS VocPrez application and the NVS infrastructure made it relatively easy to implement the I-ADOPT ontology to selected collections and concepts, and make it accessible as a separate specialised feature
- BODC PUV and CF Standard Names are already well structured vocabularies, making the implementation of the I-ADOPT Framework simple and somewhat easy to model
- There was good crossover between the semantic elements making up the BODC PUV and the CF Standard Names, and elements contained in external vocabularies (ChEBI, EnvO, and QUDT) as shown in Table 1

Table 1. Percentage coverage of P01 and P07 by QUDT, ChEBI and EnvO for property, object of interest (OOI) and matrix

EDS NVS vocabulary collections	**QUDT (property kinds)**	**ChEBI/EnvO (objects of interest)**	**EnvO (environmental matrices)**
CF Standard Names (P07)	77%	77%	67%
BODC PUV (P01)	52%	43%	70%

What did not work so well:

- Some generic terms that are used in the BODC PUV vocabulary (e.g. “Concentration”) do not align well with any vocabulary unless additional information is considered. In this instance, the reporting unit would be needed in order to indicate what type of “concentration” it is: molar, mass etc. We have proposed a

work-around solution on our GitHub repository (https://github.com/nvs-vocabs/S06/issues/76) but we will continue consulting with colleagues prior to deciding the best way forward

- We only targeted the minimum necessary set of I-ADOPT atomic elements to demonstrate interoperability on a high level and we left out the "iop:Constraint" class (Fig. 1) of the I-ADOPT Framework. However the iop:Constraint class will need to be considered in the future in order to fully represent the P01 and P07 terms. It could be a challenge to decide how to apply this element in a consistent way, one that will require consultation with managers and users of the two resources
- Some concepts such as terms related to wave spectra are very specialised and difficult to align and may require the further assistance of experts in these fields
- Due to the short timeframe of the project, it was not possible to meet with colleagues from the NGDC and PDC more often, however we have identified some areas where the use of I-ADOPT could deliver valuable results
- The time-frame of the project was also too short to obtain ratification of the approach adopted for the mapping of the CF Standard Names by the CF governance group

Overall lessons learnt:

Aligning terminologies dedicated to the description of scientific variables requires dedicated resource, domain knowledge, and access to a network of collaborators. While our prototype has successfully demonstrated interoperability between subsets of two EDS vocabularies, more time is needed to fully realise the potential for the I-ADOPT Framework in all EDS domains. This could be achieved through impactful demonstrations.

The time and effort invested in implementing the I-ADOPT common framework is a worthwhile investment as I-ADOPT is gaining continued momentum globally, with several parties showing interest in adoption (e.g. Long-Term Ecosystem Research (LTER) e-infrastructures, elements of the French Integrated Earth Observation System including e.g. Critical Zone Observatories, Pangaea). Use of the I-ADOPT ontology to bridge between terminologies and to serve as a base for reasoning and smart-mapping, is also at the centre of a number of federated search and data access pilot demonstrators developed in the frame of European projects such as ENVRI-FAIR, EOSC Future, Blue-Cloud 2026 and FAIR-EASE projects.

Recently NGDC has become aware of colleagues within the geological domain (GSEU https://www.geologicalservice.eu/) working on modelling of observable properties using an approach similar to I-ADOPT, strengthening the prospect that this solution can work in the geological EDS domain in addition to those demonstrated here. As we have seen with the impact of the FAIR principles, the more infrastructures and organisations start expressing their observed variables according to a common framework like I-ADOPT, the more impact the work will have and the more opportunities it will offer.

What recommendations would you make for the next phase of EDS commissioning:

- Dedicate more time and resources to collaboratively work within the EDS on developing mapping between resources using the I-ADOPT framework
- Target concepts and data assets that can be re-used to demonstrate real use cases of how I-ADOPT facilitates more specific discovery, interoperability and easier re-use of data assets
- Prototype an EDS vocabulary commons approach using technologies like OntoPortal to provide a central access point to terminologies from the NVS and other resources.

What recommendations would you make for the wider objective of developing a commons approach for environmental assets:

- Establish a small working group dedicated to mapping terms to I-ADOPT across the EDS and interact with other groups working on the same issues
- Explore ways to incorporate I-ADOPT to the Gemini profiles

Appendix

Acronym	Full name	Link
BGS	British Geological Survey	https://www.bgs.ac.uk/
BODC	British Oceanographic Data Centre	https://www.bodc.ac.uk/
CEDA	Centre for Environmental Data Analysis	https://www.ceda.ac.uk/
CF	Climate & Forecast	https://cfconventions.org/

ChEBI	Chemical Entities of Biological Interest	https://www.ebi.ac.uk/chebi/
ECV	Essential Climate Variables	https://gcos.wmo.int/en/essential-climate-variables
EDS	Environmental Data Service	https://eds.ukri.org/
eLTER	Integrated European Long-Term Ecosystem, critical zone and socio-ecological Research	https://elter-ri.eu/
EMODnet	European Marine Observation and Data Network	https://emodnet.ec.europa.eu/en
EnvO	The Environment Ontology	https://sites.google.com/site/environmentontology/
ENVRI-FAIR	Environmental Research Infrastructures - Findable, Accessible, Interoperable, Reusable	https://envri.eu/home-envri-fair/
EOSC	European Ocean Science Cloud	https://eosc-portal.eu/
EOV	Essential Ocean Variables > AtlantOS Essential Variables (A05)	https://vocab.nerc.ac.uk/collection/A05/current/
FAIR-EASE	(EOSC) Building Interoperable Earth Science & Environmental Services	https://fairease.eu/
I-ADOPT	InteroperAble Descriptions of Observable Property Terminology	https://i-adopt.github.io/
NGDC	National Geoscience Data Centre (BGS)	https://www.bgs.ac.uk/geological-data/national-geoscience-data-centre/
NVS	NERC Vocabulary Server	https://vocab.nerc.ac.uk/
OBIS	Ocean Biodiversity Information System	https://obis.org/
PDC	UK Polar Data Centre	https://www.bas.ac.uk/data/uk-pdc/
PUV	BODC Parameter Usage Vocabulary (P01)	https://vocab.nerc.ac.uk/collection/P01/current/
QUDT	Quantities, Units, Dimensions and Data Types	https://www.qudt.org/
RDA	Research Data Alliance	https://www.rd-alliance.org/

A-7

Consideration of governance elements of data commons

Objectives:

List of initial objectives:

1. To carry out a consideration of different governance models for data commons and their application to an EDS data commons
2. To reflect on how a governance framework can enhance the scope and scalability of an EDS data commons model and its federation with other research commons and digital research infrastructures.

To what extent have the objectives been realised:

The first objective was achieved through consideration of existing examples of data commons governance principles together with the review of the development of principles and guidance from research on self-governing systems.

This led to further reflection on development of design principles to increase the scalability of a NERC-EDS data commons and compatibility with digital twin developments.

The latter consideration brings this work together with DRI work on the NERC Information Management Framework for environmental digital twins (IMFe). Piloting of IMFe principles ran in parallel with this project and informed thinking around similar governance principles for a digital asset common capable of extending the scope of NERC-EDS to include digital twins.

Collaborations:

Internal to the project:

- With tasks related to the development of key elements of governance such as standards necessary for identification and description of a data commons
- With the whole project, most notably through the EDS Enhancement Project Workshop: Building a RoadMap for a NERC Data Commons, 25th-26th May, 2023, Lake District, UK and the EDS Futures Webinar of the 6th November 2023.

Externally:

- With the IMFe and P-IMFe projects looking at an Information Management Framework for environmental digital twins, esp. around commons architectures and cataloguing
- Australian Research Data Commons (ARDC) (https://ardc.edu.au/), incl. EcoCommons (https://www.ecocommons.org.au/),
- Elixir (https://elixir-europe.org/), Elixir-UK (https://elixiruknode.org/) and BioFAIR (https://biofair.uk/),

Summary of approach (summarise how you have gone about the research, methods used, etc,):

The approach to providing recommendations for EDS commons governance was initially as a management framework to shape the collection of appropriate key performance indicators tracking contribution of the different NERC Environmental Data Centres to delivery of FAIR data commons services (including those related to NetZero). This approach would be an extension of the current governance KPIs outlined in the NERC-EDS commissioning.
The initial approach rapidly proved to be inadequate in scope to deal with the different aspects of building a data commons that could federate to other data centres and research infrastructures such as the Digital Solutions Hub (DSH) and the Floods and Droughts Research Infrastructure (FDRI) within NERC or across research councils RIs such as the BBSRC BioFAIR. To achieve a governance model that could support such federation required a wider examination of commons governance models.

1. Consideration of wider commons governance principles:
The first element of the expanded approach was to consider the NERC-EDS as a self-organising governance model in which resources held in common (e.g. FAIR data assets) are managed and sustained by those members of the commons subscribing to the governance rules.

In this approach, the authorisation of users, representation in decision making, standards and policies, and the sanctions for misuse must be part of the governance model and determined by the members of the commons themselves. Resources held in common are governed by their members for their mutual benefit and that of authorised users. Without such rules, common resources become overexploited and undermaintained leading to the so-called “tragedy of the commons” where resources become degraded due to the lack of effective governance for their sustainable use. Garred Hardin’s influential 1968 Science paper of the same name proposed that members acting in self-interest would inevitably degrade resources open to all and that a single authority was required to protect the resource thus denying that effective self-organisation and self-regulation could provide sustainable commons governance model. This would seem to argue for a top-down authority to impose standards on the commons.

The tragedy of the commons remains a widely held concept to be guarded against in environmental resource management with parallels in digital commons such as effort and resources required in maintenance of open-source software or crowd-sourced knowledgebases versus derived commercial value from unregulated access by non-contributing users (free-loading). However, Elinor Ostrom’s Nobel prize winning (2009) work in economic governance documented several self-organising systems that demonstrated that a polycentric governance model was better suited to sustainable resource management than a simple single top-down authority model as its members could solve resource management problems more flexibly and at the appropriate scale over long periods of evolving user needs. She maintained that complex resource management issues could not be addressed by simplistic top-down governance models. She established a set of principles by which a governance model could be developed by the members of the commons:

Elinor Ostrom's 8 Principles for Managing a Commons

1. Define clear group boundaries.
2. Match rules governing use of common goods to local needs, capabilities and conditions.
3. Ensure that those affected by the rules can participate in modifying the rules.

4. Make sure the rule-making rights of community members are respected by outside authorities.
5. Develop a system, carried out by community members, for monitoring members' behaviour.
6. Use graduated sanctions for rule violators.
7. Provide accessible, low-cost means for dispute resolution.
8. Build responsibility for governing the common resource in nested tiers from the lowest level up to the entire interconnected system.

2. Consideration of existing DRI federation standards versus to the commons principles:
There are parallels in Ostrom's 8 principles with the increasing complex nature of interconnecting DRIs and associated data and computational resources (e.g. for designing federated digital twins). Creating a top-down authority to impose standards and determine access rights across these resources denies the polycentric view of how a multidisciplinary research community needs to evolve the governance rules for sustainable use and management of these interconnected DRI resources to address complex research questions.
The parallels with data commons management can be seen in the governance principles for federating elements of DRIs (such as in the those for development of digital twins via the Gemini Principles). If each DRI component follows these principles, then many of Ostom's principles are met in terms of solving issues at the appropriate level by members who have appropriate knowledge and are committed to the sustainability of the commons as a whole (e.g. supporting common findability, accessibility and acknowledgement standards).

3. Consideration of a multitier governance model across federated DRIs:
Bringing together digital resource within and across DRIs to enable multidisciplinary research communities is as much a cultural as a technical issue and this is reflected in the key challenges in creating an appropriate governance model. As potential interconnections and diversity of digital resources creates complexity, so a governance model is required that can address and adapt with that complexity. Hence, the principles of a self-governing, self-organising commons advocated by Ostrom seem to best address issues representing the heterogeneity and multiple

layers with which federated data commons and DRIs operate. Without the ability to enable inclusivity in decision making and addressing legitimate vested interest of resource managers, governance models for such complex systems of systems will neither address the changing complex requirements of their users nor enable measurable benefits for the resource managers who sustain them.

The following principles were distilled from architectural and commons governance principles developed alongside the PIMFe project but are pertinent here as principles covering both data commons and digital twin governance:

- Environmental digital twins must fulfil a common purpose:
    - They must primarily provide measurable benefit to the members of a commons or federation that operate them
    - They must provide value in addition to their individual system components such as reducing barriers to federation enabling research innovation
    - They must provide insight toward actionable knowledge to understand and manage the environmental systems they twin
- Environmental digital twins provide open and trustworthy services to all authorised stakeholders
    - Their services and assets are secure and open to levels authorised through a common authentication and authorisation standard
    - The standards for discoverability, accessibility and re-use should be defined through international standards such as those implementing the FAIR principles
    - The trustworthiness in the quality, sustainability and user focus of the services provided should be defined through international standards such as CoreTrustSeal
- Environmental digital twins must function across heterogeneous and evolving federated systems to add value
    - They must federate different systems into their architecture recognising heterogeneity but providing baseline standards for system interactions
    - They must ensure governance and regulation structures to maintain both consistent operation and ongoing relevance of services
    - They must respect existing system to system interactions while providing added value for cross-system interactions such as through minimum standards for interoperability of components and assets

What outputs have been produced (prototypes, reports, papers):

We have identified key principles for governance models for a data commons and linked these to recommendations for developing federated DRI governance. A key finding is that a governance model must be multitiered and responsive to requirements across all levels of the commons. These principles were presented and discussed at the “Building a RoadMap for a NERC Data Commons” workshop and developed into a set of recommendations given below.

Findings:

What are your overall reflections from the work (what worked, what did not work, overall lessons learned):

What worked well:

1. Consideration of general commons governance principles identified that the governance for NERC-EDS must not be top-down but be a multitiered model as advocated by Ostrom's 8 principles to support evolution and sustainability.
2. These reflections are compatible with recommendations for governance principles coming from other areas of the DRI community such as the Gemini Principles for federated digital twins.
3. In discussion at the Lake District DRI Phase 1b workshop, these principles proved a good foundation for further discussion about how NERC-EDS can enable links to other data centres and DRI including the ARDC.

What did not work:

1. The recommendations will need much further development with the user community at different levels and such work has not been factored into the NERC DRI work programme at present.
2. The governance structure currently commissioned by NERC-EDS need to be evolved to less top-down structures if this approach is to be successful.

Overall lessons learnt:

The most important lesson learned from this work is that a governance model for a NERC-EDS data commons needs to be multitiered and able to accommodate new members and federate with other DRIs. This can be done through principles established by the members of the commons that solve issues and apply standards at the appropriate level. A top-down model of governance will hinder the innovation and adaptability required to support a multidisciplinary research community relying on heterogeneous digital assets held in federated data commons.

What recommendations would you make for the next phase of EDS commissioning:

The recommendation agreed at the workshop for future work were as follows:

1. Define the common vision for the purpose, scope and principles of the commons (to be refined by EDS-MB, input from Phase 1 b EDS Enhancement project WP1 report)
2. Identify existing governance frameworks at a start point for EDS (e.g. ARDC)
3. Clearly define the scope of commons resources and who is in the 'governance club' (compliance rules and revisions / evolution)
4. Inclusion and representation of commons members in governance and transparency of decision making (EDI policies)
5. Layered approach to decision making defined to cope with local requirements
6. Conflict resolution and sanctions against member (that are recognised by higher authorises e.g. NERC/ UKRI, host institutions)

What recommendations would you make for the wider objective of developing a commons approach for environmental assets:

Consider the scalability of the governance model to cover a federated set of data commons and other DRIs in different disciplines. A self-governing, multi-level model is most likely to gain cultural acceptance across multidisciplinary research community.

References:

1. Hardin, G (1968). "The Tragedy of the Commons". Science. 162 (3859): 1243–1248. Bibcode:1968Sci...162.1243H. doi:10.1126/science.162.3859.1243
2. Ostrom, Elinor (1990). Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge University Press. pp. 90, 91–102. ISBN 978-0-521-40599-7.
3. Bolton, A., Butler, L., Dabson, I., Enzer, M., Evans, M., Fenemore, T., Harradence, F., Keaney, E., Kemp, A., Luck, A., Pawsey, N., Saville, S., Schooling, J., Sharp, M., Smith, T., Tennison, J., Whyte, J., Wilson, A., & Makri, C. (2018). Gemini Principles. CDBB. https://doi.org/10.17863/CAM.32260

A-8

EDS WP3: Building a Prototype TRE on JASMIN

Introduction

Trusted Research Environments (TREs) are secure spaces where researchers can access sensitive data without breaching privacy, to provide safe and secure data for analysis and research. They are required for any scenarios that involve analysing individual or small area population health data and other data at the individual or small area level such as social security data.

To enable the analysis of sensitive data (potentially in conjunction with 'standard' data such as NERC's environmental data), a TRE capability is needed. This is not currently supported by JASMIN or any other NERC facilities. As such, this WP has implemented a pilot TRE on JASMIN, employing a test scenario comprising synthetic health data alongside sample environmental data to demonstrate how such data could be combined and analysed.

Objectives

The following table lists the initial objectives and the extent to which these have been realised.

Objectives	Results
A prototype TRE running within the JASMIN HPC.	A prototype TRE has been deployed within a JASMIN unmanaged cloud tenancy. Being a prototype, this does not provide the operational processes or service wrap that would be needed for a Production TRE.
TRE meeting the broad technology requirements embedded in the 'Five Safes' model.	The TRE uses Alfresco to provide users with secure content management and collaboration capabilities. The following outlines how Alfresco addresses each of the Five Safes dimensions: 1. **Safe People:** ○ User Authentication and Authorisation: Alfresco provides user authentication and authorisation mechanisms, ensuring that only authorised individuals can access content within the system. Access permissions can be finely tuned to restrict access to specific folders or documents based on user roles and permissions. ○ User Management: Alfresco's user management features allow administrators to control user access, including adding and removing users, resetting passwords, and managing user profiles.

2. **Safe Projects:**
    - Folder and Workspace Permissions: Alfresco allows administrators to set permissions at the folder or workspace level. This means that access can be restricted to specific projects or research areas, ensuring that data is only accessible to those working on authorised projects.
    - Metadata and Tagging: Alfresco supports the use of metadata and tagging, which can be used to categorise content and associate it with specific projects. This makes it easier to organise and control access to project-specific data.
3. **Safe Data:**
    - Data Encryption: Alfresco supports data encryption both in transit (using HTTPS) and at rest (using encryption methods like AES), although the latter is only available in the Enterprise edition (without looking at any 3rd-party addons). This helps protect sensitive data from unauthorised access.
    - Document Versioning: Alfresco includes document versioning capabilities, which can help preserve data integrity and track changes to documents over time.
4. **Safe Settings:**
    - Access Control and Audit Trails: Alfresco provides detailed access control features and audit trails that can be used to monitor and log user activity. This ensures that data is accessed in controlled settings and provides an audit trail of who accessed what data and when.
    - Secure Hosting Options: Alfresco can be deployed in secure hosting environments or on-premises, allowing organisations to choose the level of security that aligns with their needs and regulatory requirements.
5. **Safe Outputs:**
    - Content Publishing Control: Alfresco allows administrators to control content publishing and distribution. They can ensure that only

	approved and sanitised content is made accessible to the public or shared externally. ○ Redaction and Anonymisation: While not a native feature, Alfresco can be integrated with third-party tools or custom scripts to automate redaction and anonymisation processes to ensure safe outputs.
Ability for the environment to be 'persistent'.	The TRE provides each user with a customised, persistent and siloed file system. Further, Alfresco provides the ability to create “sites”, persistent collaboration environments where users can organise, share, and work on content and documents related to a specific project, team, or purpose.
Ability to ingest and store synthetic health data.	Synthetic health data was acquired and stored in a dedicated repository for testing purposes. The process involved the following steps: 1. Establishing a connection with the TRE through a virtual desktop environment. 2. Downloading the synthetic data from the internet within the virtual desktop session initiated by the user connected to the TRE. 3. Logging into Alfresco from within the virtual desktop session. 4. Creating a site within Alfresco to simulate a scenario in which the DSH may need to manage health data or other sensitive information, requiring special handling and care. 5. Uploading the synthetic health data, previously downloaded to the TRE virtual desktop environment, into the designated Alfresco site.
Ability to ingest and store copies of selective NERC environmental data stored in CEDA archives.	An example dataset from CEDA has been obtained and placed in a dedicated repository for testing purposes. The procedure mirrored the steps outlined for acquiring and storing synthetic health data. Specifically, the dataset from CEDA was added to the identical Alfresco site where the synthetic health data had previously been uploaded. It is important to highlight that the CEDA dataset was initially downloaded using a procedure similar to that employed for acquiring the synthetic health data, which involved accessing a URL pointing to the data resource. There was no direct ingestion of data from JASMIN into the TRE environment. If the possibility of such ingestion arises, it must adhere to CEDA's guidelines

	regarding how the dataset can be directly acquired from the JASMIN file system. As of the time of writing, such information is not yet available, and the dataset was downloaded in the same manner as it would have been from any other machine, location, or environment outside of the TRE.
Remote access to the environment from a local PC using Virtual Desktop capability.	YES. Users need to connect to the TRE via a Virtual Desktop Environment (VDE). The VDE runs Ubuntu 20.04.1 LTS and provides each user with a customised, persistent and siloed file system. The TRE employs a two-step authentication mechanism. Initially, users must log in to the VDE using their username and password. Following this, they are prompted to enter a verification code generated by Google Authenticator.
Customised and persistent Virtual Desktop environment for each authorised user.	YES. See comment above.
Ability to conduct analysis on data using Jupyter Notebooks available from Virtual Desktop.	YES
Containerised deployment to facilitate portability and integration into broader infrastructure management capabilities.	The TRE is fully containerised using Docker.
Packaged IaC deployment that can be selectively recreated as required.	NO. As of the time of drafting this document, it is not feasible to replicate the TRE deployment using an (idempotent) Infrastructure-as-Code (IaC) methodology, which entails defining the desired state via code. However, it does support a semi-automated and partially scripted (non-idempotent) deployment.

Collaborations

Internal to the Project

- University of Manchester eLab team
- University of Manchester, Division of Informatics, Imaging, Data Sciences

Externally

- Health Innovation Manchester
- JASMIN
- Health Data Research UK
- Administrative Data Research UK
- Standard Architecture for Trusted Research Environments
- NHS Digital

Summary of Approach

Overall Approach

The key theme was to minimise 'reinventing the wheel' and utilise TRE best practice, insofar as was available. This included:

- Reviewing the available literature on existing TRE standards
- Speaking with owners and users of operational TREs
- Contributing to the development of the SATRE TRE Specification
- Speaking with other contributors to the SATRE specification
- Reviewing existing TRE implementations to assess their suitability for reuse.

A key challenge insofar as reuse is concerned is that many of the more recent and more relevant/updated 'exemplar' TREs are hosted on the public cloud, which engenders specific technical dependencies on these cloud platforms. This largely excluded them from reuse consideration as this prototype TRE required all technical capabilities to be installed and available within the on-premise JASMIN HPC environment.

Reusable Pattern

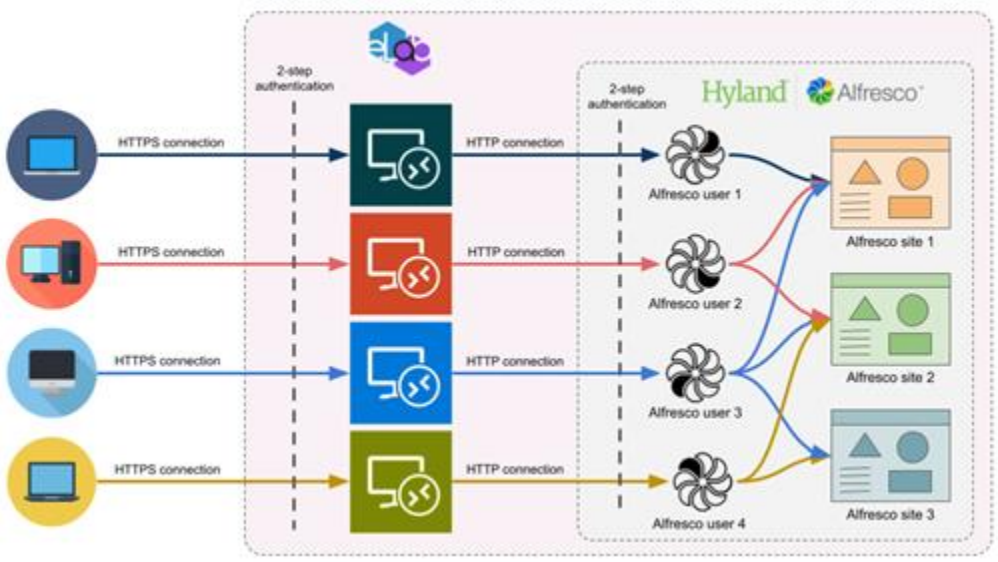
From undertaking these investigations it was discovered that the University of Manchester had independently developed its own TRE pattern (eLabs) which was being actively used within a number of sensitive environments, including the Greater Manchester Secure Data Environment (SDE). This is a secure digital space enabling authorised researchers access to view or analyse NHS and social care data for data-driven planning, research and innovation. In addition to the credibility that the SDE deployment lent to eLabs, a further benefit was that this pattern could (in principle) be installed on-premise. As such a decision was taken early on to utilise this pattern as the basis for the JASMIN TRE prototype work.

TRE Outline Description

To access the environment, users simply utilise their web browsers via an HTTPS connection. This employs a two-step authentication process; users first authenticate with their unique username and password (which they generated when creating their account), and then use a code generated by Google Authenticator. Once successfully authenticated, users are provided with a Linux-based virtual desktop environment (VDE) equipped with a variety of data analytical software, including Jupyter Notebook and RStudio, as well as development tools like Visual Studio Code.

Each user's virtual environment is persistent. This means that the environment retains its state even after a restart. It is also isolated from the environments of other users.

The TRE utilises Alfresco as a content management solution. This provides users with many tools to share, manage and perform operations on their content. Such content is typically organised in "sites", collaborative workspaces where users can organise files, collaborate on projects, and control access to specific sets of documents and data. A dataset that was in the virtual desktop environment of a specific user can be shared and managed by uploading it to that user's Alfresco account and, if necessary, to an Alfresco site for collaboration. The diagram below illustrates the connection flow from client machines to the JASMIN TRE and Alfresco, highlighting how users can share content within collaborative Alfresco 'sites'.



Alfresco also provides a multitude of additional tools, enabling tasks such as data governance, workflow automation, metadata cataloguing, and search and discovery features.

Delivery Approach

The overall work package was principally managed using Atlassian Confluence to facilitate information capture and collaboration spanning multiple internal and external stakeholders. This took the form of creating a dedicated ED3 WP3 Confluence Space (within a Confluence tenancy managed by the University of Manchester).

Regular (bi-weekly) meetings were also scheduled with key stakeholders to manage the deliverables, communicate updates, and proactively address issues.

The delivery approach comprises a number of incremental phases and associated milestones, as follows.

- **Phase 1 - Planning**
    - Initial planning of the design activities, ways of working and inter-team alignment.
- **Phase 2 - Design Activities**
    - Defining and agreeing the scope of the design effort
    - Defining the requirements to inform the design activities
    - The overall TRE design in the context of the JASMIN deployment
    - The product and technology solution to realise the deployment
- **Phase 3 - Pilot Capability Implementation**
    - Planning the pilot implementation activities
    - Deploying the infrastructure and any required changes specified by the solution
    - Deploying and updating the vanilla eLab software to align with the solution and run within the JASMIN unmanaged cloud environment.
- **Phase 4- Training**
    - How to manage the TRE infrastructure. This includes:
        - Starting up and shutting down
        - Updating, backing up and restoring
        - Whitelisting the devices that are permitted to connect
        - Docker image customisation.
    - How to use the Admin and User tools, as follows:
        - Admin Tools: Includes user management (Keycloak) and log management (ELK).
        - User Tools: The tools available to end users. This includes Alfresco and the Data Science environment.
- **Phase 5 - Testing**
    - Import sample environmental and synthetic health data and analyse using Jupyter Notebooks.

What outputs have been produced

- A prototype TRE running within a JASMIN unmanaged cloud tenancy
- Report of objectives vs results (summarised in **Objectives** section above)
- Recommendation for next steps (see below)

Findings

What worked well:

The eLab pattern was installed and configured to work on JASMIN with minimal rework and reconfiguration.

What did not work:

Some owners of TREs were not enthusiastic about sharing details of their TRE design and configuration for reuse purposes, despite building these using public funds. One reason for this is that they have monetised their TRE platform and perceive other TREs using the same design as potential competitors to their income stream.

The EDS-imposed deadline of 30th September 2023 meant that it was harder to work around unavoidable resource constraints and contention with other activities, which was an ongoing issue. Having additional time would have meant that additional results could have been achieved.

Overall lessons learnt:

A Production TRE requires much more than a fully capable technology platform with all the required storage, processing, analytics, security and privacy capabilities, etc. A key part of operationalizing a TRE to make it Production quality relates to the service delivery and management processes. This is because no matter how well designed and implemented a TRE is from a purely technology perspective, the ability of a TRE to securely handle sensitive data is only as good as the team that manages the TRE, the service that they provide, and the processes that they use. Put another way, it would be very easy for a fully capable and well designed TRE to be compromised through poor management practices, bad actors, or weaknesses in service delivery.

What recommendations would you make for the next phase of EDS commissioning:

Should the EDS be interested in evolving the TRE capability further then the following steps are suggested:

- Assess the existing TRE deployment against SATRE standard, spanning the following SATRE pillars:
    - Information Governance
    - Computing technology and Information Security
    - Data Management
    - Supporting Capabilities
- Address any required improvements identified with the existing deployment by evolving the TRE pattern to ensure compliance with SATRE specification

- Include additional technical enhancements to support advanced data processing, transformation, and analytical scenarios. For instance, enhancements to determine influences and causality in connected data scenarios.
- Include additional technical enhancements to improve the TRE infrastructure capabilities.
- Invest in a new service so that Production TREs running on JASMIN can be properly supported in accordance with SATRE specification.

The EDS should develop a hybrid-cloud strategy that supports use of public cloud environments and cloud services, alongside the existing on-premise data centre environments.

- This should include both the commercial and technical aspects of public cloud use (such as how public cloud deployments will be funded) and how public cloud and on-premise deployments should be federated and integrated.
- An agreed strategy would not only provide a basis for deploying cloud-based TREs (in circumstances where this is the better option), but also benefit other projects that could utilise the wide ecosystem of ready-built capabilities available from cloud platforms.

A-9

FAAM Airborne Laboratory data use and barriers to accessibility (as part of WP4 EDS Plus funding)

Authors: Poppy Townsend and Wendy Garland (CEDA)

6th September 2023

Objectives:

List of initial objectives:

Aim: to gain enhanced understanding of the user base, their accessibility barriers and how to overcome them.

To what extent have the objectives been realised:

We have input from a small number of the user base and gathered information on barriers and suggestions for improvements.

Collaborations:

Internal to the project:

EDS specialists in FAAM data (not directly involved in project)

Externally:

FAAM staff members, Comms colleagues, FAAM users and potential users in the research community

Summary of approach:

- Advertising lists - Lists of previous FAAM data users was drawn up from searching CEDA helpdesk conversations with FAAM in, CEDA's database of users who had previously registered for FAAM data access before the dataset was made available to all registered CEDA users, FAAM scientific project mailing lists, other usual contacts (NCAS, NCEO), members of the FAAM runways project (recent project to attract new users to FAAM)
- 2 Surveys were developed and circulated - one for existing users, one for potential new users. These aimed to gather information about challenges and barriers to

accessibility of FAAM data and asked for potential solutions. There were 20 responses from existing users and 8 from new users.

- Follow up focus groups were organised to dig into the details a bit more.
- 3 staff members (CEDA and FAAM) reviewed the information and have created a list of topics/areas that need to be improved. Along with some short/medium/long term solutions/actions.
- Focus groups were held on 22nd and 24th August 2023 on zoom with 4 and 6 data users respectively, plus 2 CEDA and 2 FAAM staff. A mix of experienced, novice and future users were present, including those directly involved in specific flights and others using approaching the data as a long-term dataset.
- Dependencies - staff availability, user availability, overlaps with job role within FAAM

What outputs have been produced (prototypes, reports, papers):

- Logistics/process summary - including;
    - Templates for emails and adverts
    - List of where/who we shared the survey with
- Survey summary (including list of questions used, and additional questions we wish we'd asked). Platform for survey - pros, cons.
- Focus group summary (and list of prompt questions, logistics of running/structuring the discussion)
- Overall list of topics/areas that need to be improved. Along with some short/medium/long term solutions/actions/recommendations for next steps.

Findings:

What are your overall reflections from the work (what worked, what did not work, overall lessons learned):

Overall we gathered useful information. However it came from a small sample of users, so may not be representative. It took a significant amount of staff resource. It has sparked new ideas for improvements to data access and will hopefully help remove some barriers for this dataset. There is now a process and template advertising text that could be reused by others for engaging with users. Areas to improve next time are: more staff resource for engaging

with users, target particular audiences (e.g. attend conferences with specific interested groups, visit universities).

What worked well:

- Engaging with data providers (at FAAM) for their input into the questions asked in survey and their attendance/support during focus groups. Extremely helpful to both parties.
- Testing out new ways to engage with users.
- Opportunity to talk to real users about their struggles, gained greater understanding.
- Highlighted areas of new work to be started to start creating solutions.

What did not work:

- Short time scales of funding for this study meant we could not plan staff resource/recruit additional resource. Difficult to engage users with limited staff resource.
- Key staff members were on parental leave or tied up with other projects, so vast majority of work had to be done in August. When many people are on holiday.
- Limited responses from the user community. Smaller focus groups than expected - still a lot of voices unheard. Difficult to reach potential future users outside of our existing communities. Needs dedicated resource/experts leading this work.
    - GDPR constraints - we don't have a specific list of users when data is open.

Overall lessons learnt:

- Engaging with unknown users is very difficult. Time consuming, needs more dedicated staff resource to do it effectively
- Little support/advice from central UKRI about how to circulate messages. No central way to contact other research council areas.

What recommendations would you make for the next phase of EDS commissioning:

- Additional staff resource or dedicated staff for user engagement and/or support from central UKRI with advice/expertise about engaging with stakeholders.

- Also additional resource for comms expertise to ensure we properly realise the impact of the projects we work on.
- Longer timescales/not over the summer for these kind of studies
- Additional funding for implementing some of the solutions proposed by the focus group discussions/FAAM staff/CEDA staff e.g. creating new tools

What recommendations would you make for the wider objective of developing a commons approach for environmental assets:

- Encourage user engagement throughout everything we do. Employ specific user engagement staff to do this
- Unify engagement processes across the EDS
- Use the valuable subject area expertise of data centre staff to feed into the engagement process and improve accessibility to the datasets.

Workshops summary and recommendations/suggestions for improvements to the FAAM dataset

The original target audience for the FAAM dataset was the specialised project teams who arranged flight time on the aircraft. The dataset was designed with their level of technical and scientific knowledge in mind, and the temporal, geospatial resolution, parameter list and file format to fit. It has now built up into a long-term archive of atmospheric observations that can be useful to non-original-project researchers and a wider more general audience, however, the high level of specialist knowledge required is proving a barrier. This is likely to be a common thread across other datasets.

The data from a FAAM flight come to the CEDA archive from several sources and descriptions (where they exist) of these data products and what is in each file are in several disparate locations. Full documentation/user guide is needed to describe the data products and link to existing tools and documents.

Documentation on the layout of the files and tools to read and access the different file types (rather than the actual parameters) need to be easily available. A shared community code base could be set up.

Existing CEDA tools could be improved or extended - eg search capability in CEDA catalogue, ability to co-locate satellite footprints in the flight finder tool. GIS layering capability would be beneficial.

Suggestions for additional data products to improve usability range from the inclusion of quicklook plots, to a sanitised, lower-resolution, reduced parameter, merged and interpolated product that is uniform across all the flights to enable assimilation into models. Or the tools to produce this on the fly. This would have to be implemented at the data centre level as the flight data come from a range of external data providers as well as the FAAM team.

DOIs should be issued on flight level datasets (this is in progress - historically this has been difficult due to additions and revisions made to the dataset for years after a flight)

There was a suggestion of creating a FAAM data working group or annual engagement/networking/exchanging knowledge/upskilling event.

The ability to tell users when there is new data/revisions to data is needed - suggestion is a opt-in notification system - this requirement is likely to be common across other data collections.

The ability to comment on individual datasets was requested (would have to be moderated). This could include a process to report bad data discovered by users.

If money/effort was no problem users would like:

- Someone to produce a merged file(s) per flight containing a specific list of essential parameters pre-interpolated to a single time series, screening out bad data flagged data, and uniform across all flights regardless of changes in the archived data files.

- Or a tool to do this on the fly with user selected parameters.
- Tools to plot and manipulate the data, and output into user-friendly format

A-10

WP4 – use case 2: Imagery and derived data from autonomous and remotely piloted aerial vehicles (UAV)

Objectives:

List of initial objectives:

To develop a workflow for data generated from an example of a new technology that complies with the data commons approach of WP1 and has the potential to utilise tools being explored in WP2

To what extent have the objectives been realised:

The project developed a workflow for imagery and derived data from autonomous and remotely piloted aerial vehicles using a data commons approach. The metadata recommendations are in line with the tools explored in WP2.

Collaborations:

Internal to the project:

WP1 and WP2
Dr Wendy Garland (CEDA)
UK PDC

Externally:

NERC NetZero Airborne Capability group on uncrewed aerial systems (NZArC)
Use case study collaborators: Dr Alvaro Arenas Pingarron (BAS), Dr Tom Jordan (BAS), Dr Barbara Brooks (NCAS) and Dr Charles George (CEH)

Summary of approach (summarise how you have gone about the research, methods used, etc,):

To facilitate UAV data management, the project used a data commons approach focusing on data users and harmonisation of data and metadata. As part of this project, we worked closely with the UAV community to give them tailored recommendations on how best to manage their data. Following the lessons from Grossman, we also looked at the core issues of the data types, metadata, and vocabularies to work towards interoperability.
As part of this project, we carried out a survey and participated in different workshops to identify and define the UAV community, their needs, challenges and vision. Then, we analysed their practices to propose guidelines and recommendations for UAV data management and comply with the FAIR principles.

What outputs have been produced (prototypes, reports, papers):

2 reports:

- UAV data management handbook - Fremand, Alice. 2023 *UAV data management handbook.* UK Polar Data Centre, British Antarctic Survey, 13pp. https://nora.nerc.ac.uk/id/eprint/536392/
- NERC report - Fremand, Alice. 2023 *Towards a data commons: Imagery and derived data from autonomous and remotely piloted aerial vehicles.* UK Polar Data Centre, British Antarctic Survey, 24pp. https://nora.nerc.ac.uk/id/eprint/536398/

Findings:

What are your overall reflections from the work (what worked, what did not work, overall lessons learned):

What worked well:

Working with the UAV community and participating in UAV-focus workshops was key in understanding their needs and challenges. Undertaking this work in collaboration with the NERC NZArC programme allowed to be in line with the latest developments in terms of UAV technology and meet scientists who use UAV for their research.

What did not work:

Liaising and using tools developed by WP2 was difficult as the outputs were not developed before the end of the project.

Overall lessons learnt:

There is a big disparity between disciplines in terms of data management practices. This makes communication with scientists sometimes challenging as some won't understand basic concepts while others are ready for next steps, speaking about interoperability, machine readiness and computational infrastructure.

The data commons approach proved useful in establishing a framework for the project and better understanding the UAV community.

What recommendations would you make for the next phase of EDS commissioning:

1. Harmonise data and metadata within NERC following the recommendations of the UAV data management handbook.
2. Include recommendations in the digital stewardship wizard to promote best data management practices for UAV data.
3. Continue the development of the UAV data management handbook by adding recommendations in specific scientific domains.
4. Investigate and promote the use of vocabularies and PIDs to be used for UAV data.
5. Get involved in international groups focusing on UAV data management best practices (such as RDA, ESIP)
6. Develop strategies to best publish imagery and large size datasets and include them in the UAV data management handbook
7. Train scientists to improve their skills in data management
8. Promote the use of open software when collecting and analysing UAV data

9. Integrate data into a computational infrastructure
10. Develop API for easy access to the data
11. Propose recommendations on privacy, licensing and security

What recommendations would you make for the wider objective of developing a commons approach for environmental assets:

As part of this project, we followed a data commons approach. This had its own challenges but was a good starting point to start from scratch in an emerging field. Challenges included the definition of the concept of the commons in itself. The commons has slightly different definitions depending on your interlocutor. The WP1 workshop was key in getting a better understanding of the concept. Because of the very early stage of the UAV community in terms of data management practices, the focus was made on the community itself and the metadata harmonisation. The goal was to better understand their challenges, needs and vision. Being close to the community, be part of it, was key in providing tailored recommendations. Communication, participation to workshops and developing the survey was a first step that proved useful in developing the UAV data management handbook. I would thus recommend this consultation for each field. It was also interesting to develop different levels of metadata: UK Gemini metadata (common to anyone), UAV common metadata (common to the UAV community) and discipline-specific recommendations. This meant that there is still a core base that is shared by everyone within the community, but we acknowledge the disparity depending on the discipline or requirements of a project adding flexibility. This project was a first step, the road is still long to achieve the commons vision but would be interesting to learn more about the challenges other groups had to overcome at different stages of the commons approach.

A-11

CMIP7 Use Case (as part of WP4 EDS Plus funding)

Authors: Molly MacRae, Charlotte Pascoe, Martin Juckes

Date: September 6th 2023

Objectives:

List of initial objectives:

To understand how CMIP data is used and going to be used by the UK research community. What features of CMIP attracts people to use it and what are the obstacles that people run into when using and accessing CMIP data. How the UK research community engages in the CMIP development process. What tools and workflows would be valuable to improve CMIP data access and useability, and to capture the data requirements of the scientific community as we prepare for CMIP7.

To what extent have the objectives been realised:

We have spoken with a number of CMIP data users at various research institutions and points in their career to gain an understanding of what the community feels could be improved about the CMIP data service and obstacles to CMIP data access. The discussions were largely oriented around experiences with data from previous CMIPs and expectations for CMIP7. A question possibly missing from the interviews would have been to ask directly how we best move forward to capture the data requirements of the scientific community as we prepare for CMIP7. However, CMIP is currently working to capture this information via the CMIP task teams as preparations are made for CMIP7.

Collaborations:

Internal to the project:

Internal collaborations were between the team at CEDA

Externally:

We reached out to a total of 18 researchers across 6 institutions and 7 early career researchers. From this we interviewed 7 researchers and 4 ECRs from the UK Met Office and universities of Cambridge, KIT (Karlsruhe Institute of Technology), Leeds, Oxford and Reading.

Summary of approach

(summarise how you have gone about the research, methods used, etc,):

We initially reached out to researchers from a number of institutions that we knew had links to CMIP data to ask them to talk with us about their experiences.

The interviews lasted about an hour and we asked questions around 5 key themes:

- How CMIP data is used by the UK research community.
- Positive features that attract people to use CMIP data.
- What makes it difficult to use CMIP data.
- Opportunities for the research community to better engage with CMIP.
- Anything else they would like us to know about.

We also asked in both the initial email and the interviews for suggestions of other colleagues to contact that work with CMIP data, including specifically early career researchers. This **'snowballing technique'** helped us to reach out to a wider range of specifically selected people.

Where availability aligned we did interviews in pairs and we wrote up a report after each interview.

What outputs have been produced (prototypes, reports, papers):

We have produced write-up reports of each interview under the 5 key themes described above. That is, 7 interview reports in total and a synthesis document bringing together all the responses and common key themes. A collection of enlivening quotes, useful phrases about CMIP and JASMIN resources.

Findings:

Our overall reflections from the work (what worked, what did not work, overall lessons learned):

What worked well:

- We found rich information from just interviewing a few specific people. We also found that even with this relatively small sample size, the same comments were appearing across the interviews

- Interviews with a couple of people at a time meant they bounced off each other and shared common experiences. There was also instances where the interviewees learnt from each other about services and data available
- Asking researchers with existing links if they could suggest ECR's to talk to meant we were able to gain the perspective of people fairly new to CMIP data who faced different challenges
- People were keen to talk about their experiences and make suggestions for improvement. They were also very positive about JASMIN and the existing services provided.

What did not work:

- Short timeframe over the holiday period meant we did not have as much time as we would have liked to invest in reaching out to people and sending reminders. It also meant it was difficult to get hold of people (due to A/L etc.)

Overall lessons learnt:

- Though the CMIP community involves a diverse range of research areas, the experiences, issues and suggestions put forward from the CMIP data users we spoke to often overlapped.
- 'Snowballing technique' of asking a contact for suggestions of other contacts is an effective way to reach key people in a research network who we did not know already
- JASMIN is an essential resource for scientists working with CMIP data in the UK for both data access and facilitating data analysis
- A centralised place for technical information about models, variables would make the data accessible to more users and save users a lot of time
- As data gets large, specifically for CMIP, centralised compute centres like JASMIN becomes necessary to access and analyse the data
- Data accessible via the archive is significantly easier to access than pulling the data directly from the ESGF node
- Detailed metadata contained within the files themselves (in the CMORized Net-CDF format) saves users a lot of time trying to scrape the information from many separated resources, increasing usability of the data

Recommendations for the next phase of EDS commissioning:

- Reaching out to members new to the community (e.g. early career researchers) as well as experienced users for ideas and suggestions in shaping the development of the EDS for a different perspective.
- In person conversations with representative members of the community (including ECRs) keen to share their experiences and ideas for preparing the EDS for CMIP7.
- More community consultation within the UK about which CMIP data should be archived with the EDS. Currently variables are prioritised for the IPCC AR6 but this could be extended to e.g. the needs of large NERC national capability projects. Perhaps a comparison between CMIP variable data usage between variables archived at CEDA and variables pulled by scientists from the ESGF node to check we are archiving the CMIP data most used by the UK science community.
- A set of tools/python scripts provided at the top level for users to run quality checks on data they download to catch errors in data
- A parallel space to the CMIP data archive hosted by the EDS for the community where people share scripts for accessing data and data analysis, catching errors etc.
- Training for use of JASMIN for data analysis and best practices. In several interviews it was highlighted that they learnt by 'trial and error' or from colleagues but would like to optimise their code and analysis techniques. As well as lowering the bar for climate data analysis, this training will help scope the needs of the CMIP community to make data available accessible to more users.
- Better help documentation for pointing to model and variable information, and useful resources for first time CMIP (or more generally EDS) users. A theme across the interviews was a need for a common place for information.
- Errors in CMIP data was a common issue raised and many suggested a reporting system for flagging these errors as they are found either in a forum or at the directory level in the archive would be useful and would encourage user engagement
- Better batch download tools e.g. to download all data across models from one variable
- Funding for implementing the improvements to CMIP data access proposed in the interviews and communicating these improvements.

Recommendations for the wider objective of developing a commons approach for environmental assets:

- Bringing valuable information about data into one clearly findable place for simplicity. It seems that as CMIP is such a huge and global resource, the information necessary for using the different datasets is inevitably spread over many places making it difficult to find, particularly with analysis using a number of models.

Some nice quotes:

"JASMIN, it's like the air, if it went away it would be a real problem"

"Access by disk is by far the best solution"

"CMIP is the best thing there is for understanding climate change on a global scale and capturing the full range of model uncertainty"

"It's been transformative" (about JASMIN)

"JASMIN is enabling loads and loads of science that would not otherwise get done."

"There's a lot of data 'a zoo of MIPs' within CMIP to investigate"

"Cf conventions are fantastic"

Synthesis

The complete feedback we received from our interviewees has been collated in a synthesis report.

CMIP7 Use Case Synthesis for EDS+

Using CMIP data, improving the NERC Environmental Data Service

Authors: Molly MacRae, Charlotte Pascoe
Date: October 18th 2023

This document supports the EDS+ CMIP7 Use Case project report.

This report is a synthesis of a series of interviews that were held to understand how CMIP data is used and going to be used by the UK research community. What features of CMIP attracts people to use it and what are the obstacles that people run into when using and accessing CMIP data. How the UK research community engages in the CMIP development process. What tools and workflows would be valuable to improve CMIP data access and useability.

The interviews lasted about an hour and we asked questions around 5 key themes:

- How CMIP data is used by the UK research community.
- Positive features that attract people to use CMIP data.
- What makes it difficult to use CMIP data?
- Opportunities for the research community to better engage with CMIP.
- Anything else they would like us to know about.

Interviewees

We held conversations with a number of CMIP data users at various research institutions and points in their career to gain an understanding of what the community feels could be improved about the CMIP data service and obstacles to CMIP data access.

Experienced CMIP users

We interviewed 7 experienced researchers, who had experience of preparing and using CMIP data. Their research interests covered historical climate, decadal climate projections, European wintertime weather, aerosols and atmospheric chemistry, stratospheric ozone, regional emissions, and teleconnections. They worked with the following models: HadGEM3, EC-EARTH, UKCA, and UKESM1.

Early Career Researchers (ECRs)

We interviewed 4 early career researchers who were all users of CMIP data. Their research covered historical climate and future climate scenarios, cloud feedbacks, large scale circulation changes, atmospheric dynamics, jet stream behaviour, interrogating model biases, sulphur dioxide and aerosol formation.

How CMIP data is used by the UK research community.

Key Points

- Data accessed via JASMIN (i) directly and worked with in gws or (ii) pulled off JASMIN to home institution
- Used for model intercomparison
- Used as a benchmark to analyse typical models
- Data not on JASMIN downloaded directly from ESGF node
- The range of research areas CMIP was used for included: atmospheric chemistry, model development, historical climate, winter weather, atmospheric dynamics, model intercomparison
- Use JASMIN jupyter notebook service for data processing
- Use google drive to share work internationally to those without access to JASMIN

Experienced CMIP users

Model intercomparison and climate science

- CMIP data is used to understand and document climate model progress and evaluate climate models to build confidence in model predictions. Almost everyone who uses climate models is interested in some process or other.
- For CMIP7, interested in projections - similar analysis to CMIP6 but taking advantage of improved capability of the models
- Exploring inter-model diversity

Model development and running simulations

- Produces CMIP data for AerChemMIP
- Model development and evaluation
- NGMS (Next Generation Modelling System)
- Model development - comparing UKESM1 to observations and other models
- Lots more work to do with CMIP6, currently running regional aerosol MIP that falls between CMIP6 and CMIP7 run under CMIP6 protocols

CMIP work using JASMIN

- When the data was new, it was pulled off MASS before it even made it to the ESGF archive, not necessarily CMORised at that stage
- Relying on shared data resources these days (accessed via JASMIN) rather than downloading multiple copies of data from the CMIP archive.

- Initial analysis of CMIP6 data on JASMIN
- Do analysis on JASMIN and can point others to scripts on JASMIN when working collaboratively (in UK)
- Use JASMIN jupyter notebook service
- Get CMIP6 data from JASMIN views CEDA disk as a local disk and runs script reads data from the CEDA archive. Then outputs post-processed files to group workspace
- Do data analysis on IDL on JASMIN, and use LOTUS, noticed higher use of JASMIN has made it slower

Downloading data

- For CMIP5, had to download own data
- Eventually became more convenient to download data to local servers
- Downloaded data that wasn't part of the JASMIN subset from ESGF

Collaboration

- A lot of work ends up on Google drive when collaborating internationally to make it accessible to those without access to JASMIN

ECR's

Model intercomparison and climate science

- Used as a benchmark to analyse typical models
- Used for model intercomparison
- In future will use to analyse how well synoptic processes are represented in CMIP models - will involve wind fields precipitation fields, pressure, temp at high resolution
- Just using CMIP in next 6 months when seems appropriate
- Investigating model projections, how the jet stream might change in the future, interrogating models in terms of their biases - what they get right and wrong with respect to observed climate. Using many models with slightly different ways of representing sub-grid scale processes as a multi-model ensemble to answer science questions e.g. how does greater warming at polar latitudes affect the jet stream.

CMIP work using JASMIN/CEDA archive

- Retrieved data through JASMIN accessed from CEDA platform
- Use JASMIN group work spaces to analyse the data
- Works on JASMIN exclusively
- Made requests for a lot of specific data not on CEDA archive

Own institution computers for CMIP processing and analysis

- Data processing and analysis done on own institution computers - for ease of collaboration, colleagues not on JASMIN
- Uses shared storage at their own institution where lots of data has already been downloaded. Also uses JASMIN. Very specific data is downloaded directly from ESGF node.

Positive features that attract people to use CMIP data.

Key Points

- Metadata rich and consistent from CF-NetCDF format
- Standard file formats makes data easy to use and lots of tools available for analysing netCDF data
- CMORisation and standard structure of archive
- Easy access from JASMIN and analysis from JASMIN means size is not an issue, also easy to download subsets from JASMIN
- “JASMIN, it’s like the air, if it went away it would be a real problem”
- “Access by disk is by far the best solution”
- Also easy to download from ESGF
- Multi-model
- “the best thing there is” for understanding climate change on a global scale and capturing the full range of model uncertainty
- Standardised variables with lots of metadata, including documenting how data was converted from raw data
- CEDA archive useful as a central data store all in one place to use different data interchangeably (e.g. FAAM, CMIP, ERA-5)
- There is a lot of data available
- Easy to find the data, all on ESGF and JASMIN has large subset
- CEDA data request for specific variables has been very useful when data need that is not already on JASMIN
- - a large number of models do the same core simulations making it great for model intercomparison
- JASMIN has been transformative in enabling science
- Easy to share work internationally by giving user access to shared gws on JASMIN for collaboration
- Contact email included within files very useful
- ES-DOC very useful and available at same time as data from modelling centre (no peer review delay) but underutilised

Experienced CMIP users

Ease of use of data and tools

- CMIP data gets easier to use with each cycle
- Use CMIP data as a teaching resource (AI for ER course)
- Lots of tools for analysing CF-NetCDF data: IRIS, CF Python, cdo, matlab, jupyter notebooks
- CF Python tools are working well
- Straightforward to set up new colleagues with using CMIP data (except watching out for differences between models)
- ES-DOC very useful as documents not peer reviewed, so can be published at the same time as data and is directly from the modelling centre. But this service was underutilised
- Download scripts for ESGF were great but difficult to customise

Standardised metadata and file structure

- Standardised data makes analysis more straightforward
- Metadata structure, file formats, various tools make it easy to use
- Checked, consistent metadata in header
- Good documentation - data is shared along with necessary core metadata
- “Cf conventions are fantastic”
- Data from lots of centres in same format - CMORisation very useful and standardised structure of the archives
- Within files there is a contact email included - very useful if notice any issues or want to ask something about the data
- Quality control for CMIP6 quite good - much better than CMIP5

Valuable multi-model resource, a large amount of data for intercomparison

- Systematic, coordinated, multi-model (lots of models not just a handful), Big
- Can compare lots of different models
- There is nothing else like it
- Great tool for evaluating model and model inter-comparison
- It’s completeness - a large number of models all doing a set of core simulations and save a lot of useful diagnostics making it great for model intercomparison and quite a unique resource for this (and most also do the tier 2 and 3 experiments too)
- Homogeneity in data - as well as a number of diagnostics provided on native grids, a set of data from each model have been processed onto a common set of grids making the data easy to use
- As model developers, model intercomparison available from CMIP is very useful for evaluating their model - especially as earth system models get more complicated.

Common experimental protocol very useful, some of models in project are specifically designed to enable understanding of why models differ, figuring out sensitivity to certain parameters etc.

JASMIN computing resources used for data access

- Access to data in the CEDA archive directly via JASMIN analysis machine
- "JASMIN, it's like the air, if it went away it would be a real problem"
- Don't have to download data that is on JASMIN, can access from there
- JASMIN is a useful resource for analysing large data volumes
- Great that students have access to JASMIN - valuable lessons for using a big HPC unit
- Data on JASMIN - can use scripts to take annual or zonal means and then pull the necessary data to local storage space (rather than downloading full data)
- Useful that the Met Office have access to JASMIN
- "Access by disk is by far the best solution"
- JASMIN has been transformative in enabling science
- Accessing CEDA copy locally has massively improved workflow, as well as being able to request CEDA to download particular variables, meant only having to keep a local copy of very specialist variables. This helped with space constraints
- Easy to set up with new colleagues and share with colleagues via JASMIN - can give access to GWS on JASMIN to international collaborators to allow sharing of data

ECR's

Ease of use of data and tools

- Easy to find the data - all on ESGF, JASMIN has large subset, can just search for the data you need
- Easy to retrieve the data from JASMIN

Standardised metadata and file structure

- Net-cdf format is metadata rich, metadata fairly solid
- Standard format - all netCDF making it easy to use
- Consistent variable names and variables well documented in metadata
- Documented how data converted from raw data, makes it easier to get back to raw data to check and re-do calculations

Valuable multi-model resource, a large amount of data for intercomparison

- CMIP is "the best thing there is" for understanding climate change on a global scale and capturing the full range of model uncertainty
- Only tool like it available for their work

- Where building on work already done, can use same datasets - don’t have to do everything from scratch
- There’s a lot of data ‘zoo of MIPs’ within CMIP to investigate any question you want to look at, wide range of variables on a range of timescales

Data access

- CEDA archive useful as a central data store all in one place to use different data interchangeably (e.g. FAAM, CMIP, ERA-5)

What makes it difficult to use CMIP data?

Key Points

- Data sizes -too large, takes weeks to download, starting to only be accessible if using centralised compute centres like JASMIN
- Data on JASMIN is not all the data available
- Finding relevant information (e.g. variable definitions, understanding directory structures) without searching for relevant papers - would be useful if this info was linked to at top level page
- Difficult to pull all data for one variable - have to click through each model - better batch download tools would be useful
- CMIP data availability incompatible with IPCC timeframes and CCMI. for IPCC AR6 this meant a lot of the data used was CMIP5 and papers published with CMIP6 data only included the data that was available at the time and were written in a hurry. These quick high-level papers blocked more detailed publications on important IPCC topics.
- Heterogeneity in data - incompatible time calendars, irregular grids, errors in data - time consuming to do own QC to check data, errors may be missed when computing bulk statistics
- Issues with download from ESGF - sometimes system is down, difficult to download large amounts of data, search engine doesn't work with too many filters, have to download full 4D fields unlike data on JASMIN where you can take a subset
- Model documentation sometimes does not include essential information about e.g. processes, modules included in a findable way
- If model physics not well documented, difficult to distinguish whether there is structural uncertainty between models or it is just noise, makes some model data unusable.
- CMIP5 withdrew and replaced many datasets with issues - meant there were papers published on previous data that no longer exists

- Sometimes difficult to know whether files are missing or not produced intentionally (e.g. specific set of ensemble members)
- Disconnect between centres about standard conventions - e.g. ensemble model physics code used differently by different centres which is not well documented.
- For model intercomparison, only half models are on BADC

Experienced CMIP users

CMIP timeline inconsistent with IPCC and CCMI

- CMIP data availability incompatible with IPCC timeframes: used raw model data for historical HadGEM simulations before fully processed (CMOR-ised) data was available
- Disconnect between CCMI and CMIP, timescales don't align, running almost identical runs again
- IPCC AR6 and CMIP6 timelines not well aligned, short timeframe between when CMIP6 data was made available and IPCC AR6. This meant some topics that were very important to IPCC were done in a hurry rather than in depth to meet the deadlines and only included the CMIP6 data available at the time. This also meant these shorter high-level papers published quickly blocked more detailed publications using the full CMIP6 datasets. AR6 ended up mostly based on CMIP5 because the deadlines were so tight.

Space issues with large data

- CMIP data "chunks" can sometimes be too large for direct use in analysis tools
- Worry that CMIP7 is going to be bigger
- Large data volumes are borderline incompatible with being able to download for local analysis.
- Data space - improved by access to JASMIN
- Memory limitations that apply to the jupyter notebook service sometimes makes exploratory analysis difficult - unclear how much memory is available and how much has been used, kernel unexpectedly dies. Some analysis too large for these notebooks to cope with

Data access issues

- Forgetting that the CMIP data that can be accessed via JASMIN is only a small part of the full CMIP dataset
- On the web interface, it is easy to find a particular field from a particular model but more difficult to gather all model data for one variable at once, much easier to just pull off JASMIN. Metagrid - rewrite of interface to make it easier coming soon (aims2.llnl.gov)

- Not always straightforward to know what data is on JASMIN for each model, wrote own script to find this out in the end.
- When workspaces get shut down, have to re-download the data all over again

Data availability

- Don't have all the variables you want from some models
- For model intercomparison, only half models are on BADC, have to download the data for the rest from the ESGF node - takes a lot of time and space
- When downloading from ESGF node, have to download the full 4D field data from each model (which becomes very large), if the data was on JASMIN you could just take zonal means and pull that to your GWS which saves a lot of space and time
- International collaboration is difficult, a modelling centre might have done a set of experiments but difficult to collaborate when the data has not yet been mirrored

Information about data difficult to find

- Struggle to find various variable definitions without papers that specifically describe them. Collection here: metoffice.github.io/arise-cmor-tables/ (not always clear what the variable is from the longname alone).
- Lots of different behaviour in CMIP for sampling structural uncertainty - but if model physics not well documented then can't tell if this is structural uncertainty or just noise - have to leave models out when information not clear/available
- Disconnect between some centres on how they use the standard names e.g. physics code in ensemble members used differently by each centre which isn't well documented. (For example, the GISS model has 3 physics versions indicated by p1, p3, p5 which should be analysed separately but they are all under the same model name. This information is explained in model paper but not in files themselves and may be easily missed in analysis.) Needs to be well documented or agreed that runs with very different physics need different model names e.g. CESM has different model versions for different physics.
- ES-DOC underutilised - difficult to find specific information from each modelling centre - what processes are included, modules included, etc. CMIP6 documentation is better than previous CMIPs but some models just cite previous CMIP papers as documentation. Model papers are held up by peer review - particularly for CMIP5.
- Sometimes it is difficult to know whether files are missing or were not produced intentionally. Would be useful to know which ensemble members a modelling centre has chosen to produce (e.g. just r1, r3, r5 etc.) documented in file or readme's

Data processing

- Inconsistency between number of dimensions between variables, e.g. not all single level coordinates have z dimension, makes it more difficult to re-use analysis code

- Converting model output to CMIP data format is challenging, would prefer if there were more configurable tools for doing this.
- “The hardest thing (after building the model and making forcing files) is converting the data”
- Data processing - have all tools needed but had to learn best ways of working through trial and error, would be great to have training on optimising data analysis techniques

Data inconsistencies

- There is some data that doesn’t follow the established norms, coding exceptions to account for that can be time consuming (“hours and hours”). E.g. met office 360 day year.
- Quality control - some of the data is clearly wrong
- Errors in data that are not flagged, it would be useful if it could be communicated when data is updated to people who have already downloaded the data
- CMIP5 - lots of centres withdrew datasets meaning datasets were being withdrawn after use in publications - quality control was better for CMIP6

ECR’s

Data inconsistencies

- Heterogeneity in data, issues or corruptions - mismatched time calendars - would be ideal if there was an agreed upon process for mapping that data to a consistent shared calendar format
- Poor quality control - even commonly used z500 one model had exactly same field for every year in January for every year, random days where geopotential height was a billion, relative humidity way over 100% in the arctic
- Extensive process of quality control to do which takes a lot of time
- Computing bulk statistics you are going to miss errors in data if you don’t take the time to perform QC yourself
- Understanding where there are heterogeneities - currently get that info from other people who have encountered the problems

Information about data difficult to find

- Understanding directory structures, e.g. some models have different grids - in directory structure but not clear
- Experiment ensemble members - base member not always clear - may be 2 if 1 was a bad run - you have to go through each model individually to check which is the right one - would be good if it was clear when you download the data

- Takes time to find model documentation and specific papers with info about the data - a centralised hub of information would be useful
- Finding model data for a given variable takes a lot of time
- Reproducibility - ‘data can be retrieved from the ESGF’ on papers there is not sufficient information to reproduce the data

Space issues with large data

- Data size is a problem - solution is centralised compute - took several weeks to download all the data, file corrupted by a second of lost connection#
- Could be a problem with size e.g. for undergraduates downloading data on their own laptop - as long as you download the data wisely is ok. Data storage has not been a problem when using JASMIN and shared storage. Can combine files and take averages rather than downloading everything

Data access

- Central compute sometimes painful compared to own institution
- Jupyter notebooks don’t work well over remote connection - makes centralised computers difficult
- ESGF download causes problems, the tools available are not easily accessible
- downloading large volumes of ESGF data is difficult
- ESGF portal sometimes doesn’t work - would be helpful if there was a message to notify when the service is down
- On ESGF search tool, if you add too many subset filters it doesn’t give you any results when the data is there
- Would be useful if there was a better way to do a batch download of data - takes longer than it should do, or more clearly advertised tools to do this

Opportunities for the research community to better engage with CMIP

Key Points

- Via own networks
- Individual model conferences - can speak to contacts in community that feed back to overall CMIP process
- Fresh Eyes on CMIP for ECRs
- Add a 'getting started' page
- Add links to useful pages and documents at top level page
- CMIP process feels big and overwhelming, not sure how as a data user I can influence it
- CMIP7 workshop on UK contributions to CMIP7 has been well advertised

- good public consultation from CMIP IPO
- ESGF help emails
- Lowering bar for climate data analysis - model specific info made available in central place so data is more accessible
- Need training in out of memory data analysis

Experienced CMIP users

- Would engage with CMIP via own networks in the first instance
- How is the data request task team reaching out to different communities?
- Add a link to the CMIP IPO information pages from the EDS website.
- A getting started page (info about useful resources and how to find them)
- Point to pages/tables that clearly show what data is available at the top level (ESGF CMIP6 Data Holdings (llnl.gov)
- Need training in out of memory data analysis, best practices, the infrastructure available e.g. optimal chunking
- Lowering the bar for climate model data analysis - model specific information made available in a central place or forum so that the data is more accessible to lots of different users
- CMIP7 workshop next week - how UK contributes to CMIP7 - well advertised
- CMIP International Project Office - has good public consultation
- IPCC AR6 involvement with CMIP6
- Running regional aerosol MIP that falls between CMIP6 and CMIP7 timeframes but is still listed with CMIP6 IPO as being a MIP
- ESGF help emails

ECR's

- Fresh Eyes on CMIP task team https://youtu.be/URXeIUP2zPY https://wcrp-cmip.org/fresh-eyes-on-cmip/
- Attending individual MIP conferences and speaking to contacts in the community who feed back into overall CMIP process
- Not really aware of opportunities to shape the process and not sure how as a data user they can have much input, feels very big, overwhelming
- Interested in shaping particular MIPs used for their research
- Incremental changes - models get a little bit better each round of CMIP but just limited improvements, maybe investing the resources in running less models at higher resolution - this is a challenge for CMIP moving forward

Other stuff

Experienced CMIP users

- How many people use the web interface vs. pulling data directly from ESGF with computers like JASMIN?
- MIP table development: working on a MIP table setup resource that would separate out generalised MIP table content (stuff you can probably leave alone) from project specific information (stuff you need to tweak).
- Advocate for more comprehensive model documentation - like was done for CMIP5 by metafor
- Recommend stringent QC for core diagnostics from the DECK
- Suggested a web tool for running own QC checks on example files.
- Concerned that increasing data volumes may become limiting
- A table that shows what CMIP data is on JASMIN and a search tool would be useful
- Problem of data sizes and data replication is a challenge everywhere with multi-centre science
- Some kind of communication that data has been updated where errors have been found would be useful e.g. an email sent out to all who have accessed the data, a forum
- ES-Doc - there is process for reporting errors in data, but it it quite separate from the directory where the data is
- Training on optimal practices with data analysis, efficient data processing
- cdo and nco are good tools but are 'black boxes' and quite fragile, if data is formatted slightly weird may not read your file, would be good for more people to have more control over data processing - training or shared forum would be useful
- A github for everyone to share their data analysis processes
- High memory jupyter notebooks would be great
- Would like a tape option for UKCA data once initially analysis of data on the disk is finished but would still like to access this data, feels wasteful and time-consuming to have to re-download it all over again
- GPU cluster being released - dedicated training on how to use this would be helpful
- Need to know - what emissions have been put into the model - is the model emissions driven or concentration driven. Technical information about certain schemes - what processes are included. Knowing the level of complexity is really important for understanding why the models behave the way they do
- When find issues with data - check with modelling centre first then report it
- Needs to be a balance between quality control slowing down data availability even more and improved quality of initial data. Experienced users have their own checks now to sanity check the data.

- Better community consultation about which CMIP data should be archived on CEDA - variables prioritised for what AR6 would need but might be good to extend further to e.g. large NERC national capability projects and what they might need
- Interesting to see what's being used by local users and what is being downloaded directly from ESGF to analyse which data would be best kept on CEDA

ECR's

- Add flags to certain files that have issues raised, (e.g. geopotential height is wrong) so that people can see this warning when they download the data and don't have to work it out individually
- Maybe create a small selection of python scripts to run with files you want to use want that flags the errors that have come up with that data
- Can more information be in the CMIP webpage of the data collection or a top-level folder?
- Interested if any outputs of storm resolution models destiny, destination Earth, next gem and dyamond high resolution global runs are going to be archived on CEDA as part of next round of CMIP? Is parallel to CMIP and going to be engaged with in the coming 5 years
- Looking at CMIP data on CEDA archive Project Record: WCRP CMIP6: Coupled Model Intercomparison Project Phase 6 (ceda.ac.uk) you have to click through each model individually - would be useful to have all in one place which data is contained by each model a for each field and resolution
- A script to collect all models with data from a certain variable
- Parallel space for the community to submit scripts that CEDA could host where people say what worked for them and share scripts for accessing data
- Even one script example of how to retrieve e.g. 'the historical projections of surface temperature' from across the data, a script to collect one variable at one resolution from the data as an example to be used by others
- Too many models to cite individually - when referencing, thank modelling community as a whole and ESGF
- Would like pressure level ERA5 data on CEDA
- More public engagement to make people aware of the data available and the tools available to download it

B

WP1: Commons Roadmap – Workshop Summary

The following notes are the outputs of a series of constantly changing, cross-cut group processes that took place over this intensive workshop which ran during two half-days in May 2023.

The above image shows some of what the participants identified at the start of the session that they wanted to focus on, that by the end of the workshop they believed they had. The final image on page twelve shows what those attending feel still needs to be focused on or addressed.

The following three sets of bullets is a capture of what small groups identified the roadmap needed to include, categorized broadly into three achievability/time frames.

Early Quick Wins:

- Consensus on what we are trying to achieve
- Understand and communicate where we are now, and what is the problem with that?
- Identify best practice/literature review
- Identify cross-domain case studies
- Find working examples in other communities
- Allow users of data to chat/message each other for support, to provide feedback on quality, usages etc. (ongoing stakeholder engagement)
- Training modules for users
- Use cases to demonstrate benefits from groups that have been successful and who may be further on in the journey e.g. Australia
- Use Hamish's slide of people, skills, tools etc. (Don't reinvent)
- Collaborate with Hamish et al :)
- Invite Hamish to visit us next year - ask him to be our Engagement and Integration lead!
- Keep the momentum
- WG lead reviews and agree discovery metadata semantics and structure
- Plan user engagement
- Community consultation (c.f. ARDC process) to focus EDS roadmap
- Produce a roadmap

- UKRI research community and data services review to discover the data assets we have (c.f. ARDC maturity) including people, standards, tools…
- Define the boundary of our commons e.g. sensor data
- Executive sponsorship behind roadmap
- Prioritise the four elements of a data commons 'thin middle'
- Abstract common infrastructure needs
- A draft set of questions to know if you are NERC commons compliant (accreditation)
- Signpost other sources of environmental data from each of the current data centres
- Write comms plan (not implement)
- Employ a comms expert who has experience in the area to communicate what EDS is, what they're doing and how great it is!
- Put roadmap milestones in an agreed order
- Have a shared vision/ elevator pitch. All communicate it to stakeholders, record reaction and put it on Slack

Achievable with a little more time/resource:

- Identify management
- Review current metadata standards (w.r.t. EDS use) and implement improvements in standards and associated tools (i.e. editing tools and catalogues)
- Process for funding business cases
- Mechanism for conflict resolution and consensus building for equitable resourcing etc.
- Process for deciding tools and technologies
- Scope agreed from requirement engineering activity
- More money for proto platform
- A prototype or working example of what we are working towards
- Decisions to be made on consultation processes – gap and requirements
- Understand our users (registration?)
- Identify communities
- Engage community for vision of commons
- Identify and 'group' our stakeholders where they have common problems
- What does the commons care about? Agree scope of asset commons
- Standardized catalogue of EDS services
- Integrated spatial data catalogue (data.gov with map!)
- Core metadata improvement/ standardisation
- Semantic interoperability of datasets in catalogue
- Develop vocabulary that is shared across the sub domains of the commons
- Establish federated access point to EDS vocabularies
- Get others on board
- Bring people together. Engage hearts and minds to inspire a shared vision

- Share use cases/examples
- Develop networks/ relationships to have cross-discipline/domain conversations
- Collaborative commons to enable stakeholder engagement (e.g. via data labs coding club)
- Establish a pilot project for EDS – wide variable search
- Carbon credits for use of DRI services
- A Skills Gaps analysis – social infrastructure/ governance infrastructure/ economic and political infrastructure to enable success
- Architecture landscape ⇒ new thing or integrating?
- Ongoing training and skills
- Collaborate globally for common principles to develop commons
- Single sign on

Down the line ambition:

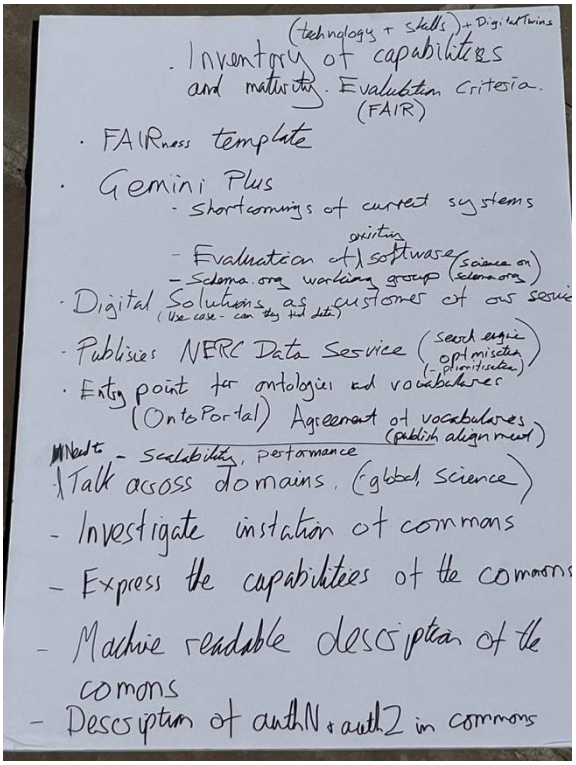
- Programme management defined
- Search for narratives across EDS
- Promote adoption of commons – UKRI, nationally and internationally
- Co-ordinating with other UKRI/gov/industry initiatives
- Getting to net zero
- Net zero by 2030
- Data lake to go across domains/institutes
- Team profiling, skills identification, recruitment
- Semantic inference and question answering from datasets (i.e. all knowledge expressed in ontologies that are aligned)
- Federated commons across UKRI
- Federated API
- Federation and integration of standards
- Integration across commons: UKRI? Australia?
- AI ready data commons/service 9e.g. Chatbot interface to EDS)
- Metadata enhancement ⇒
- Provide support for near real-time data
- ⇐ adequate ongoing funding ⇒
- ⇐ Iterative development ⇒
- ⇐ Training and support for users ⇒

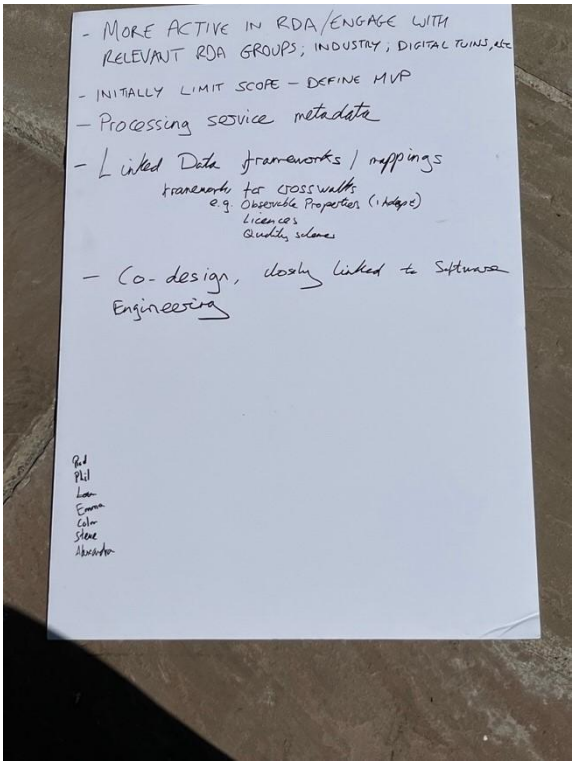
The following images reflect early thoughts from six task groups which emerged from the Achievability/Time exercise.

The task within each self-selected group was to identify project deliverables or clear practical actions to bring this section of the roadmap to life.

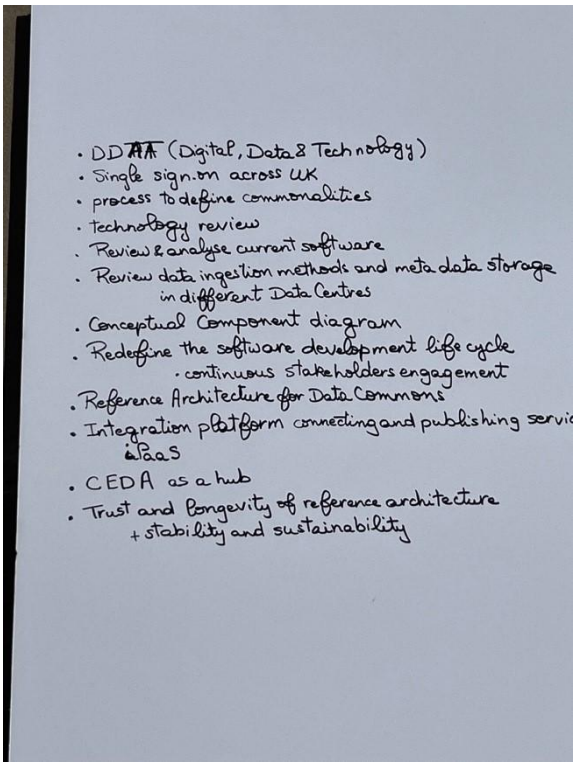
The groups were (i) Metadata and Federation (ii) Software Architecture and Commonalities (iii) Communities of Practice and Stakeholder Engagement (iv) Governance (v) Vision and Comms. and (vi) Training and Skills.

Metadata and Federation

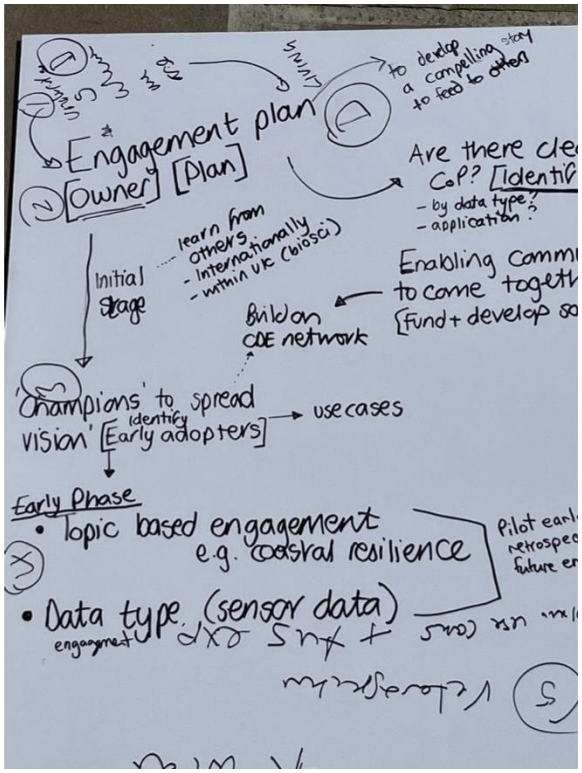

(technology + skills) + Digital Twins
. Inventory of capabilities
and maturity . Evaluation Criteria.
(FAIR)
· FAIRness template
· Gemini Plus
- shortcomings of current systems
- Evaluation of existing software (science on schema.org)
– Schema.org working group (schema.org)
· Digital Solutions as customers of our service
(Use case - can they find data)
· Publishes NERC Data Service (search engine optimisation (prioritisation))
· Entry point for ontologies and vocabularies
(OntoPortal) Agreement of vocabularies (publish alignment)
- scalability, performance
Talk across domains. (global science)
- Investigate instation of commons
- Express the capabilities of the commons
- Machine readable description of the commons
- Description of authN + auth2 in commons

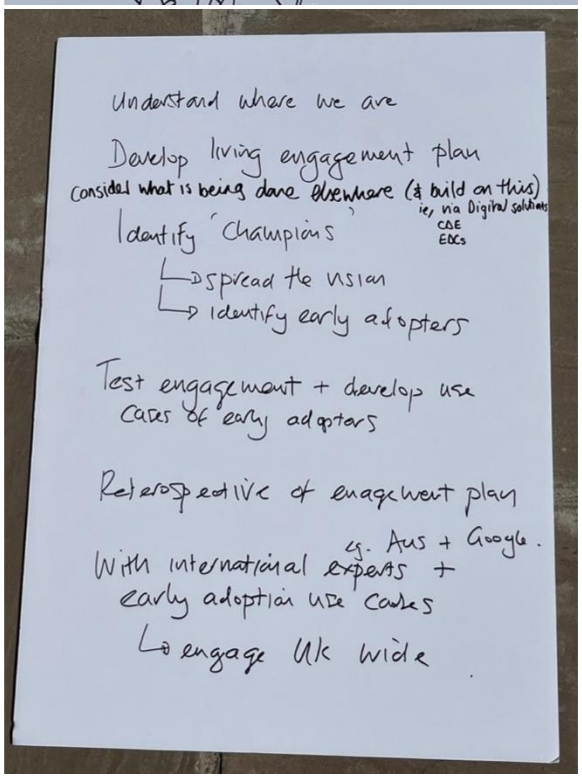

- MORE ACTIVE IN RDA/ENGAGE WITH RELEVANT RDA GROUPS, INDUSTRY, DIGITAL TWINS, etc
- INITIALLY LIMIT SCOPE – DEFINE MVP
– Processing service metadata
– Linked Data frameworks / mappings
frameworks for crosswalks
e.g. Observable Properties (+Adopt)
Licences
Quality schemes
– Co-design, closely linked to Software Engineering
Rob
Phil
Leo
Emma
Colin
Steve
Alexander

Software Architecture and Commonalities

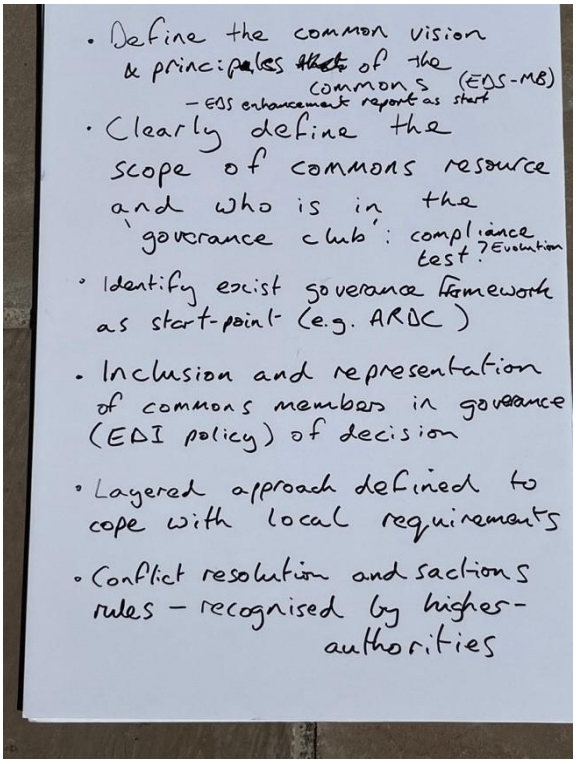


Communities of Practice and Stakeholder Engagement

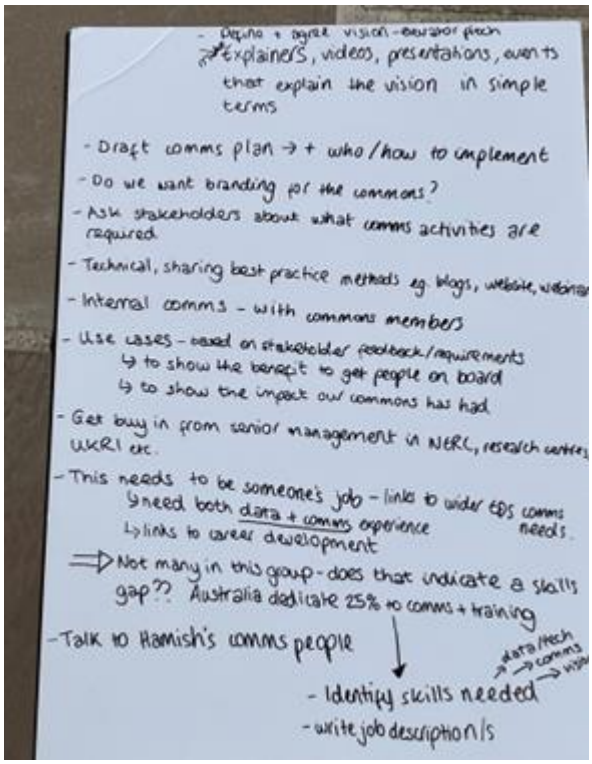

Engagement plan
[Owner] [Plan]
to develop a compelling story to feed to others
Are there cle
CoP? [Identif
- by data type?
- application?
Initial stage
learn from others
- internationally
- within UK (Blossc)
Enabling comm
to come togeth
[Fund + develop so
Build on CDE network
'Champions' to spread vision' [Early adopters]
identify
use cases
Early Phase
• topic based engagement
e.g. coastal resilience
Pilot earl
retrospe
future en
• Data type. (sensor data)
engagment


Understand where we are
Develop living engagement plan
consider what is being done elsewhere (& build on this)
ie, via Digital solutions
CDE
EDCs
Identify 'Champions'
→ Spread the vision
→ Identify early adopters
Test engagement + develop use cases of early adopters
Retrospective of engagement plan
With international experts + early adoption use cases
eg. Aus + Google.
→ engage UK wide

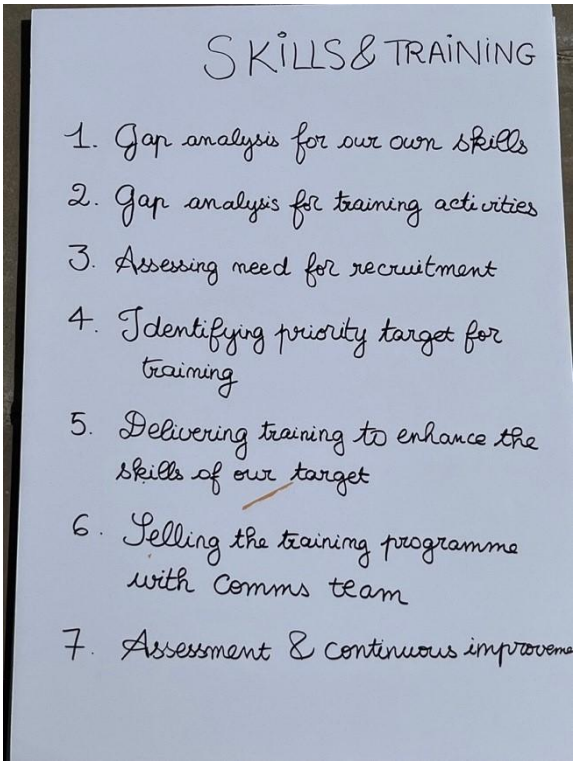
Governance



Vision and Comms.



Training and Skills


SKILLS & TRAINING
1. Gap analysis for our own skills
2. Gap analysis for training activities
3. Assessing need for recruitment
4. Identifying priority target for training
5. Delivering training to enhance the skills of our target
6. Selling the training programme with Comms team
7. Assessment & continuous improvem


LOW WOOD BAY


Discovery Metadata Standards
Element Semantics
Mark up Ontologies
Colm Walsh
Cross-domain Semantic Interoperability
RDA
Gwen
Software Architecture
Maria
Influence UKRI [ORI]
Identify challenges
Potential
Andrea
Big picture talking + writing
vision bringer!
John
Data Publishing
Petra
EDS Historical perspective + some tech stuff
Sam
Geoscience domain knowledge + user needs
Emma
Architecture Data Management Practices
Andy
domain knowledge (geophysics) - international engagement with science community
Alice
Project + people management
Task focussed & getting stuff done
Kate
Knowledge of the polar data landscape
Questioning mind
Helen
Real_time data
Sensor metadata
Alex Tate
Sensor networks, software architect / product management
Liaising with users
Lou
Semantic interoperability
FAIR Semantics
FAIR Data
Vocabularies
Ontologies
RDA VSSIG
Cross-domain interoperability
Sensor metadata
Alexandra
Software Engineering for scientific applications
Technical Leadership
James
Resource Manager
Coordination of programme
Sue
Non core users, non academic non NERC
Richard
Clear messaging, translating, bringing people together
Poppy
Data quality management + development
translation between users and developers
Philip
developing metadata systems and data APIs
Rod
Standards for metadata, vocabs, APIs, domain knowledge, ontology, ethnography, software engineering
Rachel
Large EO datasets - metadata standards/ services
Steve D
Gordon

And finally, those attending still want to see: