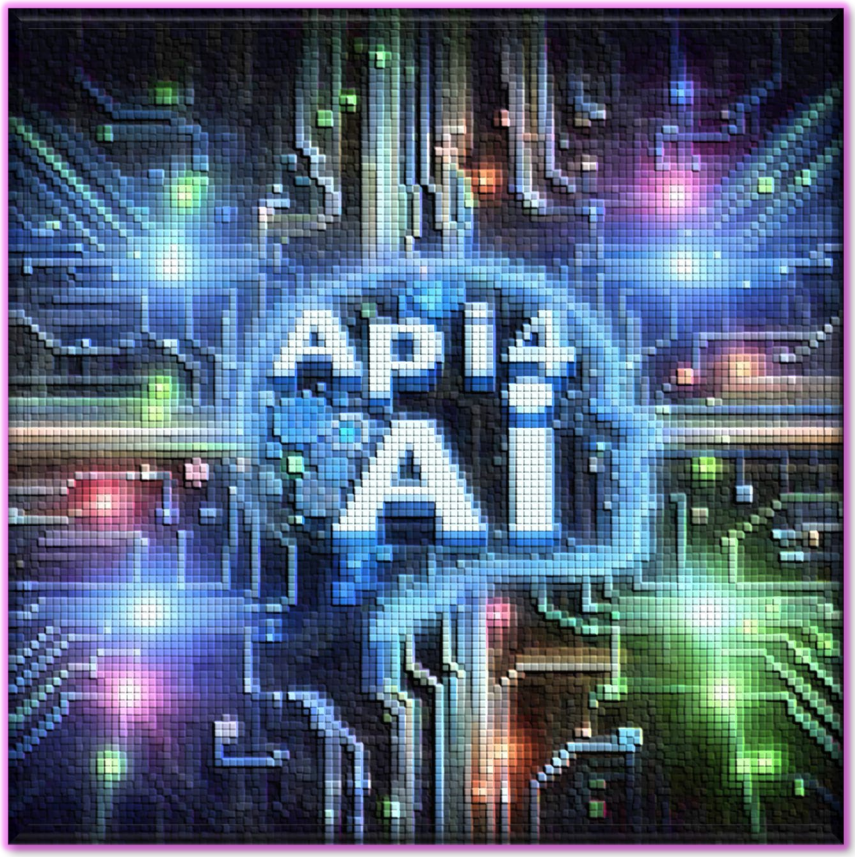
NERC Environmental Data Service

NERC Environmental Data Services: API 4 AI Project Kickoff Workshop Report



BGS, Keyworth Thursday 28th November 2024

Authors: Jonathan Booth[1], Patrick Bell[1], Andrew Kingdon[1], Rachel Heaven[1], Chris Card[1], Colin Sauze [2], Michael Tso [3]

Contributors: Emma Bee[1], Matt Cazaly[2], Tom Gardner[2], Edward Lewis[1], Matt McCormack[2], Eric Orenstein[2], Jon Cooper[3], Tim Barnes[4], Paul Breen[4], Thomas Zwagerman[4], Dave Poulter[5], Ag Stephens[5]

[1]British Geological Survey, [2]National Oceanography Centre, [3]UK Centre for Ecology and Hydrology, [4]British Antarctic Survey, [5]Centre for Environmental Data Analysis

NERC Environmental Data Service

1 Contents

2 Summary

The NERC Environmental Data Service (EDS) provides a focal point for scientific data and information spanning environmental science domains: atmosphere and climate; earth observation, polar and cryosphere; marine, terrestrial and freshwater; geoscience, and solar and space physics. Improving access to quality-assured, high-resolution environmental information and associated software tools will enable new users and communities to access environmental research by facilitating integrated analyses of environmental processes in response to societal challenges and environmental change.

The wholesale availability of Machine Learning (ML) and other Artificial Intelligence (AI) technologies is changing the way that environmental data is being used. They enable a new scale of processing and allow the identification of trends and signals in data streams that were previously impossible to identify or practically impossible to deliver due to a lack of computing power. Application Programming Interfaces (API) are a core enabling technology for AI as they provide a machine readable connection between programmes / algorithms to connect directly to other programmes and crucially data automatically without human interference. Widening the availability of API-supplied data will deliver new functionality for AI capabilities.

This project aims to share experience in creating and applying standardised APIs for utilisation in AI workflows. This will widen data access to environmental researchers and enable systematic AI analysis of multiple environmental data types to underpin the development of predictive environmental modelling and digital twins.

The project received UKRI Digital Research Infrastructure Programme funding through the opportunity entitled ‘Enhancing digital research infrastructures by trialling approaches to skills and software’. This funding provided resources across the British Oceanographic Data Centre (BODC) hosted at National Oceanographic Centre (NOC), the Environmental Information Data Centre (EIDC) hosted at UK Centre for Ecology & Hydrology (UKCEH) and the National Geoscience Data Centre (NGDC) hosted at British Geological Survey (BGS) with a deadline of 31st March 2025.

This workshop brought together staff from those organisations as well as representatives from the Polar Data Centre hosted at British Antarctic Survey (BAS) and the Centre for Environmental Data Analysis (CEDA), part of the National Centre for Atmospheric Science. The purposes of the workshop were many fold:

- To introduce staff working on the project to generate a team spirit to be carried through the project and identify common objectives and priorities
- To establish a common baseline agreement and understanding of the aims and objectives of the project
- To demonstrate previous work that was relevant and could be utilised within the project

- To begin establishing science use cases that would be used as exemplars to structure the provision of data APIs for utilisation in AI workflows
- To begin establishing the AI workflow technologies that would be utilised in the project
- To begin establishing the data APIs that would be required to address the selected science use cases
- To collaboratively co-design the project and identify its deliverables, a project timeline and participant responsibilities
- To start generating ideas for further work to build on the outcomes of this project that can be put forward in response to future funding opportunity calls

The full agenda for the workshop can be viewed in Appendix A – Agenda. The day was structured into three main sections, opening with introductions and a series of knowledge-sharing presentations across the organisations. These introduced current AI workflow technologies being utilised; the NOC Data Science Platform (DSP) and the UKCEH Datalabs platform. In addition, BGS provided an overview of their standards-based approach to API development.

The second part of the day focused on a workshop activity to identify science use cases and the AI workflows and data APIs needed to address them. This resulted in 25 use cases with ideas from across all organisations. The results of this exercise can be viewed in Appendix B – Use cases

The final part of the day encompassed a second workshop activity that generated a project timeline to address the identified project deliverables that built towards successfully developing and integrating the required data APIs and AI workflows to meet the chosen science use cases. This can be seen in Appendix C – Timeline

3 Background

3.1 Proposal Summary

Vision: Artificial Intelligence (AI) and Machine Learning (ML) workflows are increasingly used to undertake data science analysis to answer environmental questions. This proposal focuses on creating and utilising data-driven APIs that power workflow pipelines and thereby make data driven AI/ML easier to achieve. Such software technology will enrich the NERC Environmental Data Service (EDS) to enable users to explore and ingest AI of environmental data for research and to ultimately gain greater insights from this data to support societal decisions. Data-driven APIs are fundamental to the provision of an AI-powered integrated Digital Research Infrastructures. This proposal aims to develop consistent approaches to API design, development and deployment across EDS, sharing expertise and experience. This enhances both the efficiency and consistency of API provision and will enable integrated access to cross-discipline scientific data to meet existing and new use cases from collaborators. It will widen engagement in cross-discipline proposals, thereby upskilling across environmental sciences. It will enable new multi-disciplinary scientific analysis and lower the barriers to entry. The proposal will look beyond the EDS to ensure approaches are consistent with other initiatives/standards across the UK and internationally.

The project will enable:

- Better Data. All EDS data centres need to provide accurate, accessible and interoperable machine-readable data. This underpins the ‘strong foundations’ of environmental data for the UK.
- Better sharing of data to ensure better access to data across and beyond the environmental sector and through facilitating automation of these capabilities, and through enabling interoperability.
- Better analysis of the data. We need to lower the technical barriers to environmental data analysis, preventing the need for additional, often expensive, resources or specialist skills.

3.2 Aims

3.2.1 Creating APIs for use in AI/ML pipelines

A documented best practice approach for the creation of APIs with exemplar code will be created. EDS data centres will collaborate to develop best practice and implement consistent approaches to API development: API design to meet user needs and non-functional requirements; architectural patterns and development; data object management and pipelines; deployment; documentation. The partners will ensure APIs developed across disparate data centres are consistent, predictable and based on recognised standards (e.g. OpenAPI for specification/documentation and OGC API family of standards). Previous work on implementing common environmental data APIs using OGC frameworks was undertaken by BGS along with colleagues in the Marine Environmental Data & Information Network (MEDIN). This

highlighted the need for a standardised approach to implementation to support simple end-use applications.

The project will ensure APIs are quick/easy to implement by using rapid application development tools and will provide the boilerplate code, libraries and reference implementations that handle the architecture and repeatable components. Individual API providers will only need to code the parts that are specific to their dataset and specialist requirements. We will focus on APIs with simple payload data structures that can be easily and widely used by a wide range of AI programmers as well as ensuring compatibility with data from other providers and domains.

3.2.2 Creating an exemplar 'Data Science + AI Platform'

This will demonstrate that APIs provided by EDS data centres can be consistently utilised in AI/ML workflows. This will utilise NOC's Data Science Platform (DSP) and UKCEH's Datalabs, building on existing investments and creating value-added capabilities which are deployed in JASMIN (a national DRI capabilities infrastructure) to:

- Provide AI data pipelines/engineering
- Develop efficient user experience/design to facilitate the use of AI for environmental data
- Demonstrate how APIs can be integrated into AI/ML workflows to answer environmental use cases.

DSP access will be widened for users/communities beyond EDS institutions.

3.3 Workshop Goals

The workshop was designed to bring together technical experts and environmental data users to explore and co-design potential use cases that can demonstrate the potential for API-powered AI/ML workflows for environmental data.

3.4 Participants

The workshop included participants from the following EDS organisations and representative data centres: British Antarctic Survey (BAS), British Geological Survey (BGS), Centre for Environmental Data Analysis (CEDA), National Oceanography Centre (NOC) and the UK Centre for Ecology & Hydrology (UKCEH). This document refers to EDS data centres' affiliation via their host bodies.

National Oceanography Centre

BGS British Geological Survey

UK Centre for Ecology & Hydrology

Centre for Environmental Data Analysis
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

British Antarctic Survey
NATURAL ENVIRONMENT RESEARCH COUNCIL

Figure 1 Organisations

Each organisation was represented by a different number of participants spanning different roles.

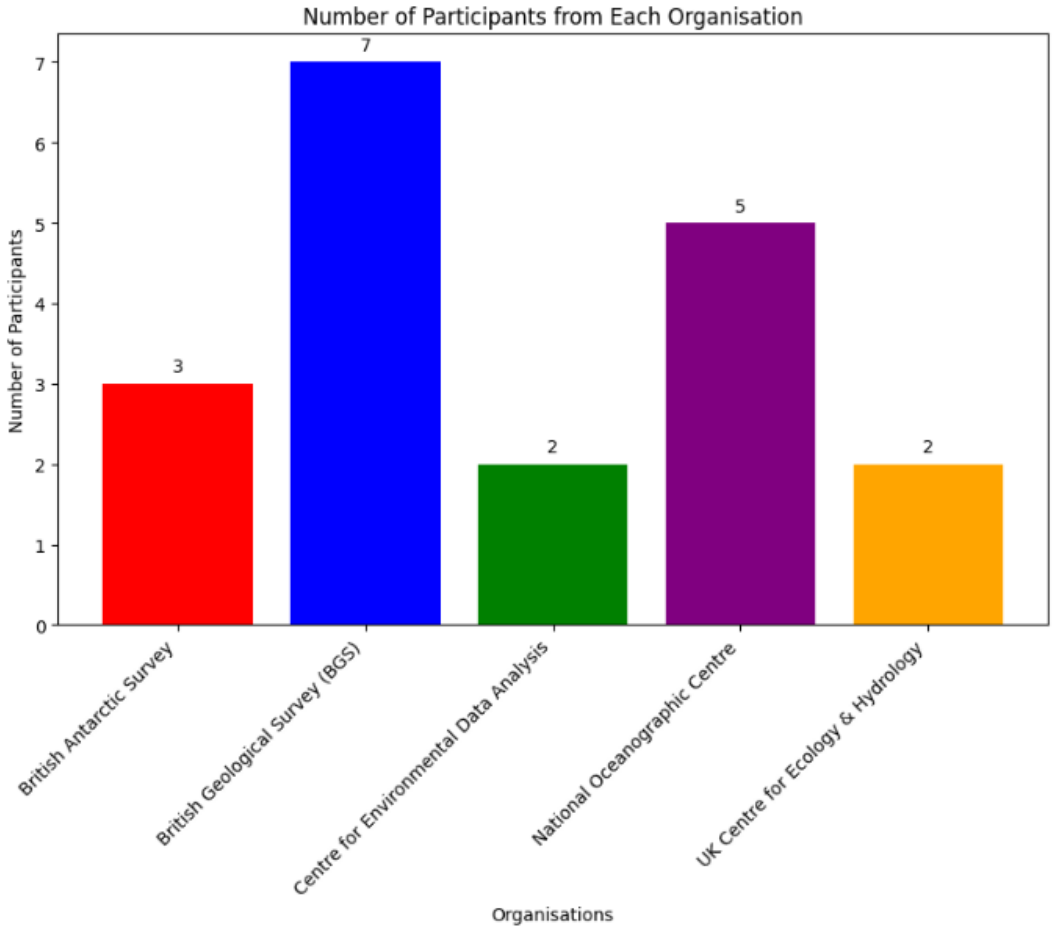


Figure 2 Participants by organisation

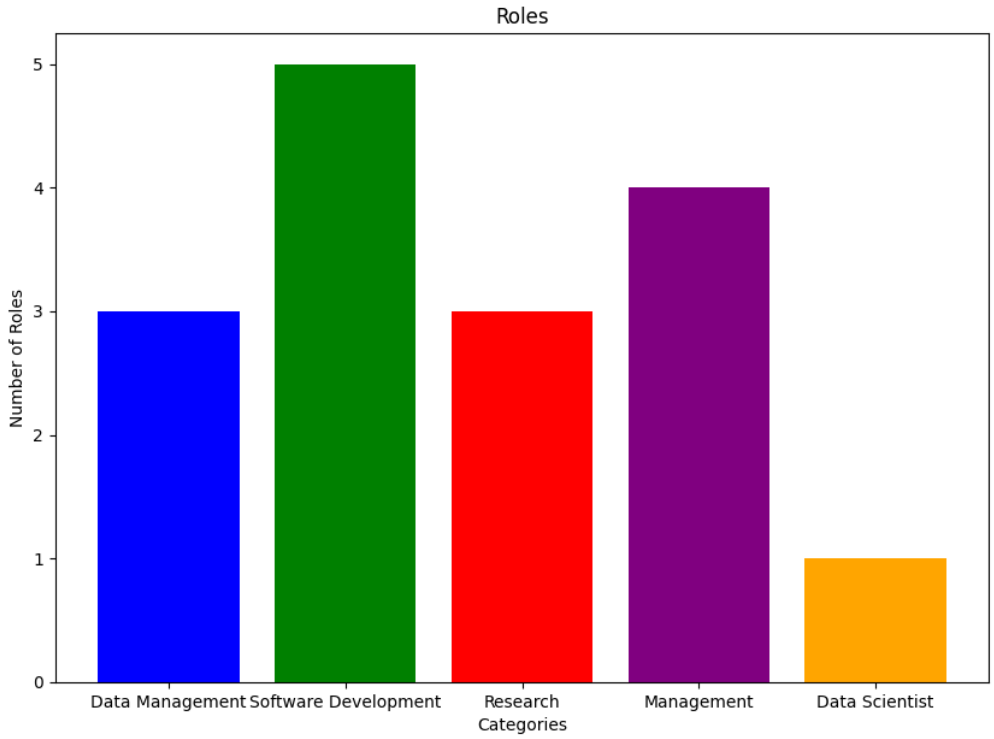


Figure 3 Participants by role

3.5 Documentation

A SharePoint space that all EDS organisations can access has been created.

This workshop report represents a snapshot of the project defined in the workshop and subsequent BGS timeline meeting. Any further updates will be documented via SharePoint.

4 Workshop

4.1 At a Glance

The workshop occurred in the BGS boardroom in Keyworth on Thursday, 28th November 2024.

Figure 4 BGS Boardroom

It consisted of a knowledge share with presentations covering the NOC Interactive Data Science Platform, UKCEH Datalab platform and proposed API standards for AI data from BGS.

Figure 5 Presentations

This was followed by a workshop activity to explore and identify AI use cases and a second activity to propose a project timeline outlining how identified deliverables would build towards demonstrating how APIs can be integrated into AI/ML workflows.

Figure 6 Workshops

Figure 7 Creating the timeline

The full workshop agenda is provided in Appendix A. Following the workshop, BGS reviewed and clarified the project deliverables and timeline generated during the workshop. This version has been shared with project partners for their consideration and agreement.

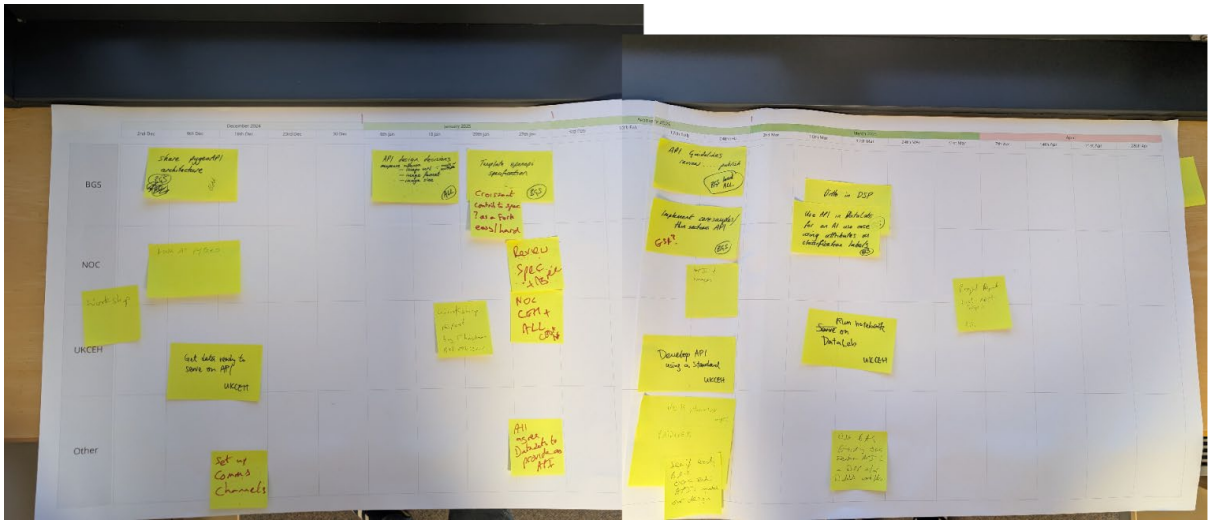


Figure 8 V2 Timeline

A detailed view of the current timeline can be viewed in Appendix C – Timeline.

4.2 Knowledge Share

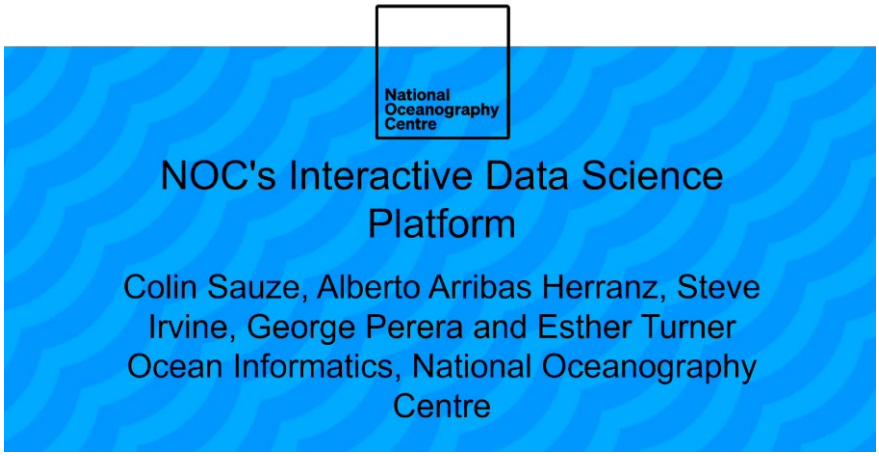
This session focused on understanding project partner capabilities relevant to achieving the project's goals. Presentations were provided on the two AI platforms previously mentioned and on proposed API standards for integrating data into these AI platforms.

Figure 9 Presentations

Each presenter had 30 minutes for the presentations and discussions.

4.2.1 Interactive Data Science Platform (NOC)

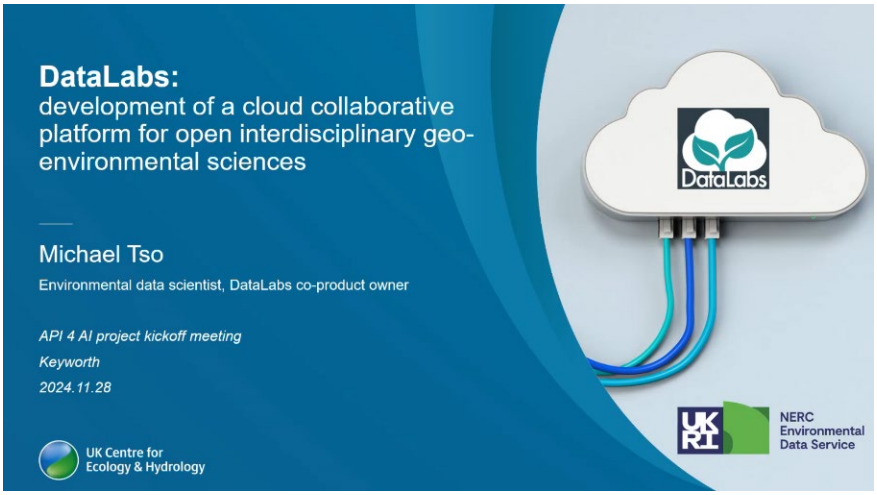
Presented by Colin Sauze, National Oceanographic Centre (NOC)



The presentation provided an overview of the Interactive Data Science Server platform, which offers remote-hosted Jupyter notebooks accessed via a Web Browser. The presentation discussed deployment using Docker, some example use cases, including Machine Learning for image recognition, and the platform's future. The platform fits into the wider Digital strategy at NOC, and as part of this, they provide training to upskill users. They also run the Carpentries workshops to provide researchers with a foundation for coding and data science skills. The final part of the presentation included more demos, including connecting to an external Met Office API.

4.2.2 Datalabs (UKCEH)

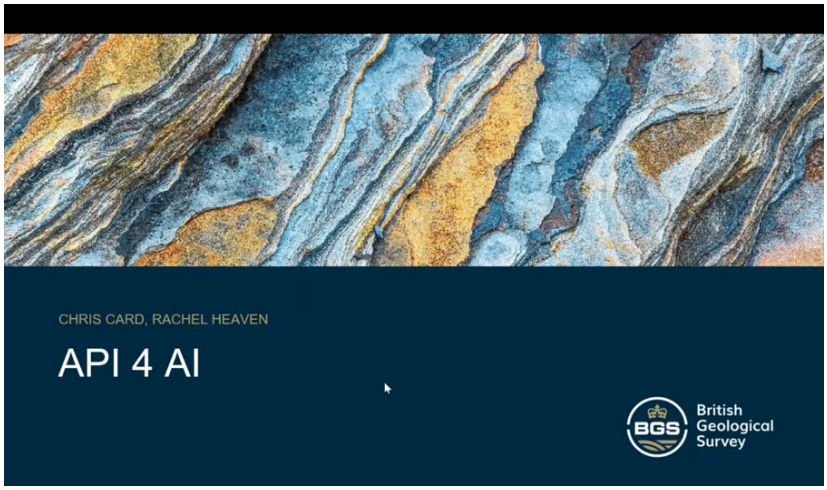
Presented by Michael Tso, UK Centre for Ecology and Hydrology (UKCEH)



The presentation provided an overview of the Datalabs platform, including motivation for design, an introduction, and a list of available tools. They discussed how the platform is open and can be deployed via docker and Kubernetes, which aids platform independence. Some use cases were demonstrated, including using APIs to create an emissions dashboard, a demonstration of using APIs to stitch drone images together and an example of accessing data through APIs to provide training on Datalabs (link to training). It was discussed how Datalabs fits into planned digital research infrastructure (DRI) initiatives and the platform's future plans.

4.2.3 API Standards (BGS)

Presented by Rachel Heaven and Chris Card, British Geological Survey (BGS)



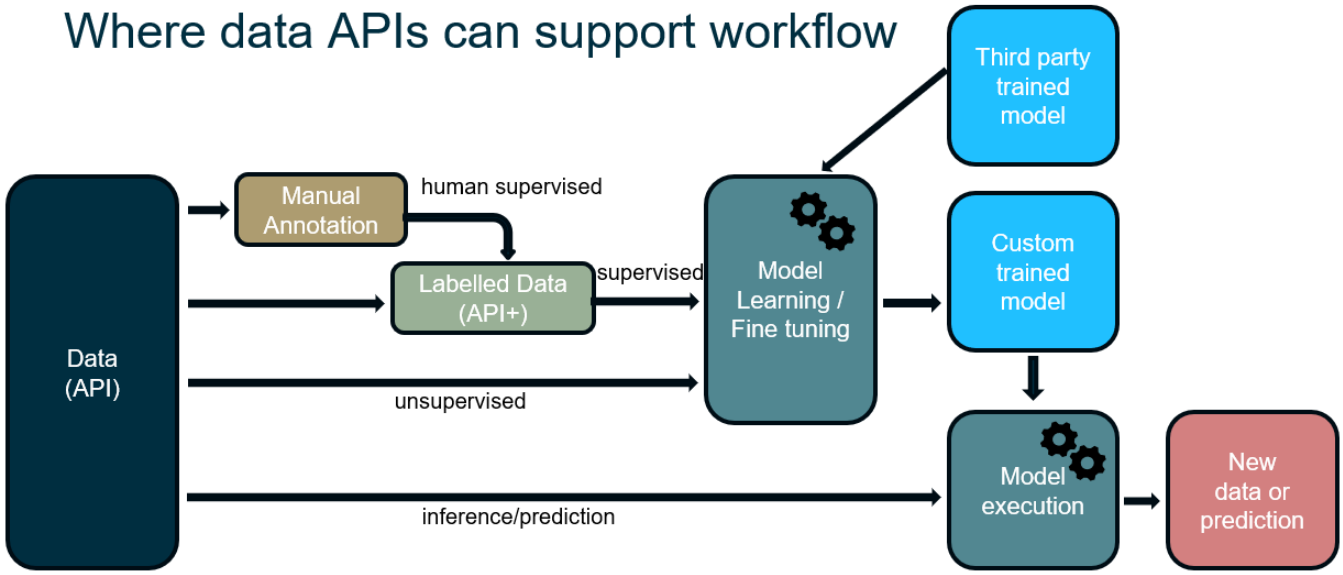
The presentation discussed the development and standardisation of APIs to facilitate AI workflows.
The Data APIs used in EDS were divided into archetypes to help categorise different types of data and their roles in AI workflows.
The following API archetypes were identified: -

- A - Feature/Resource (e.g. image) with metadata
- B - Spatiotemporal Asset catalog (STAC) – a standardised way to expose collections of spatial temporal data
- C -Time-series data - e.g. sensor data streams

This was followed by a description of AI workflows that use the data to generate models or predictions.
The following AI workflows were identified:

- Use unsupervised learning to create a model
- Use supervised learning using attributes to create a model
- Use supervised learning using markup to create a model
- Use a model on the unlabelled data in inference/prediction mode to create new attributed/annotated resources or predictions

This is represented in the following diagram:



BGS walked through some current AI use cases using the different archetypes of data and workflows. One example used Archetype A (Feature or Resource with metadata) and supervised learning to create a model trained to assess rock core fragmentation.

BGS discussed the relevant standards for implementing Archetype A using OGC Features APIs (https://ogcapi.ogc.org/features/) and the OGC Records API (https://ogcapi.ogc.org/records/) presenting a current live dataset. This highlighted some aspects of the standards that require further agreement.

During discussions the group was made aware of a new standard for describing machine learning datasets to make them discoverable and usable, croissant (https://mlcommons.org/working-groups/data/croissant/).

The next step will involve reaching a shared understanding of these standards and publishing API guidelines on GitHub or a similar platform.

4.3 Workshop activity 1 – Identify use cases

This workshop aimed to identify use cases for AI across EDS and consisted of a matrix containing the different data/API archetypes and the AI workflows based on the BGS presentation.

Figure 10 Archetypes/Workflow matrix

API Archetypes:

- A - Feature/Resource with metadata/attributes
- B - Spatiotemporal series of resources
- C - Time series data, e.g. Sensor data
- D = Gridded / continuous coverage data

Workflows:

- Unsupervised training - unlabelled data, e.g. clustering statistical analysis
- Supervised training using Metadata/attributes as labels, e.g. classifications
- Supervised training using marked-up/annotated resource, e.g. computer vision
- Human adding labels/annotation to unlabelled data, e.g. image segmentation
- Apply trained Model to unlabelled data, e.g. run a classifier

Each organisation was given 20 minutes to brainstorm applicable use cases, with each described using the following fields:

- Science question
- Dataset(s) required
- Relevant AI workflow
- Done or planned
- Organisation

Each one was captured on a sticky note.

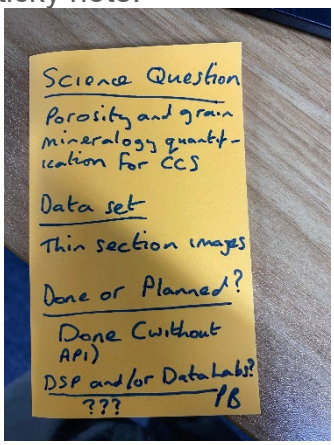


Figure 11 Use case example

At the end of the ideation session, each use case was discussed with the wider group.

Figure 12 Discussing ticket

Each use case was placed on the AI/API Matrix to build a collection of use cases that aligned with the different API archetypes and AI workflows.

Figure 13 Adding a use case to the matrix

The workshop produced 25 use cases from various organisations, covering different areas of the matrix.

Figure 14 Use case matrix

The full list can be found in Appendix B – Use cases

4.4 Workshop activity 2 – Identify tasks and timeline

The use cases generated in the previous workshop activity were used to discuss how to achieve the project's objective of providing a proof of concept that demonstrated the provision of standard-based data APIs for integration into an AI workflow.

Figure 15 Timeline workshop

Participants from different organisations worked together to create the timeline and identify the tasks required to develop the proof of concept.

.

Figure 16 Creating the timeline

The timeline V1

This V1 version of the timeline was created in the meeting.

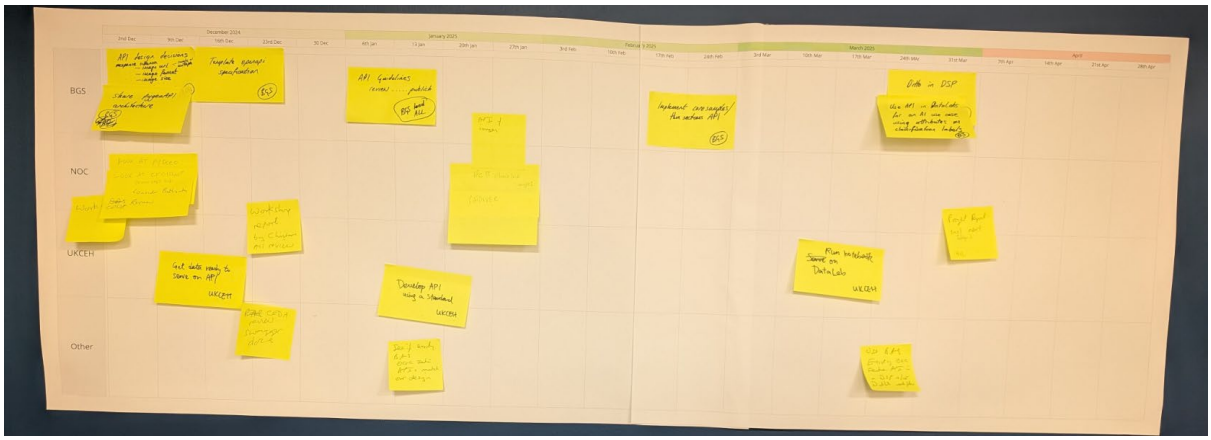


Figure 17 Timeline v1

The timeline V2

BGS reviewed the V1 after the meeting to provide a more realistic timeline for achieving the project deliverables.

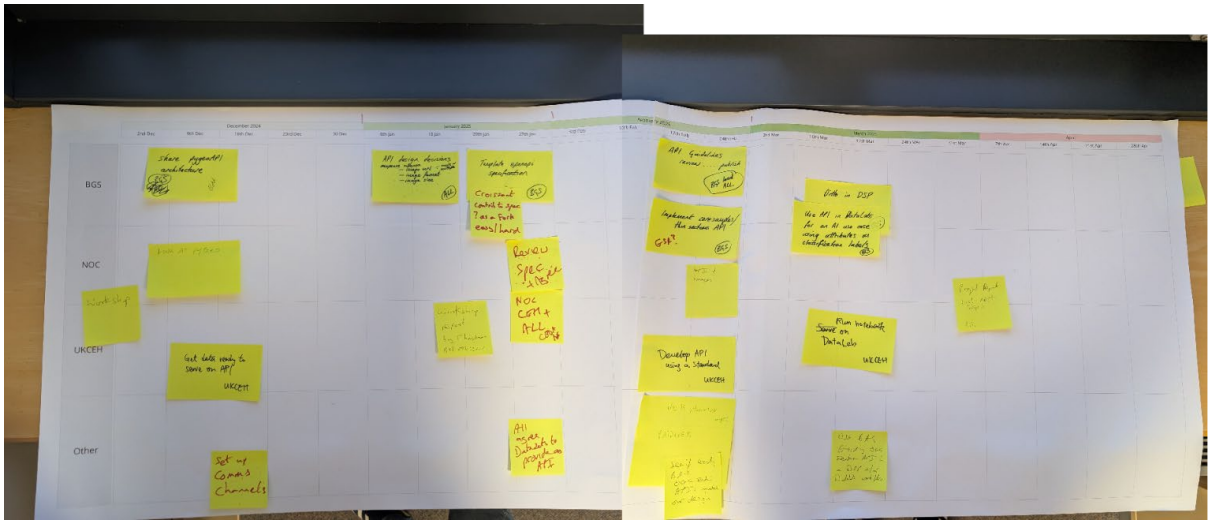


Figure 18 Timeline v2

A detailed view of the current timeline can be viewed in Appendix C – Timeline.

5 Workshop outcomes and next steps

During the workshop, we gained a clearer understanding of the current capabilities of the AI platforms available for use within this project. We also reached an initial agreement on the standards to be used to underpin the provision of data APIs for use in the AI workflows that will be developed.

The workshop also provided two outputs:

- A list of 25 AI use cases from across EDS
- A timeline to implement the identified project deliverables.

The timeline represents a high-level understanding of the next steps of the project, which can be viewed in Appendix C – Timeline. Any subsequent changes will be captured in SharePoint.

An overview of the next steps are as follows:

December 2024 – March 2025
December 2024 – Preparation for API specification delivery. Information sharing relating to agreed API standards and enabling technologies to deliver conformant services.
January 2025 – Delivery of API Specification. Agree on API design decisions. Share, review and confirm API specifications.
February 2025 – Agree on the selection of science use cases that will be addressed in the project proof of concept demonstrator. Publish API guidelines and standards. Create data APIs conformant with the agreed API specification.
March 2025 – Create AI workflows utilising the data APIs to answer the agreed science use cases. Workflows will be implemented across both the NOC DSP and UKCEH DataLabs AI platforms. Write the end of the project report.

April and beyond
The API workshop unveiled 25 AI use cases from across EDS, each presenting a valuable opportunity for implementation with future funding.

6 Acknowledgements

We would like to thank all participants for their valuable insight. Thanks to the NERC Environmental Data Service for its interest in this topic and the UKRI Digital Research Infrastructure programme for funding this project.

7 Appendix

7.1 Appendix A – Agenda

Location: BGS Boardroom, Keyworth

Morning (11 am – 1 pm)
11:00 am - Introductions

- Introductions from all participants across the 5 EDS organisations.

11:30 am – Presentations

- Interactive Data Science Platform - Presented by Colin Sauze (NOC)
- Datalabs - Presented by Michael Tso (UKCEH)
- API Standards - Presented by Rachel Heaven and Chris Card (BGS)

Lunch (1 – 1.30 pm)
1:00 pm – Lunch (1/2 hour)
Afternoon (1:30 pm – 4 pm)
1:30 pm - Use cases workshop

- Workshop activity 1 – Identify use cases
- Workshop activity 2 – Identify tasks and timeline

2:30 pm - Coffee
2:40 pm - Next steps

- Discuss next steps

4:00 pm - End

7.2 Appendix B – Use cases

The use cases were digitised and captured in Miro from the physical boards.

The following sections contain the use cases as captured in the meeting. Each use case has been numbered, and the text has been copied to this document.

This document represents a snapshot of what was captured during the meeting. Any additional use cases or changes will be documented in SharePoint.

7.2.1 API Archetypes A

Use cases that require APIs that provide details of a resource (such as an image) relating to a feature of interest along with its associated metadata attributes.

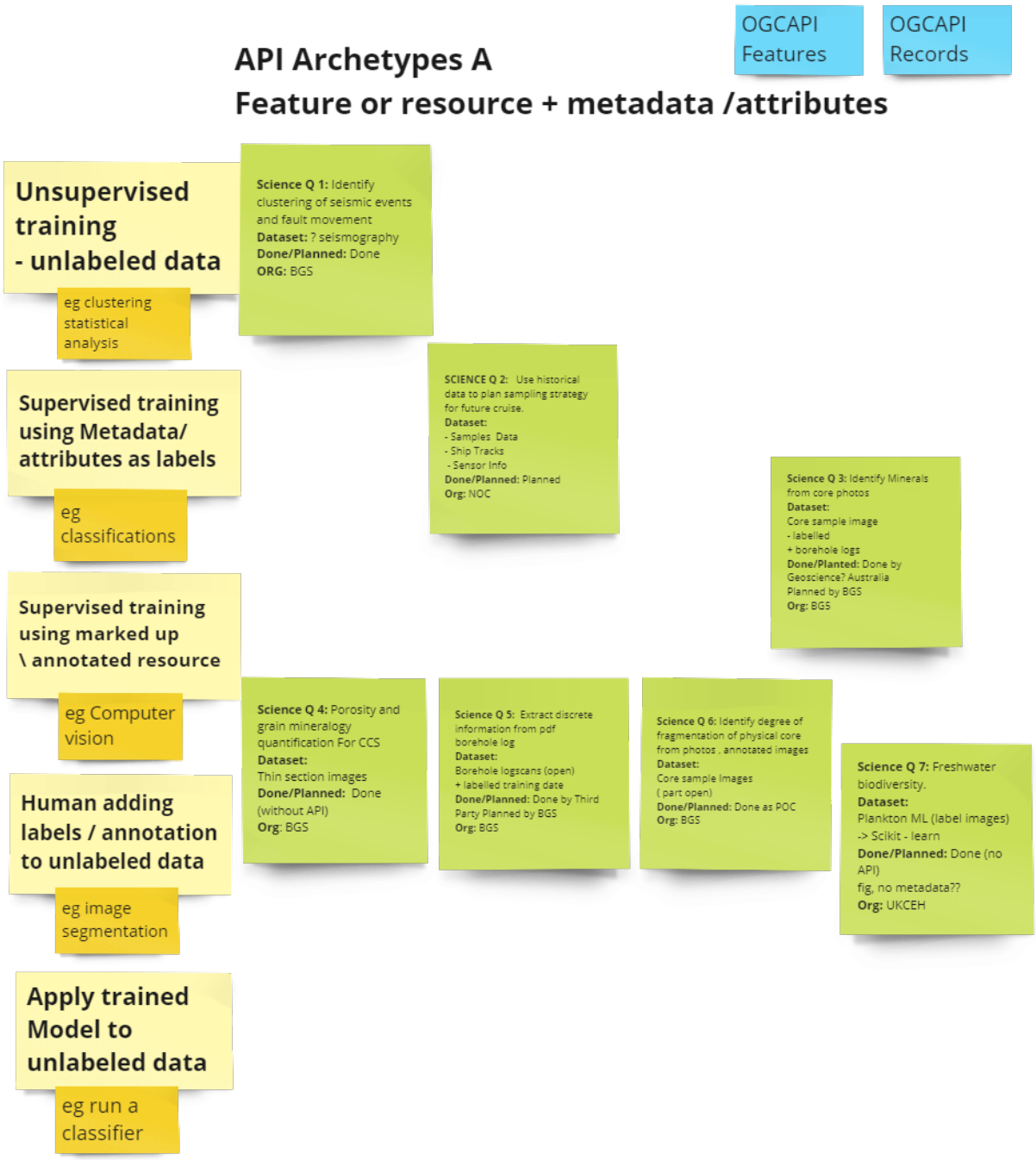


Figure 19 API Archetypes A

List of use cases:

Science Q 1: Identify Clustering of Seismic Events and Fault Movement
Dataset: Seismography (details unspecified)

Status: Completed
Organisation: British Geological Survey (BGS)

Science Q 2: Use Historical Data to Plan Sampling Strategy for Future Cruise
Dataset:
Samples Data
Ship Tracks
Sensor Information
Status: Planned
Organisation: National Oceanography Centre (NOC)

Science Q 3: Identify Minerals from Core Photos
Dataset:
Core Sample Images (labelled)
Borehole Logs
Status: Completed by Geoscience Australia, Planned by BGS
Organisation: BGS

Science Q 4: Porosity and Grain Mineralogy Quantification for CCS
Dataset: Thin Section Images
Status: Completed (without API)
Organisation: BGS

Science Q 5: Extract Discrete Information from PDF Borehole Log
Dataset:
Borehole Log Scans (open)
Labelled Training Data
Status: Completed by Third Party, Planned by BGS
Organisation: BGS

Science Q 6: Identify Degree of Fragmentation of Physical Core from Photos
Dataset: Core Sample Images (partially open)
Status: Completed as Proof of Concept
Organisation: BGS

Science Q 7: Freshwater Biodiversity
Dataset: Plankton ML (label images) using Scikit-learn
Status: Completed (no API), Figure available, No metadata
Organisation: UK Centre for Ecology & Hydrology (UKCEH)

7.2.2 API Archetypes B

Use cases that require APIs that provide details of a spatio-temporal series of resources, e.g. images

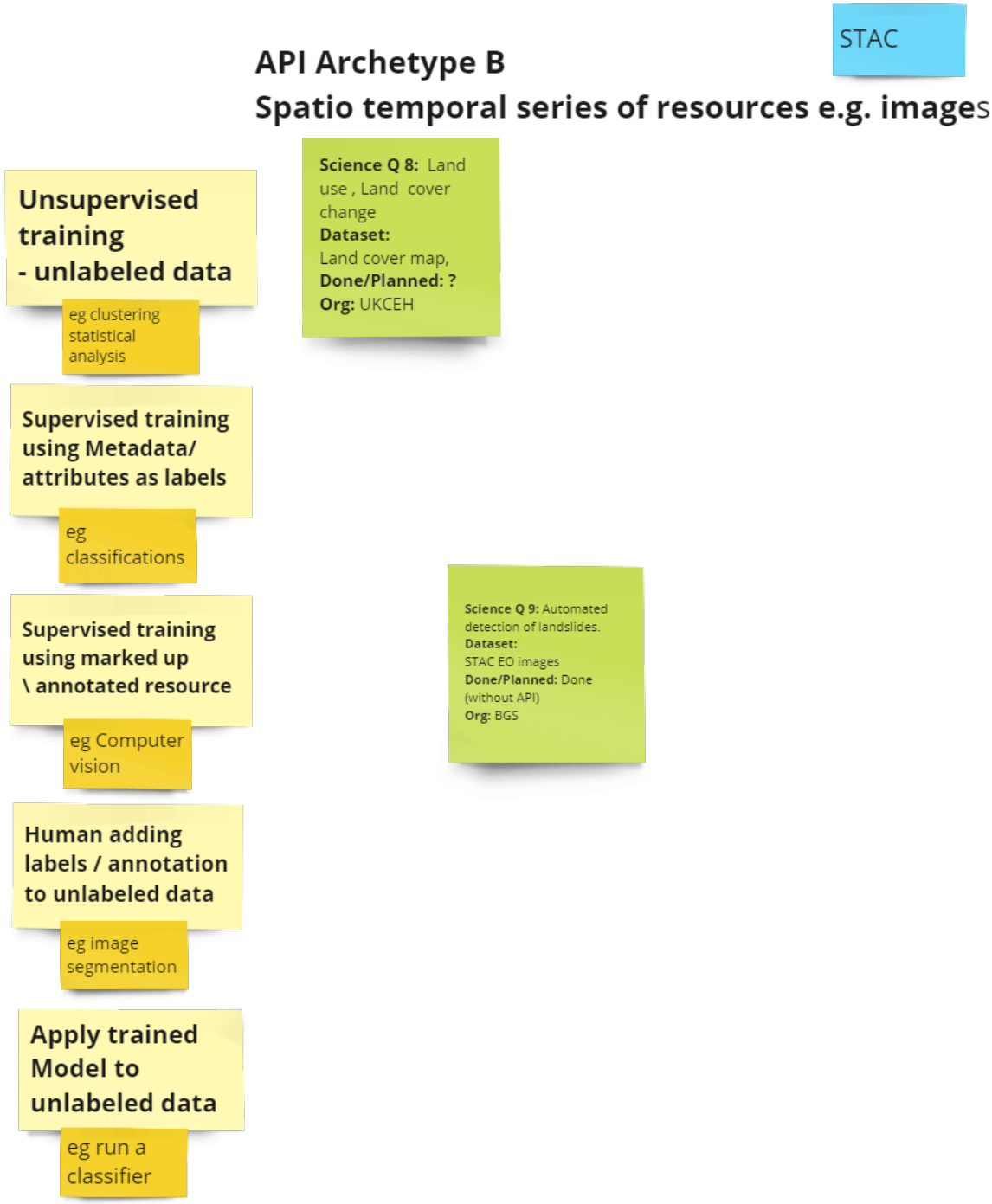


Figure 20 API Archetypes B

List of use cases:

Science Q 8: Investigating Land Use and Land Cover Change

Dataset: Land cover map
Status: The status of completion is not specified.
Organisation: UK Centre for Ecology & Hydrology (UKCEH)

Science Q 9: Developing Automated Methods for Detecting Landslides
Dataset: SpatioTemporal Asset Catalog (STAC) Earth Observation (EO) images
Status: Completed, but without an API
Organisation: British Geological Survey (BGS)

7.2.3 API Archetypes C

Use cases that require APIs that provide details of time series data, e.g. sensor data

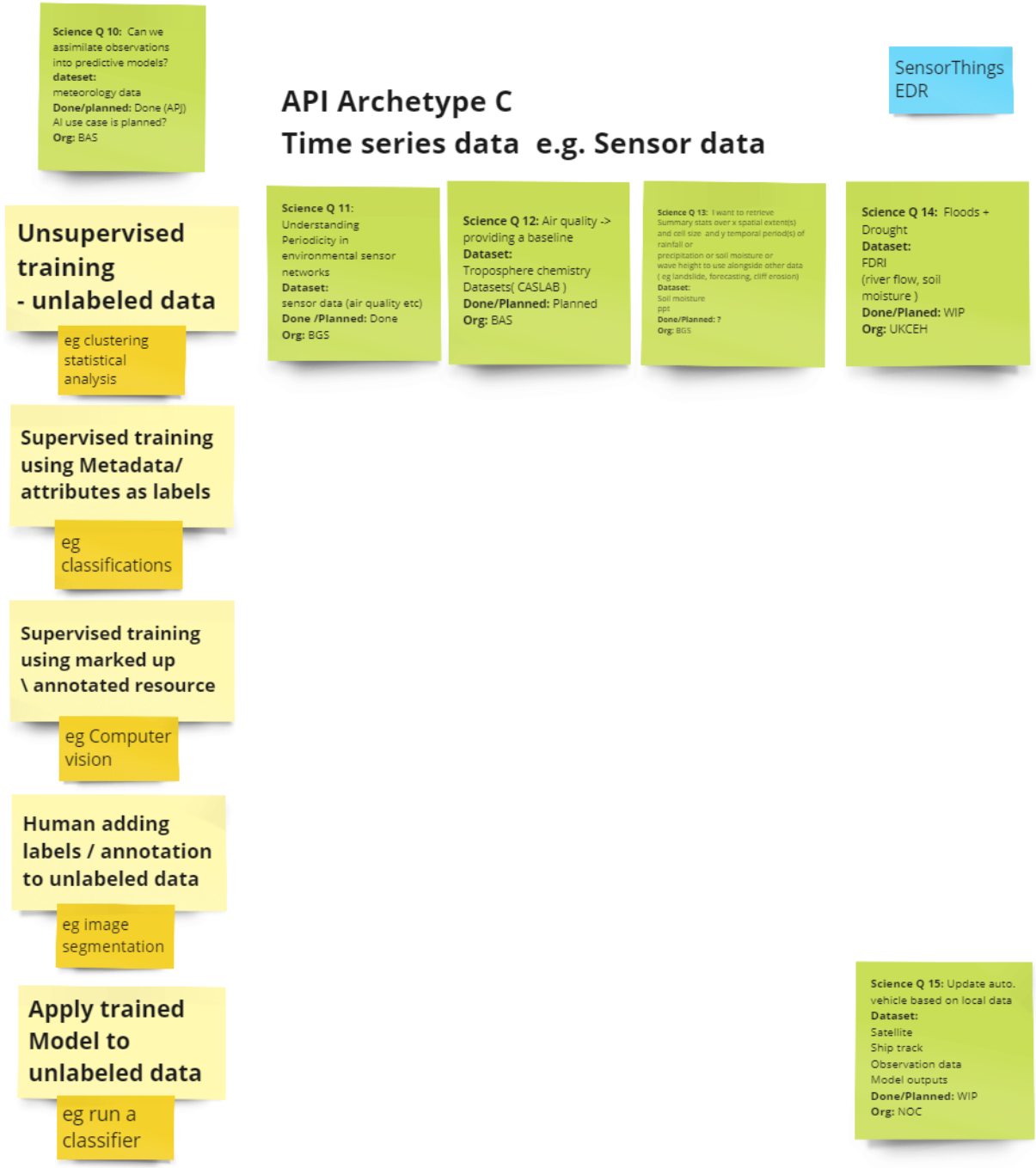


Figure 21 API Archetypes C

Figure 22 API Archetypes C

List of use cases:

Science Q 10: Assimilating Observations into Predictive Models

Dataset: Meteorology data
Status: Done
Organisation: BAS
AI Use Case: Planned

Science Q 11: Understanding Periodicity in Environmental Sensor Networks
Dataset: Sensor data (air quality, etc.)
Status: Done
Organisation: BGS

Science Q 12: Air Quality Baseline
Dataset: Troposphere chemistry Datasets (CASLAB)
Status: Planned
Organisation: BAS

Science Q 13: Summary Statistics for Environmental Data
Dataset: Soil moisture ppt
Status: Unknown
Organisation: BGS

Science Q 14: Floods and Drought
Dataset: FDRI (river flow, soil moisture)
Status: Work in Progress
Organisation: UKCEH

Science Q 15: Updating Autonomous Vehicles with Local Data
Dataset: Satellite, Ship track, Observation data, Model outputs
Status: Work in Progress
Organisation: NOC

7.2.4 API Archetypes D

Use cases that require APIs that provide details of gridded / continuous coverage data

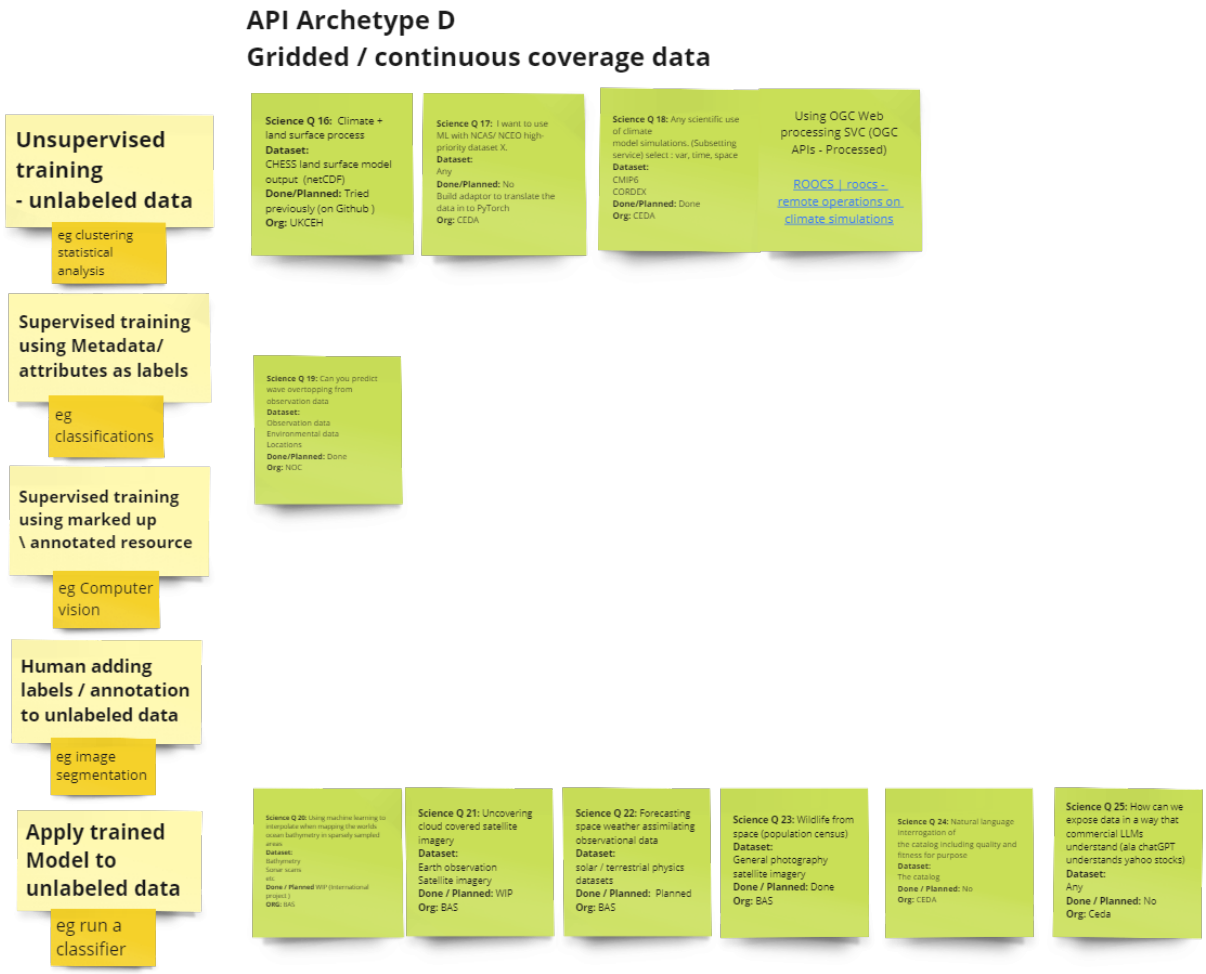


Figure 23 API Archetypes D

List of use cases:
Science Q 16: Climate + land surface process
Dataset: CHESS land surface model output (netCDF)
Done / Planned: Tried previously (on Github)
Organisation: UKCEH

Science Q 17: I want to use ML with NCAS/NCEO high-priority dataset X
Dataset: Any
Done / Planned: No
Additional Note: Build adaptor to translate the data into PyTorch
Organisation: CEDA

Science Q 18: Any scientific use of climate model simulations (Subsetting service)
Dataset: CMIP6, CORDEX
Done / Planned: Done

Organisation: CEDA

Science Q 19: Can you predict wave overtopping from observation data?
Dataset: Observation data, Environmental data, Locations
Done / Planned: Done
Organisation: NOC

Science Q 20: Using machine learning to interpolate when mapping the world's ocean bathymetry in sparsely sampled areas
Dataset: Bathymetry Sonar scans, etc.
Done / Planned: WIP (International project)
Organisation: BAS

Science Q 21: Uncovering cloud covered satellite imagery
Dataset: Earth observation Satellite imagery
Done / Planned: WIP
Organisation: BAS

Science Q 22: Forecasting space weather assimilating observational data
Dataset: Solar / terrestrial physics datasets
Done / Planned: Planned
Organisation: BAS
Science Q 23: Wildlife from space (population census)
Dataset: General photography satellite imagery
Done / Planned: Done
Organisation: BAS

Science Q 24: Natural language interrogation of the catalog including quality and fitness for purpose
Dataset: The catalog
Done / Planned: No
Organisation: CEDA

Science Q 25: How can we expose data in a way that commercial LLMs understand?
Dataset: Any
Done / Planned: No
Organisation: CEDA

7.3 Appendix C – Timeline

The timeline was digitised and captured in Miro from the physical version captured in the meeting. Following the workshop, BGS reviewed and clarified the project deliverables and timeline generated during the workshop.
This document represents a snapshot of what was captured during the meeting and the subsequent BGS refinement. Any additional changes will be documented in SharePoint.

7.3.1 Timeline V1

The timeline, as captured in the meeting

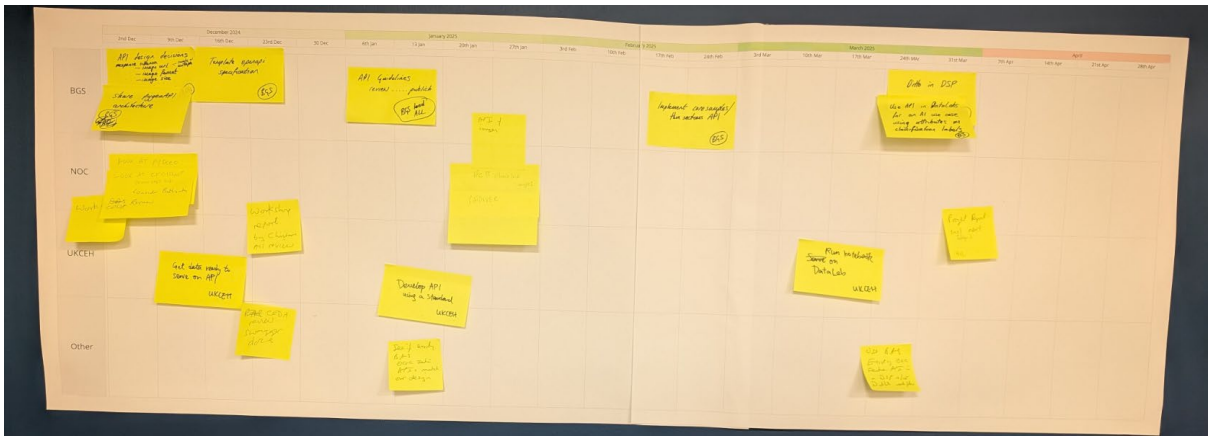


Figure 24 Timeline V1

7.3.2 Timeline V2

Updated timeline to reflect clarified project deliverables and schedule.

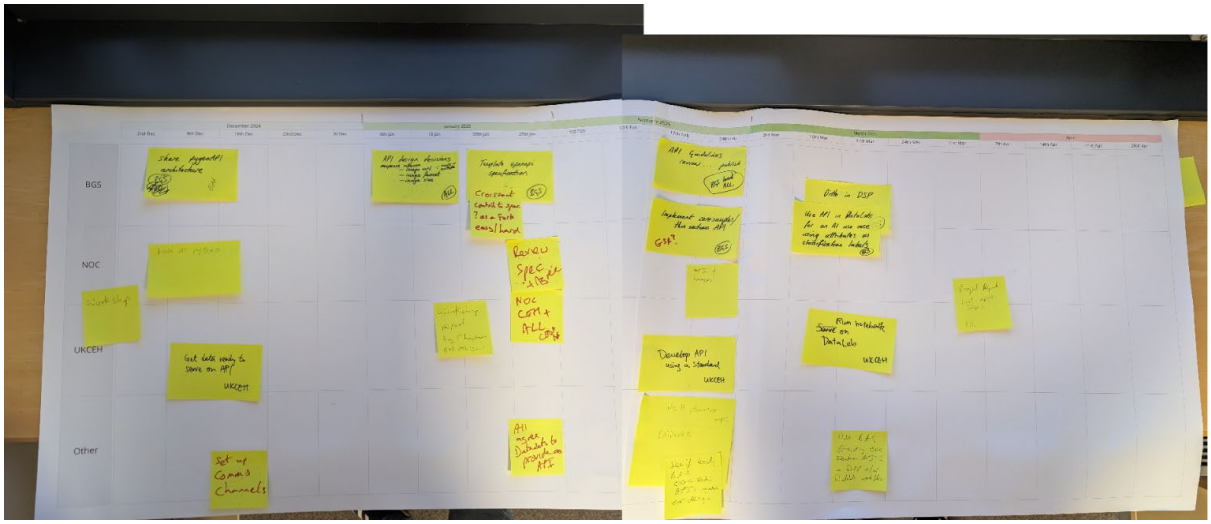


Figure 25 Timeline V2

The digital version of the timeline

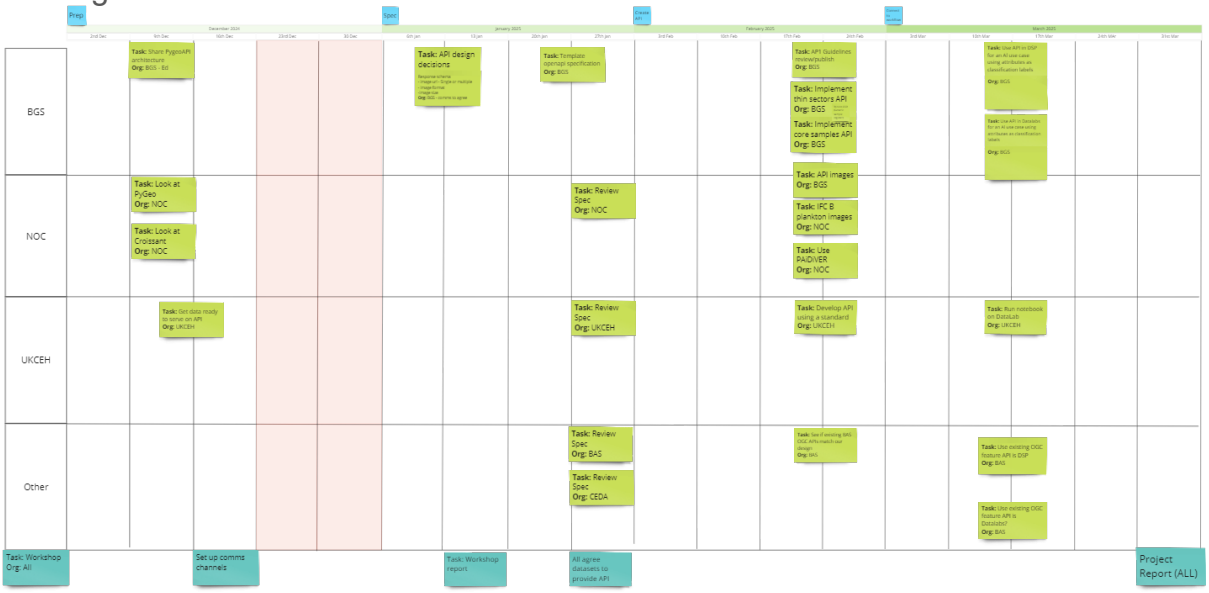


Figure 26 Timeline V2 digital version

A monthly timeline view can be seen over the next few pages.

Dec 2024

Figure 27 Dec 2024

Jan 2025

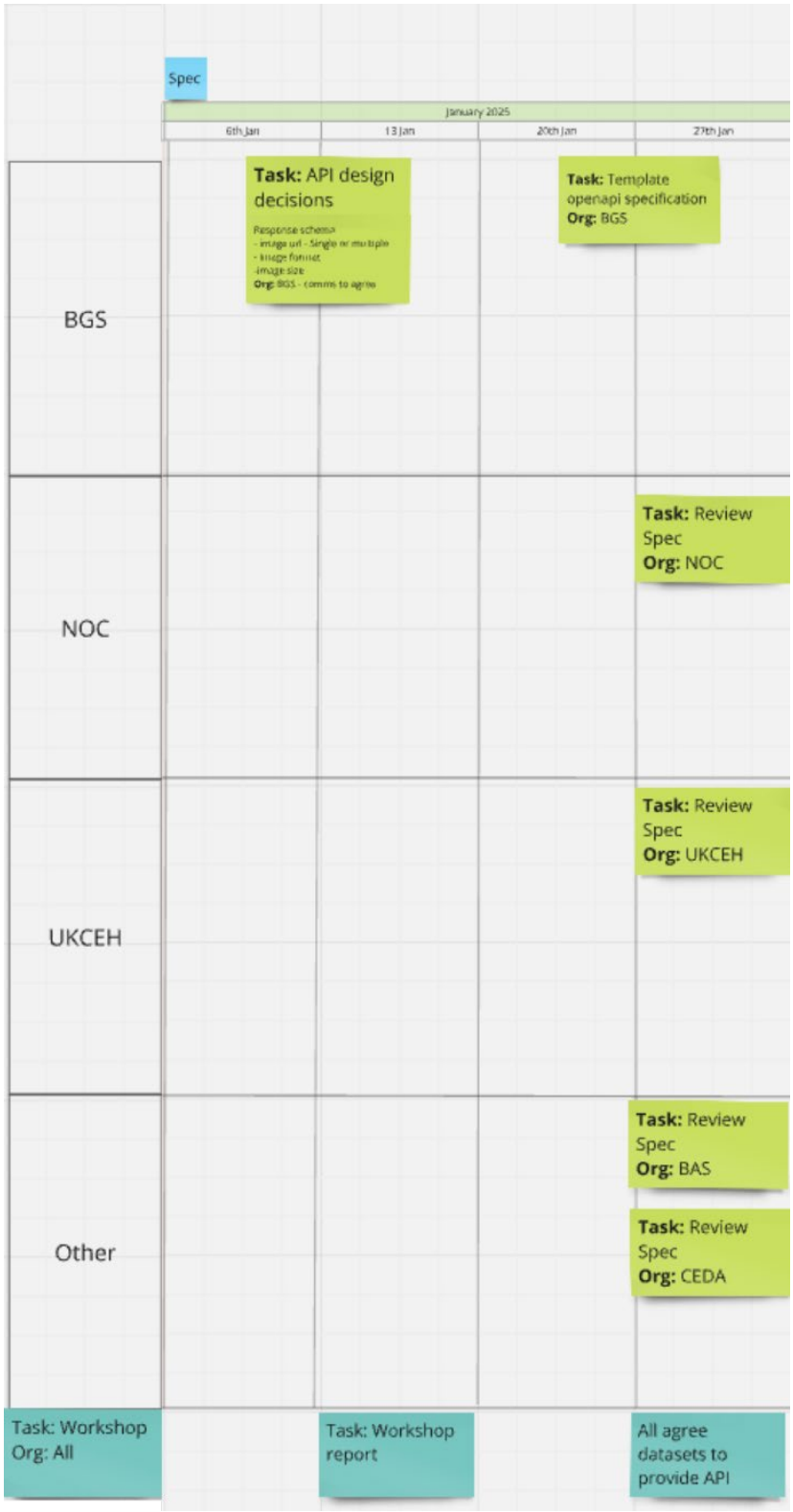


Figure 28 Jan 2025

Feb 2025

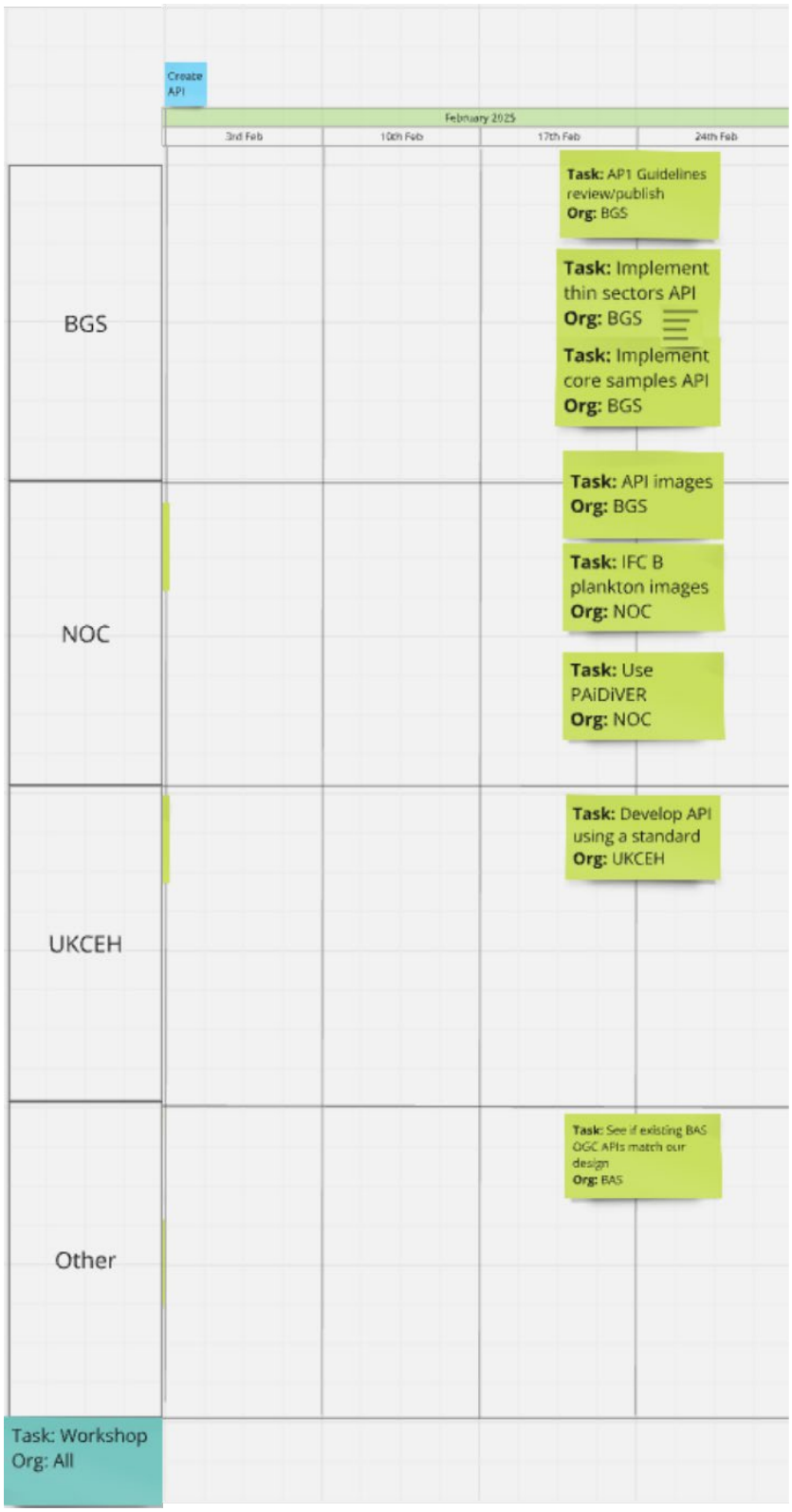


Figure 29 Feb 2025

March 2025

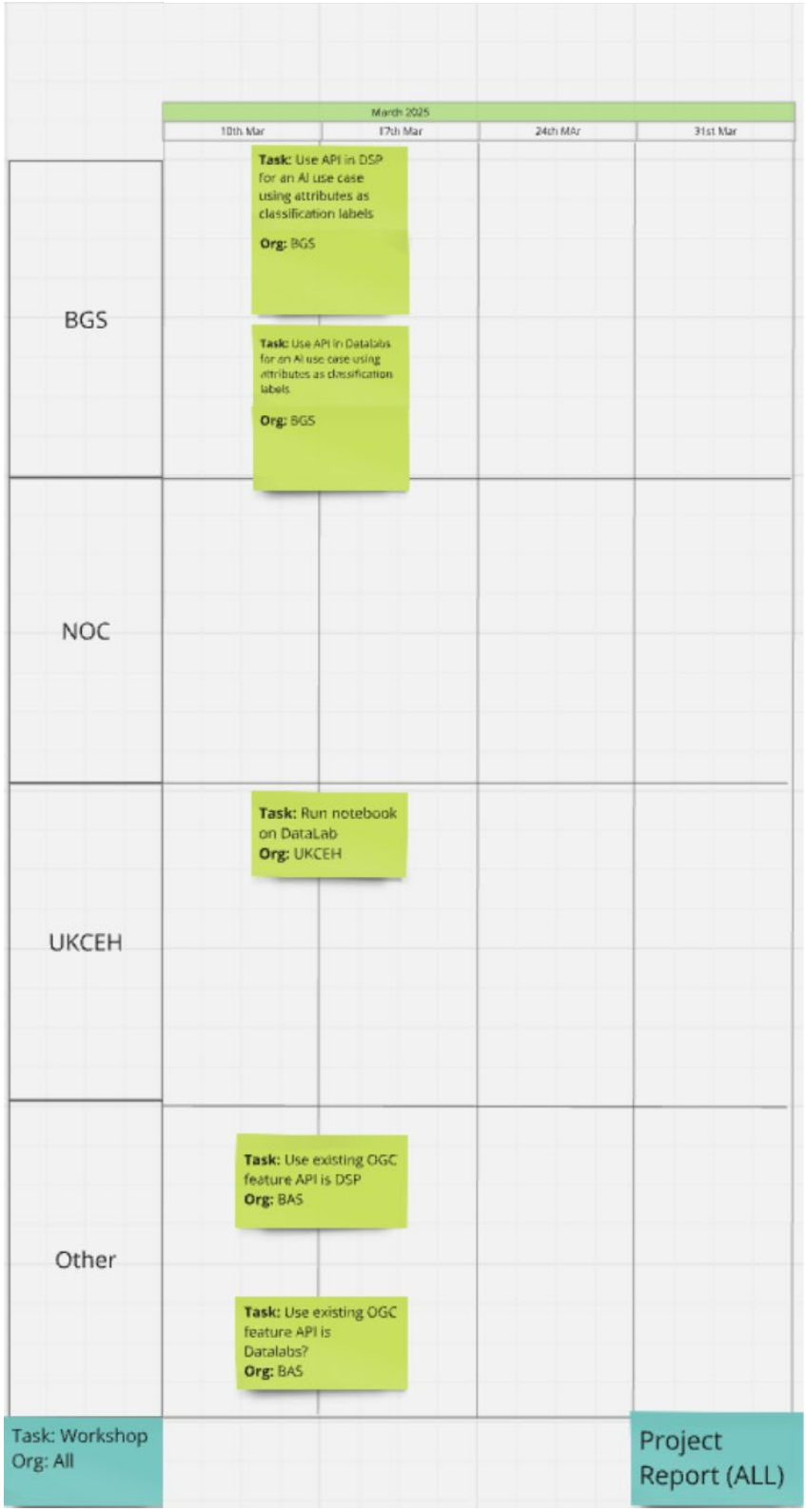


Figure 30 March 2025