High 5 for AI
User Research Report
Combined Analysis:
Quantitative and qualitative findings
March 2026

Report number OR/26/050_subreport2

Authors:

Ola Podgorska https://orcid.org/0009-0002-3722-7658

Shwetha Raveendran https://orcid.org/0000-0002-4757-498X

Carl Watson https://orcid.org/0000-0002-9026-7432

Affiliations:

3 - British Geological Survey
4 - British Antarctic Survey
5 - UK Centre for Ecology and Hydrology
6 - National Oceanography Centre
7 - Digital Solutions Hub, University of Manchester

Executive Summary

The High-5 for AI project, led by the UK's Environmental Data Service (EDS), aims to make NERC environmental datasets (from BODC, CEDA, BGS, CEH, and BAS) more accessible to non-traditional users - particularly in clean energy (eg: offshore wind siting) and public health (eg: blue-green infrastructure planning) - by developing APIs and AI-powered tools.

The High-5 for AI user research reveals that while APIs and AI-ready metadata are essential, they are insufficient on their own to unlock the value of NERC environmental datasets for non-traditional users.

Our research (focused on the clean energy and public health sectors) identified data fragmentation, inconsistent metadata, and the lack of a "single source of truth" as the primary bottlenecks, causing users to spend days on data wrangling rather than analysis.

Crucially, trust in AI outputs remains low without transparent audit trails and human-in-the-loop validation, while data discovery continues to rely heavily on fragile informal networks rather than robust digital tools.

To transform EDS from a siloed asset into a shared, actionable resource, future projects should shift focus from merely building interfaces to fixing foundational data infrastructure. Success depends on adopting cloud-native formats like Parquet for efficiency, but more importantly, on implementing semantic harmonisation layers that allow diverse datasets to interoperate without forcing rigid standardisation on providers.

By prioritising "preview before download" workflows, federated indexing, and tools that digitise and improve collaboration for communities of knowledge, the EDS can significantly reduce time-to-insight and support the UK's environmental missions (eg: clean energy and public health) with trustworthy, AI-ready data.

Contents

Executive Summary ..... Page 2

Contents ..... Page 3

Strategic Recommendations ..... Page 4

Research Objectives and Background ..... Page 5

Primary Research Questions ..... Page 6

Methodology Overview ..... Page 7

Research Impact within the Project ..... Page 9

User Flow Diagram ..... Page 12

Key Findings Summary ..... Page 13

Findings and Recommendations in Detail ..... Page 14-26

User Archetypes ..... Page 27-31

Ecosystem Service Map and Pain-Points..... Page 33

User Stories ..... Page 36

Methodology in Detail ..... Page 37

Strategic Recommendations

About the Environmental Data Service

The Environmental Data Service (EDS) provides a focal point for scientific data and information spanning all environmental science domains: atmosphere and climate, earth observation, polar and cryosphere, marine, terrestrial and freshwater, geoscience, solar and space physics. The EDS is made up of a network of distributed data centres, with domain specific expertise.

Our main goal is to ensure that environmental data are made available, accessible and re-usable for the long-term in order to fully realise their value. We are funded by the Natural Environment Research Council (NERC), part of UK Research and Innovation, to advise researchers on how to prepare data for long term storage and dissemination.

The EDS is a fundamental part of NERC's digital strategy and works with the Digital Solutions Hub, and other partners, to break down disciplinary barriers and facilitate data sharing beyond academic use. **While the EDS aspires to transform environmental data from a siloed asset into a shared, actionable resource, we believe this report is useful as the research revealed some gaps in the initial approach, as well as new areas of opportunity.**

We go into more detail about these as they relate to the High 5 for AI project in the "Findings and Recommendations" section on Page 13 of this report. The three insights below are strategic-level recommendations, which have been based on our research findings:

1. Digitise and coordinate the "human glue": The research highlights that data discovery is currently a social act dependent on individual relationships, creating a single point of failure when experts leave. Beyond just "expert directories," the EDS, NERC or others could form **collaborative communities, cohorts, or workspaces** that allow users to co-curate datasets, share context-rich "playbooks" on how specific data was combined, and maintain an institutional memory log. A "digitisation" of the informal trust networks would ensure that critical knowledge about *where* data exists and *how* to use it would be democratised and would remove or reduce knowledge "silos".

2. A curated federated index supported by semantic harmonisation: Adopt a dual approach to make data easily findable and accessible across sectors. First, establish a curated federated index where data custodians are incentivised to maintain "fit-for-purpose" badges (eg: "AI-Ready") on their datasets, ensuring quality and trust. Utilising a semantic harmonisation layer will allow these diverse datasets to be used in conjunction with each other, without forcing rigid standards (which even when in place, are also not always adopted by users in reality). By combining active curation with flexible semantic mapping, this model shifts the focus from simple discovery to active collaboration. This structure specifically lowers barriers for non-specialist users and enables seamless cross-discipline data use, effectively accelerating overall data maturity. Also see UNITED project *WEBSITE CASE STUDY*,
EPSRC project (CEH), and FMRI Digital Infrastructure Scoping study findings.

3. Transparency is a core product feature, not an add-on: Users distrust AI without audit trails - suggesting that explainability must be baked into the product architecture from day one to any AI tool used for Research. Generating a human-readable "data lineage receipt" for every query or AI output would be the recommended route. This receipt would explicitly list the source versions, processing steps, and any assumptions made by the AI, positioning EDS as the trusted broker that

bridges the gap between raw data complexity and the need for rapid, defensible decision-making in government and industry.

Limitation: As the project was specifically focused on the two primary use-cases, there is a lack of evidence across more diverse use-cases and scientific research areas, Further user research is recommended for a more inclusive "Environmental Data Service" view.

Research objectives and background (Why did we do this?)

PROBLEM STATEMENT

Environmental datasets from the UK's five NERC data centres (BODC, CEDA, NGDC, EIDC, PDC) are under-utilised outside the environmental science community.

Despite being 'high-value' for users across sectors (eg: government, policy making, urban planning, insurance, offshore wind, public health) face significant user challenges in discovering, accessing, and integrating these datasets due to:

- inconsistent metadata
- incompatible formats
- lack of AI-ready structures
- the overwhelming variety and volume of user-interfaces

This results in excessive time spent on data wrangling (often days to months), low trust in automated outputs, and missed opportunities for cross-sector innovation – preventing sectors beyond environmental science from using public assets to solve for broader government missions like clean energy and public health.

WHY THIS RESEARCH WAS CONDUCTED

The High-5 for AI project was initiated to address the growing demand for accessible, interoperable, and AI-ready environmental data particularly for use cases beyond traditional environmental science. With increasing reliance on AI/ML for decision-making, stakeholders need tools that reduce manual effort, ensure reproducibility, and build confidence in data-driven outcomes. This research was conducted to understand real-world user behaviours, pain points, and expectations: to inform the design of APIs and AI-powered interfaces that truly serve diverse users, not just environmental scientists.

ORGANISATIONAL AND PROJECT GOALS

- **Enable broader adoption** of EDS datasets by non-environmental sectors (eg: construction, insurance, urban planning, health).
- **Develop APIs and AI tools** (eg: agentic chatbots and workflows) that simplify access, integration, and analysis of complex environmental data.
- **Support UK government missions**:
  - Accelerate offshore wind deployment (clean energy)
  - Improve public health via blue-green infrastructure planning
- **Position EDS as a trusted, future-ready data infrastructure**
  Aligned with FAIR principles, cloud-native formats, and user-centric design.

PRIMARY RESEARCH QUESTIONS

1. How do users currently discover, access, and use EDS datasets: what are their biggest needs and pain points?
2. What formats, tools, and interfaces do users prefer for integrating environmental data into their workflows?
3. What does "AI-ready data" mean to different users/ user archetypes: what barriers prevent adoption?
4. How can APIs and AI tools reduce friction in data access while maintaining trust and transparency?
5. What role do human networks and institutional relationships play in data discovery: how can tools support this?

INITIAL HYPOTHESIS
In the EDS DRI funding proposal for the High 5 for AI project this was the initial scope proposed for the project:

"Reaching beyond business as usual, a substantial part of the project (approx. 50%) is devoted ***to providing improved access to these datasets through APIs*** *with a particular focus on ensuring the data is ready for use in AI applications* ***by utilising common metadata formats such as Croissant*** *that describe datasets for use in AI."*

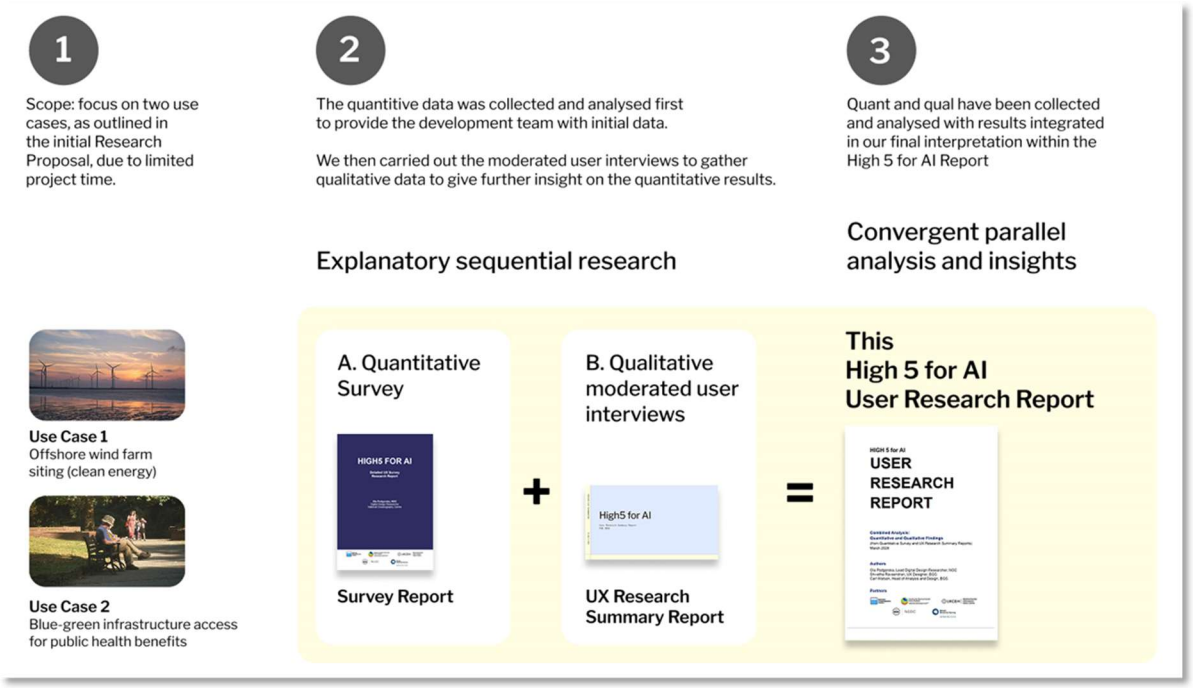
Our user research reveals that while APIs and AI-ready metadata are essential, they alone cannot solve the data access challenge – for example, scientists also struggle with data discovery and access, inconsistent and interoperable formats that render the datasets unfit for purpose. Consequently, this report expands beyond the original hypothesis to address these critical gaps in location, standardization, and data quality.

Challenge Category	Key Issues
1. Location & Discovery	Poor search, inconsistent naming, multiple interfaces, no single "source of truth" / portal
2. Fit for Purpose	Insufficient metadata, outdated data, overly specific datasets
3. Access & Governance	Licensing, permissions, robotic processes, lack of human support
4. Formats & Interoperability	Multiple grids, projections, lack of STAC/ARCO, no R/Python packages
5. Readiness for Use	Not "analysis-ready" (eg: lot of pre-processing needed)

METHODOLOGY OVERVIEW

This user research was conducted between November 2025 and January 2026 to inform the design and development of APIs and AI-powered tools for the High-5 for AI project: an initiative by NERC's Environmental Data Service (EDS) to improve accessibility of environmental datasets beyond the traditional scientific community.

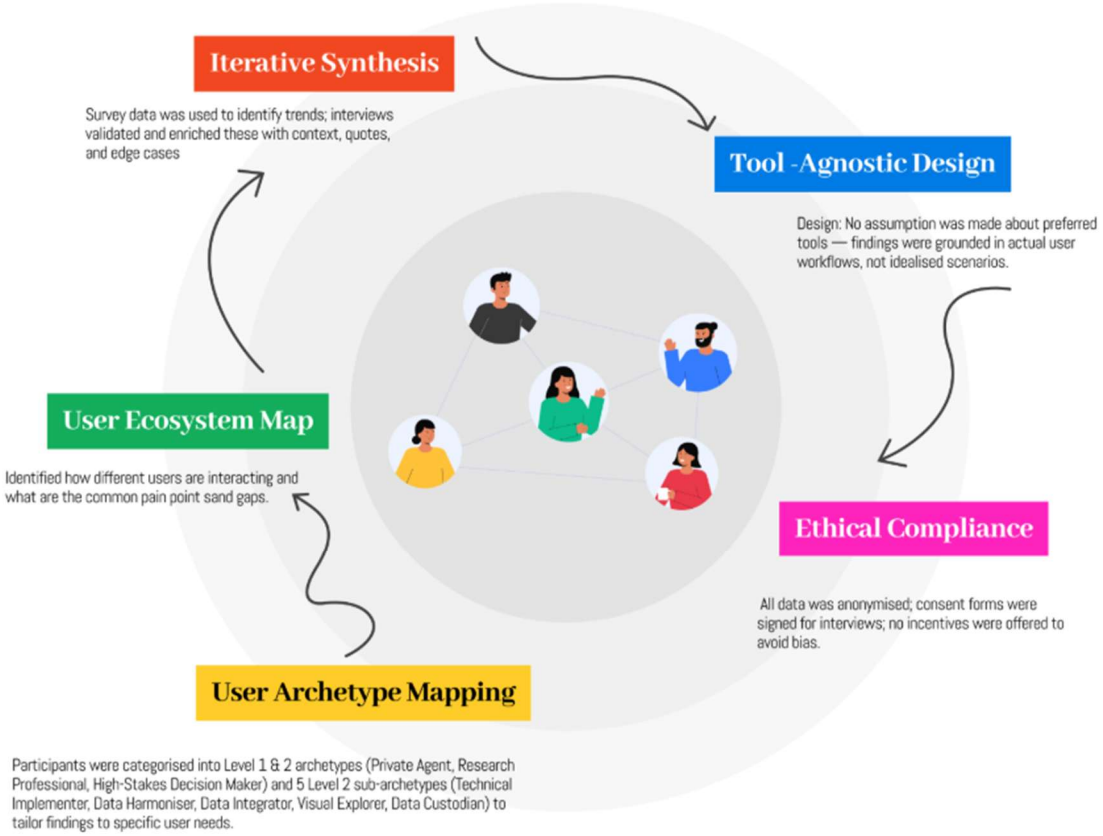
The research employed a **mixed-methods approach**, combining **quantitative survey data** with **qualitative, moderated 1-2-1 interviews**, to capture both breadth and depth of user needs, behaviours, and pain points.



A. UX Survey Report
B. UX Research Summary Report
C. Consolidated UX Research Report

This methodology ensures a balanced, user-centric, and evidence-based foundation for the High-5 for AI project.

It captures both the scale of user challenges (via survey) and the nuance of their experiences (via interviews), to guide the development of truly useful, trustworthy, and adoptable tools, supporting real-world needs.



For further information on how we carried out the research, please see the "Methodology in detail" section - Page 41 in this document.

Research impact within the project: in response to user challenges (What do we know now?)

1. Users spend most time on foundational tasks like cleaning data than analysing it: due to poor metadata and format inconsistency.

Validated through our research. Among others, we found out that Parquet is a desired format in the survey and then confirmed this through the user-research interviews.
Some of the Development teams have consequently focused on delivering Parquet through Croissant, while others continued (as originally planned) in delivering APIs.

GeoParquet is a cloud native format, which adds native support for geospatial vector data. With a columnar storage architecture, and highly efficient compression algorithms, it reduces cloud storage costs and Input/Output operations - enabling faster analysis, and scalable eco-system interoperability. Both of the High 5 for AI use cases involve analysing massive datasets.

Use Case 1
For offshore windfarm siting – a user may want to include certain variables (eg: wind speed, water depth and seabed composition) across millions of gridded cells – but exclude turbine maintenance logs (for example). Parquet allows the system to read **only the specific columns required in the dataset**, drastically reducing operations and speeding up queries, compared to row-based formats like Shapefiles or GeoJSON.

Use Case 2
Calculating public health benefits (eg: air quality improvement, heat island reduction) often requires aggregating data across millions of population points or land parcels. Parquet's columnar nature makes these aggregations significantly faster.

Limitations: Because Parquet files are fixed, they require full rewrites to update individual records, they are not suited for real-time telemetry or live sensor feeds, which instead demand time-series databases or append-optimized formats (eg: complex structured files like R NetCDF, semi structured flat files like .JSON, or FlatGeobuf, optimised for fast reading/writing of vector features where you might need to continuously add new data points efficiently).

2. Based on the findings from our user research interviews, we recognised that the "B. Scientists/Researcher "Technical implementers" are our focus primary API users: but aside from the API functionality they *also* require technical API stability, documentation, and cloud-native formats (see point 1).

How have we implemented this? The recommendations of this report have been passed onto the Development team on the project who have created example notebooks* as part of the project, which will be situated on a public repository, for user access. Parquet was chosen as it's a cloud native format (see Point 1).

** Notebooks facilitate interactive exploration and reproducible analysis, allowing users to visualise data, test hypotheses in real-time, and share their findings without requiring extensive programming expertise.*

3. During our research, user evidenced their distrust of AI outputs unless provenance, audit trails, and multi-model validation were provided.

The existing High 5 for AI systems do not generate AI outputs, so the specific requirements for AI audit trails and multi-model validation are not applicable to the current production environment.

Recognising the critical importance of trust in autonomous systems, exploratory work has been conducted under the HIGH5 initiative by Lucy Killoran and Ag Stephens, and while the prototype addresses the user-identified need for validation and trust assessment, these capabilities remain in the research phase. They are not ready for public release or integration into the user-facing community tools.

Please refer to the WP4 Agentic AI Evaluation Report for more information.

4. Legacy and sensitive data remain inaccessible: requiring new pathways beyond traditional APIs.

How have we implemented this? This is currently out of scope for the High-5 for AI project, but to support future projects to address these challenges, we've written some User Stories which can address these challenges. One example:

As a Research Professional, I want to discover and access hard-to-find environmental data through a trusted network of colleagues and domain experts ***so that I can*** leverage collective knowledge to locate, verify, and contextualise datasets that are not available in formal catalogues or repositories.

For further context, please see:
- Page 19 in this report [Finding 3]
- User stories [WP2_and_5_High5forAI_UXResearchSummary.pdf - Page 5]
- Future aspirations covered in High 5 for AI case study WEBSITE LINK

5. Discoverability and usability are the top pain points: users want us to properly curate the data offered as they're looking for something specific (the user research provided us with examples of one type of data being represented in a range of datasets causing confusion).

“It's in the magnitude of hundreds of disparate data sets due to the fact that... a lot of it's held by NGOs, some of its being held by natural England, there's different regulators around the country. There's a significant different number of repositories that they have to gather this in and ring together”
- User Voice

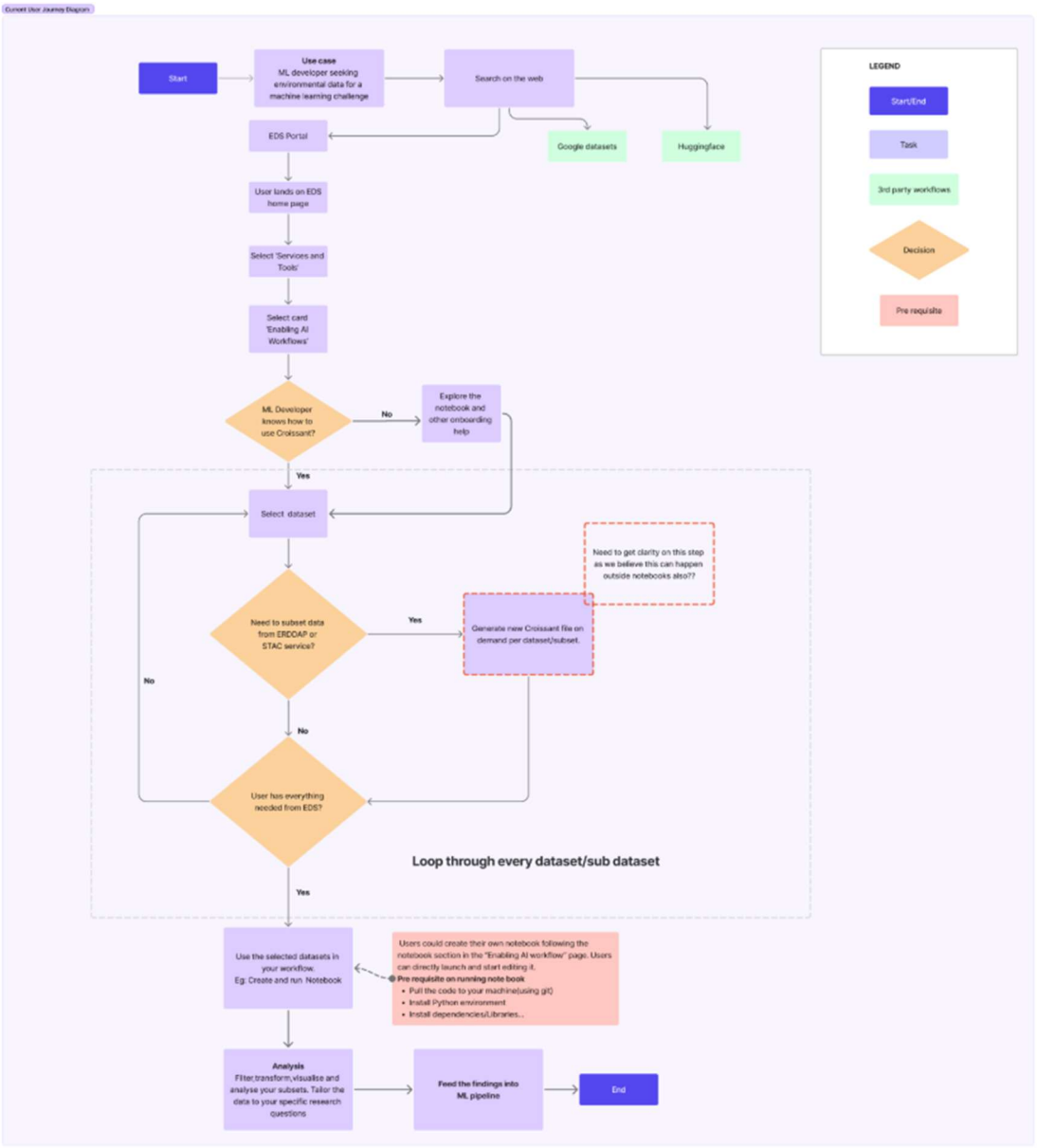
A “single source of truth” (eg: aggregated portal) with standardised formats would significantly reduce time-to-use and increase trust.

How have we implemented this?
The implemented workflow guides users through accessing and preparing environmental data for AI projects via the NERC EDS platform. Eg: the current user journey begins when an ML developer lands on the EDS Portal, navigates to "Services and Tools," and selects "Enabling AI Workflows," the developers can also search environmental data through external sources like Google Datasets or Hugging Face, loops through available datasets and sub-datasets, and manually determines if data subsets are needed from ERDDAP or STAC services before generating new Croissant files on demand.

The process concludes only after the user verifies, they possess the necessary Croissant knowledge, preferably launches a notebook to analyse, filter, and visualise the tailored data, and finally feeds these findings into their ML pipeline.

User Flow Diagram

The below diagram shows the journey through the EDS portal, The datasets can be found in via Google Dataset or through Hugging Face and that journey may differ.



This journey is a working draft at time of publication. It has not yet been fully validated with stakeholders and users. It should be treated as indicative rather than definitive. There could be gaps in the information.

Key Findings Summary

1. APIs are context-dependent, not a complete solution — Page 14-15
2. Data fragmentation is universal. Harmonisation and interoperability are the key to AI workflows — Page 16-17
3. The “human glue” dichotomy — Page 17-19
4. Trust requires a “Human-in-the-Loop" — Page 19-20
5. “AI-Readiness” is determined by metadata depth and context of use — Page 20-22
6. Discoverability & access: the “Single Source of Truth” is a curated index, not a central repository — Page 22-23
7. Missing workflows: “Preview before download” — Page 24
8. Legacy data recovery is a manual, high-effort bottleneck — Page 25-26

Findings and Recommendations in detail

FINDING 1
APIS ARE CONTEXT-DEPENDENT, NOT A COMPLETE SOLUTION

While APIs exist, they are niche tools for system architecture. They do not solve the foundational problems for the Private Sector (access vs. security), Policy Makers (speed), or Consultants (usability) - the delivery mechanism must match the data scale and security context.

In the survey (*Q10 – Do you prefer to download data or access via an API?*) we found out that the majority of users said they prefer to download data directly (44%), but in combination with those who said they prefer the option of both an API (application programming interface) and a download (9.2%) - this suggests API's are very popular. However, the survey is **current state –** the results are not an accurate representation of **potential future needs.**

To understand the 'why', we questioned API use cases **in moderated interviews.** Users reported that while APIs are useful for light metadata queries or web mapping, they are often viewed as insufficient for large-scale scientific modelling or heavy industrial use cases (like seismic surveys). Issues such as firewalls, security risks, and the sheer volume of "big data" (terabytes) mean that many organisations (esp. in private sector) still prefer bulk downloads or direct "in-place" cloud access where they can run proprietary models securely. During our interviews we also noted that **a differentiation between Web API's and System (Developer) APIs was not always clear to end-users.**

The difference is that some API's will be accessible to users, and others won't be 'discoverable' as they are intentionally built purely to allow internal systems to function (the technical architecture). Ultimately, API usage was related to the user's role, and the task they were trying to achieve.

- **User Voice:** *"We're talking terabytes worth of data... you need to pull the whole data set down... APIs wouldn't work."*
- **User Voice:** *"For running hydraulic models driven by multiple river gauges, APIs were too fragile. When one gauge went offline (returning "no data"), the whole automated chain broke, requiring "hands-on person fiddling..."*
- **User Voice:** *"Some locations they run for one day… next day for a couple of hours… disruptions in the observation should not be there."*

What users reported APIs are useful for	What users think API's currently struggle with
Accessing specific datasets, providing direct access to lower resolution data efficiently (where a direct download of a dataset is not a requirement for research)	Highly detailed, high-resolution data (unless the user knows EXACTLY which part of the dataset is relevant to them upfront, which would require a visualisation mechanism before data download/API use)
	Accessing large volume datasets (terabytes)
Automated data collection (eg: weather station data)	Those who need to manipulate the data locally (on their computers / local servers) eg: private sector end-use
Providing interoperability between different datasets	Notifying end-users who have installed the API of upgrades or disruptions in observations (which can

	cause data transfer interruptions which severely impact research projects)
	Clean user documentation (often not provided, causing users to waste time during install)
	Harmonising inconsistent data, and data formats, unit types etc. - a requirement for interoperability to be possible

For detailed information, refer to the
User Research Summary Finding 6 and 10
(WP2_and_5_High5forAI_UXResearchSummary.pdf)

RECOMMENDATIONS

- Build API's which help support the harmonisation of datasets (to enable AI readiness and interoperability across different datasets)
- Build reliable, well-documented system APIs, with AI-ready cloud formats supported (eg: Parquet) as an alternative to "cloud-native" direct access.

While APIs are valuable for quick verification and lightweight integration, **direct downloads** remain crucial for heavy processing, modelling, and secure offline environments (eg: firewalled networks), which can also be achieved by APIs. API documentation was called out as a specific need across at least three sources.

Some potential solutions:

1. **Adopt cloud-native formats:** Transition towards cloud-optimised formats like Parquet, Zarr, and Cloud-Optimized GeoTIFF (COG). Benefit: These formats support efficient sub-setting and streaming, reducing the need to download entire datasets for small queries. Current State: Users currently spend hours/days converting legacy formats (NetCDF, GRIB, HDF5) into analysis-ready structures. **User Research Summary Finding 3**
   (WP2_and_5_High5forAI_UXResearchSummary.pdf)

2. **Pilot Cloud-Native Delivery:** Demonstrate a workflow where users can query and analyse a large environmental dataset (eg: ERA5 or bathymetry) via Zarr/Parquet without downloading the full file.

3. **Publish best practice guidelines:** Release a "Playbook" for data providers on how to structure metadata and licences to maximise interoperability and AI usability.

FINDING 2
HARMONISATION & INTEROPERABILITY ARE THE FOUNDATION FOR AI WORKFLOWS, BUT DATA FRAGMENTATION IS UNIVERSAL

Every user type, from Students to Data Custodians, is connected to the central "Data Fragmentation" hub.

No one is immune to the problem of siloed, incompatible data. Lack of standardisation is the most significant barrier to automation and adoption of AI, and interoperable data use. Users spend disproportionate time "wrangling" data rather than analysing it.

In the survey *(Q3: How many distinct datasets would you typically use for any specific work task?)* we found 89%* of users surveyed use multiple datasets per task (datasets are used in combination to carry out complex, multi-source workflows), and in Q5. (How much time do you spend on data once sourced?") common themes emerged:

- "Too much time"
- "Significant time to create ML/cloud-friendly data"
- "Quality checking, format adjustment, documentation review"

*59% of respondents said they use 2-5, an additional 11% said they use 5 or more
Please see WP2_High5forAI_Survey_Detailed Report.pdf

This was confirmed in the user research: 8 out of 11 participants we spoke with, highlighted the importance of metadata for data utility, specifically regarding how it provides necessary context, enables inter-comparison and harmonisation, and is a prerequisite for AI and machine learning readiness. Inconsistent data structures, the manual labor currently required to link metadata, grids, naming, and units were the most cited barriers to effective interoperable data use.

It's data wrangling that causes the bottleneck (not the analysis). However, a few mentioned that focuses on the wrong level, noting that "data architecture and infrastructure" matter more than standardising formats, which "will change all the time."

User Voice:

- *"AI was going to solve the problem... [but] it doesn't pull back the data... with a sufficient level of confidence...*
- *"I'm talking about standardisation, not standards... to make sure that they are standardising across their data model design and development..."*

User Research Summary Finding 1 & 2
(WP2_and_5_High5forAI_UXResearchSummary.pdf)

RECOMMENDATIONS

The creation of pre-processing tools (eg: APIs, auto-formatting, metadata enrichment, or "Data Health Check" tools (AI-assisted validators that ingests sample datasets and report on compliance with FAIR principles and AI-readiness: missing metadata, format issues, interoperability) could potentially help users solve for this.

A. **Develop an API which introduces an intermediary semantic / ontology layer to bridge the gap:** Create a mapping service that translates domain-specific jargon (eg: "benthic survey") into standardised tags usable by search engines and LLMs, to support standardisation of ontologies.

B. **Develop automated cleaning tools**: Create APIs or tools specifically designed for data cleaning, categorisation, and labelling. *Use Case:* An AI tool that scans a dataset and flags missing mandatory fields (eg: "missing wave state") or suggests standard ontological terms.

   AI Context: Large Language Models (LLMs) and AI agents require semantic clarity to interpret data correctly. Inconsistent naming (eg: "first_name" vs. "fname") breaks automated pipelines.

C. **Harmonise licensing:** Move towards standard, clear licenses (eg: Open Government Licence, CC BY 4.0). This is directly in response to pain points bought up during the moderated interviews: ambiguous or complex licensing agreements prevent data combination and delay projects, especially when aggregating data from multiple NGOs or commercial entities.

FINDING 3
THE "HUMAN GLUE" DICHOTOMY

Data discovery is not driven by search engines: it's a social act, driven by personal networks, institutional relationships, and domain experts. This is a community strength that's rarely recognised or given prominence in the development of digital tools. But it also creates a critical risk: knowledge of where data exists (and how to access it) rests on individuals.

A successful tool must digitise and support this social layer of discovery, not just the technical catalogue.

Sometimes the system only works because skilled individuals are manually fixing broken data pipelines. In addition, gatekeeping by private utilities and/or commercial entities and stonewalling of requests for specific datasets, even when it is essential for public good creates significant blind spots. Changes in the international policy can lead to sudden loss of vital dataset. If the "human connection" networks fail, silos occur and significant issues may arise. **Relying on these networks to patch gaps alone is inadequate**, unreliable and unsustainable: people move on, critical knowledge is lost, systems and processes shift and change over time.

In the survey (*Q8: How do you choose which datasets are relevant to you?)* Participants responded "Task relevance – matching the dataset to their specific research question and project requirements) - 20 responses (32.3%). The second highest response was "ease of findability and availablity of data" - 9 responses (14.5%).

The third most popular option was "*finding it through expertise and existing knowledge (individual or community subject matter expertise) - 8 responses (12.9%).* These open-ended responses revealed licensing and conventions/events were also very useful, and one respondent mentioned a protocol used to search for data:

- *"We develop a protocol to search for data, then review lists of what we found to decide whether to include/exclude."*
- *"I take advantage of the CF **conventions,** the DRS conventions used by CMIP…"*
- *"--> **licensing** --> ease of use"*

In the Moderated user interviews, we found that the reliance on informal networks is not just a preference, but a necessity driven by converging findings. Everyone we spoke to would on some level spend time networking within their community in the search of authoritative data sources, because each project required different, specific datasets.

1. We discovered that practitioners already act as "data detectives," using trusted personal networks and institutional relationships (on projects where team members have overlapping expertise, for example). **This is one of the primary mechanisms through which scientists actually locate, trust, and successfully integrate environmental data.**

2. Currently, users leverage these human connections to build shared understanding, as well as ensuring continuity in the face of "gatekeeping" or geopolitical instability. Commercial sensitivity, geopolitical shifts and cost to transfer/host hinder and prevent access to critical data. **Geopolitical data interruptions are particularly damaging to AI workflows, which require consistent data streams to maintain reliable predictive environmental models.**

User Voice:

- *"We are... very, quite well connected into the academic space... that allows us, I suppose, access to leading figures... who can help us say, right, actually we know that data exists here..."*
- *"Looking at data spaces where you could have some secure way or sharing data with trusted partners... sounds like it would be the the right way to go..."*
- *"I know that lots of companies are fighting it and trying to retain ownership of that when it really has quite small value to them... there's a blockage there because of the because of the data ownership."*

User Research Summary Finding 9, considered in conjunction with 7 & 8 (WP2_and_5_High5forAI_UXResearchSummary.pdf)

An additional privacy challenge: Due to the commercial confidentiality, granular level search data is often held back (eg: end-user marine activity data, health data, fisheries data).

User Voice:

- *The difficulty with fisheries data: vessels under 10 meters are active close to shore (critical for wind farm siting), but their data cannot be shared at high resolution due to **commercial confidentiality** (identifying specific fishermen), in practice, they all know what each other are doing anyway... but legally, it's their business confidentiality."*
- *"An example of one that I'd love to get a hold of is Strava activity data... we were looking for marine activity data... we got a block there where we went pursuing the data at the highest level we could and they wouldn't release it to us. It was just a stone-wall"*

Recommendation
To address the reliance on informal but organic data discovery processes, which are often at the mercy of institutional gatekeeping, we could build simple, digital tools that capture this expertise and support the Research community and support knowledge transfer – creating open, sustainable ecosystems that collect and catalogue critical knowledge and vital relationships/connections.

- **"Ask the Community" features** - remove user's barriers to access collective wisdom (instead of hunting down individuals manually, needing formal introductions etc.), reducing the socio-economic barriers which exist between sectors, democratising knowledge sharing.
- **Expert directories with clear tags** so you can instantly find who knows what, without needing an insider introduction.
- **Project-based data sharing with provenance** to track exactly where data comes from and who handled it.
- **Institutional memory logs** that record "who worked on what, where" ensuring knowledge survives even when teams change and programmes or projects finish up.
- **Privacy-preserving methods or mechanisms**: Develop robust methods for handling sensitive data (eg: commercial fisheries locations, individual health records). For example: implement secure data spaces or aggregation techniques (eg: fuzzy masking, differential privacy) that allow analysis without exposing individual identities.
- **Clarify authority**: Clearly distinguish between authoritative source data and derived/third-party republishing to prevent the propagation of errors or outdated information

FINDING 4
TRUST REQUIRES A "HUMAN IN THE LOOP"

Trust is not built by confidence in AI tools, but by transparency of process. Users demand visible audit trails, explicit processing steps, and the ability to trace data lineage (provenance) back to the original source.

Currently, the path from "Data Custodians" to "Policy Makers" is blocked by a lack of validation and visualisation tools. Without trusted, curated data, Policy Makers cannot get the "rapid insights" they need. Digital, and specifically AI tools can help in solving this challenge successfully, *if* the 'human in the loop' component is factored in.

In the survey (*Q17: How much do you trust the output of your AI workflows?)* the majority of participants (63.6%) rated their trust in AI workflows as a neutral 3 on a 5-point Linkert scale, with a weighted average score of 3.09. Responses showed some clustering around the middle of the scale, indicating mixed levels of trust.

We followed this up with Q18: What would give you peace of mind that the results you receive are trustworthy? What do you need for assurance?
22 responses out of the 43 participants, 14 (63.6%) emphasized the need for explicit processing steps or an audit trail for trustworthy results. A few also mentioned the provision of data sources and "overall transparency".

The clear preference for transparency through explicit processing steps indicates a requirement for methodological rigor and reproducibility, and there is clearly opportunity to enhance trust by consistently providing detailed documentation.

In the interviews, despite of the eagerness for the AI output, there was a concern regarding the reliability of the outputs, and they consistently demanded for the “human-in-the-loop" element to verify the results.

User Research Summary Finding 5 & 7
(WP2_and_5_High5forAI_UXResearchSummary.pdf)

User Voice:

- *"I firmly believe that 'AI' systems are nothing more than a 'get rich quick' scheme for the companies that develop them. LLMs are non-deterministic, frequently inaccurate, reinforce existing biases in datasets, and have a far higher environmental impact than comparable systems.No responsible organisation should be considering their use..."*

- *"Most of my job is is finding new sources of data, updated data, get new connections that might find these connections."*

Recommendation

Provenance and Audit Trails: Any AI-generated insight or processed dataset must include a visible **audit trail** showing:

- Source data trails or origins
- Processing steps applied, prompts used.
- Versioning of models and algorithms used.

Human-in-the-Loop Validation:
Maintain mechanisms for expert validation. AI should be positioned as a decision *aid* (providing options, highlighting anomalies) rather than a black-box decision *maker.*

FINDING 5
“AI-READINESS” IS DETERMINED BY METADATA DEPTH AND CONTEXT OF USE

AI-Ready" is currently an ambiguous term: users have different understandings (or mental models) of the term “AI Ready”. From the survey and contextual inquiry, we found that users think data is rarely “AI-ready” because metadata is often missing, incomplete, or lacks the semantic context required for machine learning. Users emphasized that metadata is as valuable as the data itself; without it, datasets cannot be compared, combined, or trusted. “AI-ready” effectively means “context-rich and machine-readable.”

We found this out by asking Q12 and Q16 in the survey. When we asked the participants in how they interpret the term "AI-read data" (Q12: What is AI ready data mean to users) it revealed a dominant focus **on data structure and AI/ML compatibility**, with 18 respondents (the largest group) emphasizing the need for **standardized formats, clean values, and sufficient metadata to enable machine learning applications without manual intervention**.

A secondary significant group of 12 respondents expressed scepticism and ambiguity, often viewing the term as marketing hype, meaningless buzzwords, or simply as "sources of information" without clear applicability – which **revealed a need for clarity and context setting whenever utilising AI.**

In Q16 we asked participants to explain their key challenges in AI workflow creation. **Option 2 (Data collection, preparation and management) is most mentioned out of the 19 responses collected.**

AI Lifecycle Stage	Challenges	Responses
1. Problem definition, method selection	Unclear goals, methods, workflows	3
2. Data collection, preparation, management	Quality, scale, formats, labelling, misuse	6
3. Model development and prompting	Skills gaps, prompting, coding	3
4. Evaluation, validation, reproducibility	Inconsistent, unreliable outputs	3
5. Deployment, operations, constraints	Cost and usage limits	1
6. Governance, trust, human factors	Time, trust, accountability	3

From the user interviews – across the board all users mentioned they "don't trust" AI **[Finding 5]** and we consistently heard about data which isn't ready for API operation **due to the lack of metadata information.** Users also shared the lack of confidence in using APIs.

User Voice:

- *"the metadata is sort of king in that space... it's that metadata which is as important as the data itself to give context to the LLM."*

- *"To train on not ideal. It also completely lacks nice metadata, which makes it into comparison of data sets virtually impossible. We have some some of our sort of operational gold standard data sets... so these are better,* ***but they're still not ideal, certainly for the certainly for training on large data sets, very large data sets."***

- *"Yeah, I mean, the short answer I'd say is no, they're not AI-ready. That's because things are have not been standardised. The the metadata for all of these different surveys is not standardised. It's you've got missing information all over the place."*

These is captured in detail in the User Research Summary: Finding 2 & 5 (WP2_and_5_High5forAI_UXResearchSummary.pdf)

Recommendations:
Define "AI-Ready" by Use Case: Recognize that "AI-Ready" is context-dependent. Provide guidance or presets for common AI tasks (eg: "Computer Vision Ready" vs. "Time-Series Forecasting Ready"), specifying required resolutions, formats, and metadata completeness.

Be clear and explicit about context of use and how AI is involved:
When building any AI-tools, it's important to be clear about which tasks it helps to solve, and exactly how automation or AI plays a part on the landing page, or throughout the user-flow of the interface.

Support Data Fusion: Facilitate the combination of disparate datasets (eg: combining bathymetry with wind resource data) by ensuring consistent spatial/temporal grids and coordinate systems.

Address the "Long Tail" of Data: Prioritize the digitization and standardization of legacy data (eg: old seismic tapes, paper records) which are often critical for long-term climate baselines but currently inaccessible.

FINDING 6
DISCOVERABILITY & ACCESS: THE "SINGLE SOURCE OF TRUTH" IS A CURATED INDEX, NOT A CENTRAL REPOSITORY

Users consistently report that finding the right data is their primary bottleneck. The current landscape is fragmented across multiple repositories with inconsistent search interfaces. A "single source of truth" was desired but also acknowledged as something which may be "naive" to consider (as it would have to be a cross-Organisational endeavour).

Many researchers expressed frustration over "scattered" and "fragmentary" data that requires navigating multiple, disparate repositories and is sometimes restricted.
The time it takes to find and access these, increases time-to-use. In addition, current search hierarchies and naming conventions are often non-systematic, making it hard to find the "most up to date" or "fit-for-purpose" datasets. This issue occurs at two levels: when working with datasets from different providers, but also when it comes to internal data use (eg: data duplication on internal servers within one organisation).

The research interviews revealed that the challenge doesn't lie in centralising environmental data or standardising formats, but in addressing it at the technical architecture and infrastructure levels. Focusing here enables the seamless integration of modular models, ensuring scalability and rapid adaptability while accommodating the inherent fragmentation across organizations.

User voice:

- Researchers waste weeks manually hunting for data. In one case, a seismic survey took 5 months to locate because the metadata was missing from public portals. *Critical datasets (eg: small-vessel fisheries, marine mammal observations) are often known only through informal networks ("who you know") rather than searchable catalogues.*
-
- *"For marine activity data (eg: sailing/windsurfing) bounced around between multiple emails. Despite the data being available for urban cycling. Without a central index, researchers rely entirely on "personal connections" and "topic experts" to know data exists".*

User Research Summary Finding 2, 4 and 9
(WP2_and_5_High5forAI_UXResearchSummary.pdf)

Recommendations:

- **Build a Federated Index ("single source of truth").** Develop a centralised federated search layer, or curated platform that indexes authoritative sources without physically hosting the data (point to verified sources with clear provenance, rather than a physical centralisation of files). This provides a unified view of available assets with a unified view and clear licensing, solving fragmentation while respecting existing data sovereignty and storage architectures (without requiring unified storage). Currently even sites like data.gov or the EDS catalogue are not consolidated – along with hundreds of other disparate catalogues, there is no UK/Europe wide dataset portal that is the definitive 'authority' on all environmental data.

- **Integrate filtering and tracking:** Implement prompting functionalities or granular filtering on the platform, to allow users to find data with the correct temporal resolution, variables, and licence type, ensuring that every search result includes an auditable trail back to the original source to maintain trust and verify data quality.

While a curated index solves discoverability, it does not solve the semantic fragmentation caused by human resistance to rigid standards. Therefore, we recommend augmenting an index like this with:

- **AI-mediated harmonisation layer. A**n "intelligent translation engine" that works as an automated conversion layer, mapping data from various provider schemas (eg: varying variable names, units, or structures) into a single, consistent format. It would achieve this by translating each provider's unique data structure in real time, ensuring a common user view without requiring any changes to the providers' current systems or workflows. This would bypass political deadlocks over rigid standards by allowing providers to keep their existing formats while delivering a synchronised, standardised output to the user.

- **Enable "smart" semantic intent search:** Replace reliance on exact keyword matching with AI-driven semantic search that understands user intent rather than specific terminology. This allows users to find data based on concepts (eg: "wind speed at 10m") regardless of the provider's specific naming conventions, neutralizing "emotional commitments" to local standards.

- **Establish a "living ontology" with feedback loops:** Create a dynamic, community-driven metadata framework where AI learns from successful user queries and data combinations to suggest emerging best practices. This shifts standardisation from a top-down mandate to an organic, bottom-up process that evolves with actual usage patterns.

FINDING 7
"PREVIEW BEFORE DOWNLOAD" IS A MISSING WORKFLOW

A major pain point identified is the inefficiency of downloading massive files only to discover they are irrelevant. Users strongly desire to visualise and understand what's inside the dataset before use. Without this, they follow a "download first, check later" workflow, and "understanding" the data before analysis takes significant amounts of time. A preview tool would reduce this understanding phase significantly.

In the Survey we asked two open questions:
• Q5: "How much time do you spend on the data once you've sourced this?
• Q6: "What's your biggest challenge finding relevant datasets?

Individuals provided open text responses noting that datasets were often not specific enough and that ReadMe files were often empty or incomplete, which demonstrates that users struggled to determine if a dataset fits their needs (scale, recency, structure) without downloading it first or reading often-missing documentation.Out of 23 responses in Question 5, all of them stated it takes them "hours", of those - 15 exceeded to "days", and 7 exceeded that, taking from weeks to years.

User voice:
• *"Often quite some time, **firstly to understand the data using the supporting documentation** and then integrating data sources."*
• *"It depends on the complexity of the data and the length of the supporting documentation. Up to a day per dataset..."*
• *"Most datasets are not designed to be generally applicable... they are usually super specific...so, it can be difficult wading through data sets about foraging habits of a species of moth on farmland in York in May 1998 when **I'm trying to do something more generic like map UK biodiversity."***

We probed into detail further during the moderated research, to understand what would be most useful in terms of "pre-viz". Users were specifically interested in pattern recognition and finding out which variables were included in the datasets.

User voice:
- *"We often have to spend ages writing the code to kind of read in different data sets, and often what you want is something that maybe you just want to kind of have some kind of tool that enables you to **visualise the data very quickly to see whether it's what you want**."*

- "...looking at distributions. So *how do the patterns vary spatially and visualising that?"*

User Research Summary Finding 10
(WP2_and_5_High5forAI_UXResearchSummary.pdf)

RECOMMENDATION
Visual Previews: provide quick-look visualisation tools within the portal:
Build an interface or API which allows users to preview using features such as thumbnails, summary plots, spatial coverage, sample rows, variable distribution. This would allow users to verify data relevance, before committing to a download, and allow for rapid relevance checking without writing custom code, before moving large data volumes.

FINDING 8
LEGACY DATA RECOVERY IS A MANUAL, HIGH-EFFORT BOTTLENECK

A significant portion of valuable data (historical climate, seismic tapes, maps etc.) are trapped in non-digital formats or obsolete media. Recovering this data is not a technical API problem but a manual investigative process involving hunting down former employees and physical media. Any strategy for "comprehensive" data access must account for this non-automatable, human-intensive phase of data rescue.

In the survey, several open-ended responses and statistical trends highlight the symptoms of missing or inaccessible legacy data and imply the difficulty of bridging old baselines with new needs. In Q1 (Role Diversity): the presence of roles like "Hydroclimate extremes," "Natural catastrophes," and "Soils" suggests a user base heavily reliant on longitudinal historical data (decades of records) to model trends, **making the absence of legacy data a critical failure point.** Further on, in Q6 (Biggest Challenge): **multiple respondents explicitly identified the lack of legacy data as a primary barrier.**

User voice:

- *"Data publication is patchy and people have only been making data available for about the last 20 years so there's a huge lack of legacy data. Some people don't bother to publish their data."*
- *"Just finding the most up-to-date datasets... While it is good to have a baseline comparison for 1961 to 2010, the data is not useful in the wider impact assessment [if recent data is missing]."*

We questioned this further during the qualitative interviews. Users described the extreme lengths they go to for historical data, which surveys alone might not have captured in detail.

User voice:

- *"Field-tapes of seismic... bottleneck for us... a lot of the time what's missing is the actual field data, the raw data that was originally acquired before anything's been done to it.*
  And it's really that that you need in order to extract the most information from it. Because if you want to reprocess it and use all the modern algorithms and you have to go start from the beginning and take it..."

- *"I use **historical records**, future projections and forecasts."*

- *"I was tasked with collecting all the registers of green space for every local authority... there were 22 ways to record this data. So harmonising that data then at across anything at scale sounds really challenging, yes... different formats, you know, people sending me data sets in PDF and you think no, what are you doing?"*

- *"One of the biggest like just first steps that would make a huge difference... in combination with accepted standards, to be able to say, **even if it's identifying what you don't have in a historical data set**, that means that it can't meet that standard... you're able to plug a spreadsheet into an AI tool that comes out and says right, you're missing glare, you're missing wave state... that would be incredibly handy."*

- *"We noticed that because of the **reanalysis data** with the gauges, there's like discrepancies. So we always need to bias correct all the data that we're seeing, like also bias correct to ERA5 data and everything. So I think this is... something that we see as a big*

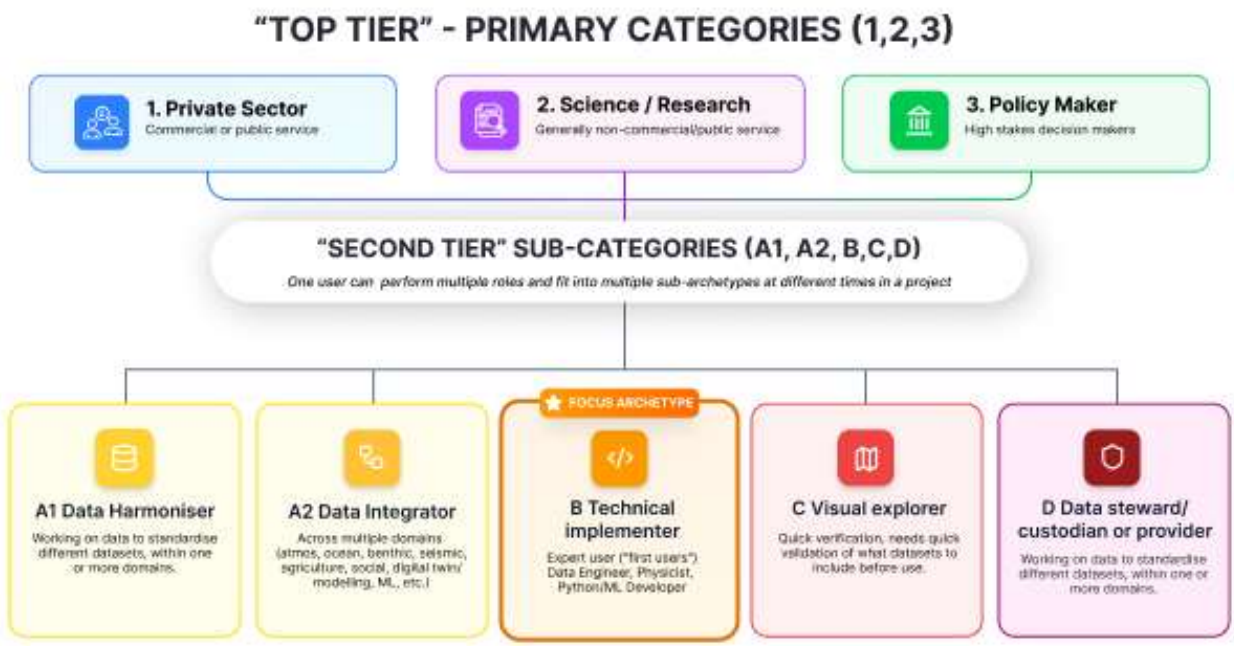
challenge." [Context: "Reanalysis data" often combines historical observations with models. The need to manually "bias correct" these datasets because they don't match current gauge data illustrates the high effort required to recover and use historical/legacy information.]

User Research Summary Finding 9
(WP2_and_5_High5forAI_UXResearchSummary.pdf)

User Archetypes

We created User Archetypes from the research: by classifying user types we spoke to, based on their use cases, needs, pain points and characteristics.

Generally, the users we spoke to fit into one of the three “top tier” categories (1,2 or 3) and then within this, fell into one or more of the “second tier” categories (A1 or A2,B,C,D). One user can perform multiple roles, and therefore fit into multiple sub-archetypes at different times during their career, or even on a project.) **See diagram below:**



Based on this we determined that our “typical" High5forAI user (or "focus user group”) is:

Based on the research and user archetypes, the most likely ‘type of user’ who will utilise the High 5 for AI services is:

TOP TIER

2. Research Professional

Generally non commercial, academic researcher or scientist

SECONDARY TIER

B. Technical Implementer

Potentially also: data integrator or harmoniser (A1 or A2)

This is because they are the individuals responsible for creating, building or implementing standardised data formats and APIs. These users are most likely to build new connections or pathways to different datasets from across different services, and are most likely to use an API to facilitate this.



THREE PRIMARY USER ARCHETYPE CATEGORIES:

1. Private Sector: commercial or public service

"particularly in Europe [we] have to bring all the data together into one place before you can actually get a full picture... and then you can see the whole data and make more useful products."

Workflows and needs

- Cannot freely use APIs: security policies prohibit external calls
- Must download data manually or use "approved endpoints only"
- Data ingestion more complex, often 4+ datasets used in conjunction to arrive at results required
- Would like more open-source SDKs made available across commercial entities

Pain points

- (Currently restricted) access to commercially sensitive datasets would create efficiencies.
- API access blocked by organisational firewalls and vital security measures.
- Multi-dataset harmonisation is entirely manual (eg: spreadsheets, CSV files)
- No standard method(s) to evaluate risks on assets (across industry)
- Data retrieval (esp. physical records) add effort & time to delivery

Goals / Desired outcomes:

- Multi-dataset harmonisation for required outputs
- Working within approved data endpoints only
- Complex data ingestion across multiple sources
- An increase in open-source data sharing between Private and Public sector

2. Science / Research: generally non-commercial/public service

Workflows and needs

- Open data-source tools (Python, Jupyter, HadCM3 APIs, HBF3)
- Prefers downloadable bulk datasets, API calls useful for specific datasets (but less often)
- Expects full metadata transparency
- Actively merges, cleans, re-formats datasets and looks for automated pipeline functionality
- Believes data should be openly accessible
- Working under budget constraints, often cannot afford paywalled/commercial datasets

Pain points

- Often frustrated by fragmented discovery platforms ("data silos")
- Dataset sizes (NetCDF/HDF5) too large, no compression guidance.
- Downtime when receiving data (eg: API changes, observation stations going offline, non-continuous across certain areas)
- Lack of metadata makes ML/DL dataset comparison significantly challenging.

- Fragmented discovery: no unified source across data providers and repositories
- Gridded data is inconsistent

Goals / Desired outcomes:

- Performing standardised ML/DL datasets, accessible via API calls
- Automated, harmonised analysis pipelines (development and maintenance)
- Trustworthy, credible open-access data
- Clear provenance and processing history of data products

3. Policy Maker: high stakes decision makers

Workflows and needs

- Only acts on data with clear provenance, peer-reviewed methods, and authoritative sources (eg: Sea Office, ECMWF, ERTRE)
- Needs dashboards, summaries, or scenario zones, not raw datasets or model code
- Uses or requests data from external sources (data must be peer reviewed, provenance is vital)
- Time constrained: wants rapid, reliable insights
- Traceability and accuracy are critical: decisions affect public safety, policy, budgets, and occasionally legal compliance

Pain points

- When it comes to any modelling (including ML/AI) - model accuracy is questioned
- No widely-available tools to translate scientific outputs into actionable policy currently in use
- "Spreadsheet hell": no aggregated dashboards or scenario summaries
- Systems across agencies (even within the same industry eg: water/utilities) are not interoperable.
- Size of project and geographical distance challenge workflows (distance, conflicting schedules, language etc)
- Mental models differ across different disciplines (eg: terminology for "policy people" vs "science").

Goals / Desired outcomes:

- High-stakes decisions on environment (eg: habitats), resilience (eg: climate), resources (eg: fishing quotas) etc.
- Traceable data and evidence-based decision making (eg: policy development, regulatory impact assessments)
- Stakeholder consultation with reliable, rapid insights

SECONDARY USER ARCHETYPE SUB-CATEGORIES: A1, A2, B,C and D

Note: One user can perform **multiple roles** and fit into **multiple sub-archetypes** at different times in a project.

A1 Data Harmoniser

Working on data to standardise different datasets, within one or more domains.

- Suffers when datasets lack consistent schema/format (spends time ensuring data is harmonised). More data cleaning focus)
- Frustrated by fragmentation (desires consistency and interoperability)
- Occasionally needs to access legacy data (eg: from the 70s–90s, on magnetic tapes, paper logs, or in non-standard formats)
- Manually traces ownership, contacts former employees, hunts down physical media (USBs, tapes), and submits formal requests to data owners
- This is manual and time intensive: relies on human networks, persistence, and domain knowledge (ideally would like “one unified source of data”)

A2 Data Integrator

Across multiple domains (atmos, ocean, benthic, seismic, agriculture, social, digital twin/modelling, ML, etc.)

- More technical focus: standardisation of data infrastructure / architecture
- Performs cross-discipline calibration, combination, and harmonisation - serves as the connector and translator between communities and disciplines
- Bridges scientific modelling and decision-making (manual)
- Frustrated by fragmentation. Needs cross-discipline harmonised metadata/data (combines environmental, geospatial, atmospheric, meteorological datasets eg: integrating seismic, oil/gas, government, and/or NCAR data into workflows)
- Licensing costs block open access to subsurface datasets
- Occasionally mines legacy data (File-70k, magnetic tapes, paper logs, finding documentation)

B Technical implementer

Expert user ("first users") Data Engineer, Physicist, Python/ML Developer

“Spent a week figuring out how to use the (API) key…"

• Utilises machine learning and/or data modelling deliberately for specific, interpretable tasks (eg: clustering, uncertainty estimation).
• Lack of metadata makes ML dataset comparison virtually impossible (requires domain clarity to interpret columns, variables, or classifications).
• Writes code collaboratively with version control and shared environments: a pain point is a lack of metadata, calibration info, or QA/QC (cannot validate data).
• Pre-processes datasets to reduce coding overhead, integrates APIs and engineers’ data pipelines: therefore API integration complexity, missing documentation is a challenge
• Balances performance/resolution, cost, and environmental impact

C Visual explorer

Quick verification, needs quick validation of what datasets to include before use.

- Archetype looks for patterns in data (sometimes using visualisation), needs quick verification.
- Wants tools for rapid prototyping and sharing (to "plot a map" without scripting, web maps, tile services)
- Often not a subject matter expert (eg: Policy maker, "end end user)"
- Needs pre-processed datasets.
- Currently no streamlined export for client-ready visualisations
- Desire a "a single, trusted portal" to find, download, or API-access data (eg: "data.com", "ArcGIS Online APIs for proximity tools", "Geopandas.explore")
- Sometimes has to review maps, documents (eg: PDFs, PowerPoints) for quick client-facing comparisons (consultancy end use)
- Catalogues and secure data sharing access are desirable (eg: Delta Sharing, Unity)

D Data steward/custodian or provider

Working on data to standardise different datasets, within one or more domains.

- Curates, validates, translates data to ensure fitness, reproducibility, defensibility, value.
- Trust-focused: Validates with 2–3 independent sources.
- Needs tools for rapid validation (eg: cross-source comparison, bias correction, advises policy but doesn't make it).
- Technically "digitally fluent": Uses Python, R, SQL, Databricks, APIs
- Builds custom tools, often employed by data Provider.
- Data detective: spends significant time sourcing, validating, and reconciling data from multiple providers (Copernicus, ERA5, government sources, universities).
- Deadline-driven: works under short timelines (2 mnths.), needs tools that reduce ingestion time.
- No cross-agency tools for shared state or real-time/emergency coordination. Works across regions (EMEA, Africa, US) - data availability gaps are a challenge.

Ecosystem service map

This is a **stakeholder ecosystem map** that visualises the relationships between different user groups, the core challenges they face (Connection Hubs), and the specific unmet needs or "gaps" identified during the High 5 for AI research.

It effectively tells the story of **why** the current system is failing: users are stuck in silos, struggling with fragmented data, and relying on manual workarounds because the tools don't match their needs. The diagram is organised around "thematic areas" (blue “connection hubs) that define the user journey:



Legend	
Blue Connection Hubs	These represent the central themes or problem areas where multiple user groups converge. They are the "pain points" of the ecosystem.
Green User Types	These are the individuals interviewed and the archetypes we developed based on the research.
Red Identified Gaps	These highlight critical failures in the current system: areas where user needs are not being met.

Core thematic areas/connection hubs:

A. Data Access & Availability (Top Left)

- **Who:** Students and the Private Sector.
- **The Problem:** Data is often **"Restricted or expensive."**
- **The Gap:** There is a specific **"Private sector gap"** related to **security restrictions**. As noted in the interviews companies need compliant solutions that allow them to bring data within their secure firewalls and therefore can’t access external resources (so APIs cannot be used).

B. Data Fragmentation & Integration (Centre)

- **Who:** Scientific Researchers, Data Harmonisers (Type A1), and Data Integrators (Type A2).
- **The Problem:** Data is **"Siloed & incompatible."** Users must manually stitch together datasets from different sources that don't speak the same language (different formats, units, naming conventions).
- **The Gap:** The diagram highlights **"Technical users bridging data to applications"** as a gap. This confirms the interview finding that highly skilled humans are currently acting as the "glue" to make broken systems work, rather than having automated tools do it.

C. Tooling & Workflow (Bottom Left)

- **Who:** Development Teams
- **The Problem:** Workflows are **"Manual."**
- **The Gap:** There is a desperate **"need for automation tools."** This aligns with the survey result that "Cleaning the data" is the most time-consuming task. The diagram also points out that **"APIs are catered to [Technical Implementers and Data Integrators],"** suggesting a mismatch where APIs are built for coders but fail to serve the broader needs of non-technical users or heavy-data users who need downloads.

D. Trust & Provenance (Top Right)

- **Who:** Policy Makers and Data Custodians (Type D)
- **The Problem:** Decisions rely on integrity, but **"Need trustworthy data."**
- **The Gap:** There is a **"Data custodian gap"** (need for curation/validation tools) and a **"Policy maker gap"** (time-constrained, needing rapid insights). This illustrates the tension: Policy makers need fast answers, but Custodians can't provide them quickly because they lack the tools to validate and curate the data efficiently.

E. Communication & Usability (Bottom Right)

- **Who:** Environmental Consultants, Local Authorities, and Visual Explorers (Type C).
- **The Problem:** Data comes in **"complex formats"** and models are too difficult to interpret.
- **The Gap:** There is a **"communication gap"** in translating **"complex models to actionable insights."** Consultants and local authorities need simple visualizations to make decisions, but the current output is often too technical or opaque.

Pain-points clustering (*affinity mapping)*

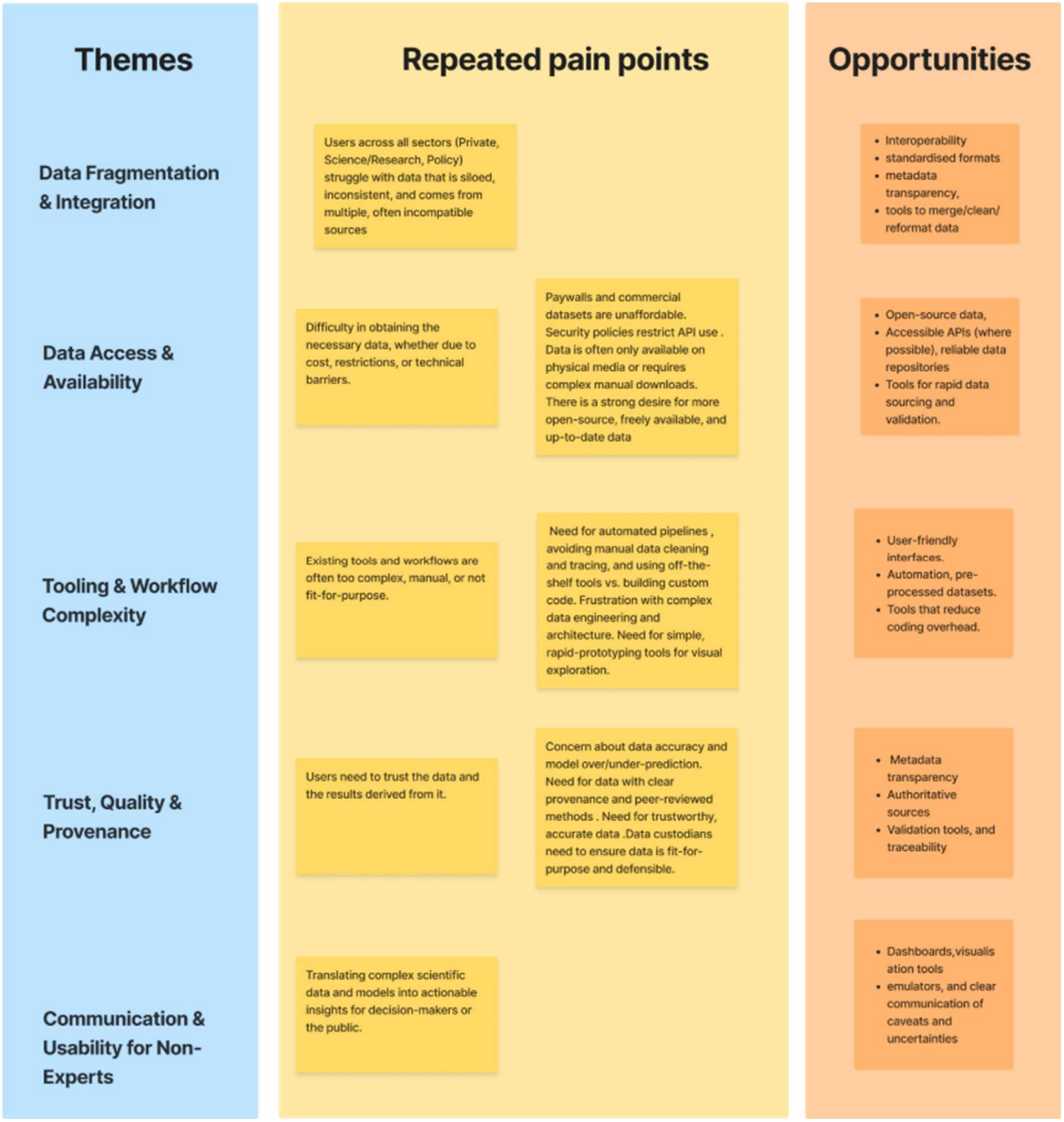
The synthesis of the High5 for AI survey and in-depth user interviews helped to cluster the pain-points and emerging user needs. The matrix shown below translates research findings into an "opportunities" roadmap by mapping five core Themes (Data Fragmentation, Access, Tooling, Trust, and Communication) to specific Repeated Pain Points and corresponding opportunities.

As initially outlined on Page 10 and 11 of this report: users across all sectors are currently bottlenecked by siloed, inconsistent data and manual, complex workflows that hinder productivity and erode trust in model outputs.

The primary opportunity lies not merely in deploying advanced AI, but in fixing the foundational data infrastructure.

Implementing interoperable standards *or* building infrastructure in a way that is standard-interoperable (by working on the technical architecture layer to enable interoperability between different standards) and automated cleaning tools to resolve fragmentation, providing secure, open

access/ access with clear licensing and usage conditions to overcome availability barriers, and developing user-friendly visualisations with clear provenance tracking to bridge the gap between complex scientific data and actionable decision-making for non-expert:



User Stories

What are user stories?
User stories are development tasks often expressed as “archetype person + need + purpose.” They are concise, user-focused descriptions of desired outcomes that drive collaboration and value in development teams. They help teams understand the "why" behind their work and break down large projects into practical outputs and manageable tasks.

- **Stories keep the focus on the user.** A to-do list keeps the team focused on tasks that need to be checked off, but a collection of stories keeps the team focused on solving problems for real users.
- **Stories enable collaboration.** With the goal defined, the team can work together to decide how best to serve the user and meet that goal.
- **Stories drive creative solutions.** Stories encourage the team to think critically and creatively about how to best solve for an end goal.
- **Stories create momentum.** With each passing story, the development team enjoys a small challenge and a small win, driving momentum.

Source: Atlassian

Purpose of user stories on the High 5 for AI project:
We wrote these user stories based on our research (our findings and understanding of the challenges, pain points and needs of the Users we spoke to). These user stories can be used to align the High 5 team around features that address real user needs by clarifying goals and benefits.

They also establish a reusable foundation for future teams to build upon this research – any other development team can pick up any of these user stories and decide to build new solutions knowing they will address a real-world problem.

Each user story is based on a typical user role [based on the Archetypes from users we spoke to diagram on Page 27] and has been cross-linked to Findings in the report.

The user stories can be found on Page 5 in the UXResearchSummary.pdf

Methodology in Detail

PRIMARY USE CASES

These use cases defined the target audience and survey design focus (from EDS DRI Funding Proposal).

Use Case 1: Clean Energy

"Support UK net zero targets through improved data access for decarbonisation and energy security."

- Focus: Offshore wind farm siting
- Key datasets needed: Bathymetry, sediment composition, wind resource, weather patterns, sea level, ocean physics
- Stakeholder need: Consistent, integrated data access (referenced: NORA report)

Use Case 2: Blue-Green Infrastructure for Public Health

"Support NHS Fit for the Future mission by enabling planning of healthy built environments."

- Focus: Access to open spaces, water, and air quality
- Key datasets: Landcover maps, sea level, air quality, geological maps
- Goal: Enable planning of parks, housing, beach access to support mental and physical health

EXPLANATORY SEQUENTIAL USER RESEARCH

PART 1: QUANTITATIVE SURVEY (WORK PACKAGE 2)

Due to time constraints within the project timeline, the team opted for a rapid market-style user survey rather than in-depth interviews or ethnographic studies. This survey was conducted to inform:

- How users currently access, use, and struggle with EDS datasets
- What data formats, tools, and workflows are most common
- How AI is being adopted, and where barriers exist
- What would make data truly "AI-ready" and trustworthy

Goal: Design data services that reduce friction, increase value from insights, and empower users, from researchers to policy-makers.

Platform: Conducted via Hubble (trial version), a GDPR-compliant platform that automatically anonymises user data.

Deeper engagement: Participants were given the option to provide their email at the end of the survey to be contacted for follow-up moderated user research interviews.

Collaborative design: The survey questions were reviewed and refined by all Work Package Leads to ensure alignment with project goals and user needs.

DURATION

- Launch date: 19 November 2025
- Initial closure: 19 January 2026 (after 40 business days)
- Reopened briefly: 27 January 2026 to collect 2 additional responses
- Final total: 43 responses

SURVEY CONSTRAINTS

The project's scope was shaped by funding priorities and work package focus.
Final qualitative survey sample: 43 survey respondents. While diverse, not statistically representative of all potential users. We primarily reached out to audiences known within our existing networks and did not use any external recruitment agencies. Based on the NOREX framework, we focused on the "Regular" and "Occasional" users. The "Never, Extreme or X" users were **not** well represented in the sample.

- **Data access:** Cannot mandate changes to existing data centres or formats - research focused on how to interface with them, not how to restructure them.
- **Security & sensitivity:** Cannot override commercial or governmental data restrictions - recommendations focus on secure, anonymised sharing pathways

SURVEY DISSEMINATION

Initial reach: approx. 2,500 pax. (across academia, private sector, and government).
To ensure diverse and relevant responses, the survey was distributed using multiple channels, targeting professionals aligned with the two use cases.

Digital channels:

- EDS Website: https://eds.ukri.org/projects/high5-ai
- EDC Website: https://ukerc.rl.ac.uk (courtesy Peter Holt, Energy Research Unit)
- LinkedIn: Shared by Ola Podgorska (NOC), Carl Watson (BGS), and reshared within professional networks
- Online platforms: Blogs, websites, and relevant forums

Direct email outreach:

- **Custom distribution list:** Curated by Work Package Leads and Lead Researcher, targeting individuals aligned with the two use cases.
- A survey log spreadsheet was maintained throughout the project during the survey period *(please note: this is an internal document, and can only be shared with participating Organisations Research Project Members, according to the User Research Consent forms signed, to ensure compliance with GDPR)*
- JASMIN user list: Distributed via Jesse Alexander (NCAS)

Events & conferences:

- UK Environmental Exposure (UK-EEX) Hub workshop
- MEDIN Monthly meeting
- Data Sharing Working Group (x2 meetings)
- Digital Twins of the Ocean Event – Ditto Ocean Decade

SURVEY STRUCTURE

19 questions covering:

- Role and expertise
- Data usage frequency and sources

- Challenges in discovery, access, and formatting
- AI readiness, trust, and workflow integration
- Preferred formats (CSV, NetCDF, Parquet, etc.)
- Interest in co-design

PART 2: QUALITATIVE INTERVIEWS (WORK PACKAGE 5)
Moderated interviews are structured, one-on-one conversations between a researcher and a participant, guided by a predefined set of questions — but flexible enough to explore unexpected insights, emotions, and context. Unlike surveys, they allow for deep, qualitative understanding of user behaviours, motivations, frustrations, and workflows. In the High-5 for AI project, moderated interviews were used to complement survey data - going beyond "what users do" to understand "why they do it," "how they feel about it," and "what they really need."

User Research:
Ola Podgorska, Shwetha Raveendran

Duration:
Interviews Conducted: December 2025 – February 2026

Participants:
- 12 professionals across 10 organisations, including Marine Spatial Planners
- Data Scientists (Met Office, Ordnance Survey)
- Environmental Consultants
- University Researchers (Swansea, Manchester Met)
- Private Sector (insurance, seismic, fisheries)

Format:
- Online, 45-minute - 1 hour sessions
- Semi-structured guide (flexible to follow user's narrative)
- Recorded with consent (GDPR-compliant)
- Transcribed and anonymised

Focus Areas:
Deep-dive into workflows, decision-making processes, and pain points related to the two use cases. **Please see: High54A_UserInterviewDiscussionGuide.pdf**

- How users currently find, access, and use EDS datasets
- Pain points in workflows (discovery, cleaning, integration)
- Attitudes toward AI, APIs, and trust
- Use cases: offshore wind siting + blue-green infrastructure

Synthesis:
Transcripts were **manually** coded (by the Researchers) using the affinity mapping method in Figma.
Link: User Experience Research Synthesis Board

When you click on the link, the Moderators will be able to provide you view-access to the file for review (we review access on an individual basis).



Affinity mapping: insights were manually grouped into:

- Challenges / Pain Points
- User Needs
- Technical Requirements
- Benefits / Value Add

Document was supplemented by AI-assisted summarisation of the data with Euria AI. Euria is hosted in Switzerland, leaves no trace in ephemeral mode (*"Nothing is collected for training. Any data exchanges are encrypted in transit and on our servers in Switzerland"),* and is powered with renewable energy. 100% of the energy use is revalued and heats thousands of homes.