ENVIRONMENTAL DATA SERVICES

HIGH 5 FOR AI

High5 for AI

NERC Environmental Data Service

Report number OR/26/050_subreport4

User Research Summary Report

FEB 2026

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

Increase the availability of high-value data in accessible, AI-ready formats through the development of APIs.

- Showcase the use of such data provision through the creation of demonstrator AI applications (such as an agentic AI chatbot) that would enable easy access and utilisation of complex data through an innovative interface

- Learn from and engage with the wider community so the EDS can deliver value beyond the environmental science community by a broad range of disciplines and sectors to address key government mission challenges

To understand how end-users currently find, source and use datasets, and to understand if the development of APIs and AI tools (such as an agentic AI chatbot) would benefit them in accessing EDS datasets.

For **WP2** a survey approach was chosen to ensure user research insights were delivered with time to utilise.

For **WP5**, to better understand what brings value to a broad range of disciplines and sectors beyond the EDS, we focused on 1-2-1 moderated user research interviews.

Work Package 2

Quantitative Survey

Over 2,500 potential end-users were reached to engage through: direct individual emails and email groups (JASMIN), conferences and meetings (DITTO, Data Sharing Working Group, MEDIN and more), blog posts as well as the EDS website.

- 41 responses
- Anonymised
- Not incentivised
- Hubble survey platform

Work Package 5

Moderated user interviews

11 potential end-users and stakeholders working in the environmental field, across 10 different organisations.
Our focus was on two use cases (offshore wind farm siting and blue/green spaces). These interviews were recorded for research synthesis, and in line with GDPR, consent forms were signed before each interview.

- 11 online moderated interviews
- Anonymised
- Not incentivised

APPROACH

Research synthesis

We colour coded insights from interview transcripts. Using both our manual research synthesis process (transcribing > colour coding > affinity mapping), in combination with the AI summary tool in Figma, we **collated feedback and compiled user stories based on:**

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

Direct (anonymised) quotes from research interviews have been provided where relevant.

BASED ON OUR RESEARCH (Findings - Page 17)

Some user stories for High 5 for AI

FORMAT
As a < type of user >,
I want < some goal >
so that < some reason >

- As a Technical Implementer, I want to access data through a single, well-documented API or cloud-native data lake (e.g., Parquet) that supports direct querying, so that I can avoid the friction of downloading large files and can integrate the data directly into my scalable cloud-based analysis environment.
  [Finding 3: High 5 for AI - our focus user]

- As a Marine Spatial Planner, how can I trust that this API or digital tool is verified, so that I know the data is trustworthy and credible for use. [Finding 4 & 5: Data is often scattered, republished, or re-hosted by third parties (e.g., on ArcGIS Online, GitHub, or other platforms). It's difficult and time-consuming for researchers to determine if the data they've found is the original, authoritative version or a potentially outdated or modified copy. This creates uncertainty and risk for their analysis.]

- As a Data Scientist, I want to understand the provenance and processing history of an AI-generated data product or insight so that I can assess its reliability and explain its limitations to stakeholders, ensuring that I don't blindly trust a "black box" output. [Finding 2 & 5: lack of transparency, uncertainty around which original data sets (or versions of them) were fed into the AI model, unknown processing history the ability to assess how realiable a dataset is becomes more difficult. Blockchain / logs and UI explicitly showing audit trails could be used to trace API origin and updates].

- As a Geospatial Scientist, I want to access environmental data in a consistent, machine-readable format (like NetCDF, Parquet) with complete metadata (e.g., schema.org, DCAT, which includes publisher, version, and provenance) so that I can integrate it into my Python-based workflows without spending significant time on manual cleaning, reformatting, or deciphering the data structure. [Finding 1, 2 & 3: inconsistent, poorly documented data formats that force scientists to spend excessive time on manual data wrangling instead of analysis. It specifically tackles the pain point of having to clean, reformat, and reverse-engineer data structures due to missing or unclear metadata, which hinders reproducibility and efficient integration into automated workflows. For example, an API can return structured metadata (schema, units, spatial bounds, temporal range, resolution, source) + a small, representative sample (e.g., 100 rows or a 1km² tile) in a lightweight format like GeoJSON or CSV.]

- As a Climate, Hydro-climate, or Hydrological scientist I want to combine disparate environmental data sets (e.g., satellite-derived NDVI, in-situ sensor data, model outputs, and potentially historical data) into a single, harmonised analysis pipeline so that I can focus on answering my research questions rather than on the technical challenges of data integration and transformation. [Finding 4 : scientists waste significant time wrestling with incompatible formats, structures, and scales across disparate data sources (satellite, sensor, model, historical) . They often have to manually integrate data, which takes time and effort away from actual research.]

- As a Data Scientist (Data Harmoniser/Integrator), I want to contribute my own processed, high-quality data (e.g., from a local observatory or specific project) to a central, authoritative repository so that it can be easily discovered and used by other researchers, thereby increasing the impact of my work and fostering collaboration.
  [Finding 4 & 9: a lack of a trusted, standardised, and incentivised mechanisms / curated platforms for researchers to share their own processed or locally generated data.]

- 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. [Finding 9: data discovery is not, and has never been a purely technical or search-based task, but as a social and collaborative act, and is often found through human networks, not just APIs or metadata. A tool supporting this might include features like: expert directories with domain tags, “Ask the community” forums or Q&A, data sharing through trusted connections with provenance tracking, integration of institutional project histories to surface “who worked on what, where etc.]

- As an Atmospheric researcher working on air pollution modelling, I want to use AI/ML tools for data cleaning and quality control (e.g., identifying missing variables, inconsistent units, or metadata gaps) so that I can efficiently assess the fitness of a data set for my analysis and reduce the risk of introducing errors into my models.
  [Finding 2: inconsistent, incomplete, or poorly documented data quality, which forces researchers to manually inspect and clean datasets before use - and increases risk of introducing errors into models]

- As a Technical Implementer working on large-scale, compute-intensive environmental modelling (e.g., climate, hydrology, or air quality), I want to access high-volume, high-resolution environmental data directly within my secure, cloud-based analysis environment (e.g., via a cloud-native data lake in Parquet format with direct query support), so that I can avoid the friction and security risks of bulk downloads and the fragility of traditional APIs: enabling scalable, stable, and reproducible analysis without being disrupted by external service changes.
  [Finding 6: Issues such as firewalls, security risks, and the sheer volume of “big data” (terabytes) mean that many organisations (especially in the private sector) still prefer bulk downloads or direct "in-place" cloud access where they can run proprietary models securely.]

- As a Research Professional, I want to quickly visualise the key variables, spatial coverage, and temporal range of a dataset through an interactive web interface or API endpoint before downloading it, so that I can assess its relevance and quality without writing custom code or transferring large files: enabling faster, more informed decisions on whether to invest time in deep analysis. [Finding 10: An API can power this user story by providing lightweight, on-demand data preview capabilities eg: pre-generated thumbnails or summary plots.)

TOP LINE RECOMMENDATIONS

1. Prioritise "harmonisation of data" through synchronisation. Instead of enforcing rigid, static "standards" that people often resist or fight over, **focus on tools which support the harmonisation of disparate datasets**. For example, building an AI-supported platform to develop ontologies and infer column meanings can bypass human ego-clashes and manually intensive "cleaning" processes to provide a foundational tool for the standardisation of various datasets.

2. Because practitioners currently express a resounding "No" when asked if they trust autonomous AI results, **any new tools or systems must include auditable trails, signposts to code, and multi-model comparisons to build confidence.** Define the use-cases the tool will be able to support, be clear about which tasks it helps to solve, and exactly how automation or AI plays a part on the platform's landing page, or throughout the user-flow of the interface.

3. Recognise that personal networks and institutional relationships are currently the primary engines for finding high-value data. **Reframing data sharing as a social and collaborative act, and incorporating this into the new tool - to enable or support a collaborative flow -** will help bridge the silos between tech experts, scientists, and policy-makers.

4. **Signposting and how the tool is promoted is just as important as the tool itself.** Fragmentation is a universal pain point, especially in Marine and Global South. "Single source of truth" is the #1 wish from a lot of the users we spoke to. Users will only use a tool if they know it's there, and trust it .

5. **Trust = provenance + standardisation.** For users to trust any digital tool or database, they need to know it's from a trusted source, and know it's secure/safe to use. APIs are not enough - users also need to know the original source of the data and understand *what* they will get (eg: metadata, documentation), before they even think about connecting (to an API) or downloading a dataset.

6. **Confidentiality blocks data sharing - especially in fishing, social and insurance industries.** Extend the reach of the EDS, by building connections with Industry, Organisations and Environmental Data creators - beyond NERC - using this new tool as an invitation: to support the creation and connection of previously disconnected datasets. This will build on the future vision of UK's "one single source of truth."

Research Summary

FEB 2026

HIGH 5 FOR AI

NERC Environmental Data Service

OUR FOCUS USE CASES

To focus the user research remit on the project, we recruited users who worked with these two specific environmental use cases:

DATA USE CASE 1
Clean energy eg. siting of wind farms to improve decision making related to offshore renewable assets.



Image source

DATA USE CASE 2
Access to blue-green infrastructure for benefit to public health. Inform decision making that enables the planning and provision of built infrastructure.



Image source

FEB 2026

HIGH 5 FOR AI

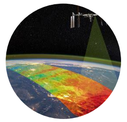
QUALITATIVE USER RESEARCH INTERVIEWS: WHO WE SPOKE WITH

Role	Use case	Notes	L1 User archetype	L2 Archetype categories
1. Research Scientist, Met Office	One	Quantifying near urban exposure to climate hazards under socioeconomic pathway/focus: green belts. Integrates multi-source data.	**2. Science/Research**	**A, B, C, D**
2. Private Insurance Broker	One	Sources, cleans, validates data for clients; uses Databricks, APIs not used	1. Private Sector	A, B
3. Fisheries Data Manager, Marine Institute Ireland	Two	Manages fisheries data, produces advice for Gov/EU, confidentiality & data privacy constraints.	**2. Science/Research**	A, D
4. Marine Spatial Planning, Marine Institute, Ireland	Both	Sources fragmented marine data, "single source of truth" would reduce time-to-source, QMF/FAIR principles applied to data.	**2. Science/Research**	**A, B**
5. Technical Project Data Coordinator, Marine Institute, Ireland	Both	Builds APIs, prototypes, data pipelines; focuses on technical architecture, data access.	**2. Science/Research**	**B**
6. Managing Consultant, Private Seismic Data Consultancy	Two (seismic surveys)	Using data from surveys, cross-referencing acquisition reports, matching seismic to wind farm locations, identifying gaps in historical data and recovering data from archival formats eg: tapes, multi-role eg: SME infrastructure tasks (eg: IT)	1. Private Sector	**A, B**
7. Epistemologist, Swansea Uni	One (linking environmental and health data)	Sources, processes, and links environmental data with health data, builds exposure models, interprets for policy.	**2. Science/Research**	**A, B, D (although mentioned C favourably)**
8. Senior Data Scientist, Ordnance Survey	Two	Sources, processes, and publishes geospatial data; leads automation/ML in data capture; data custodian.	**2. Science/Research**	**A, B, D**
9. Marine Data Manager, Crowne Estate	One (marine delivery, Data Exchange Platform)	Ingests, validates, and models multi-source environmental data for offshore wind, data custodian.	**2. Science/Research**	**A, B, D (Org level)**
10. Geomorphologist, Manchester Met. Uni.	One (digital twins / river & waterways modelling)	Builds digital twins; sources multi-source data; leads model development; University data custodian.	**2. Science/Research**	**A, B, D (Org level)**
11. Private Sector Consultant (Strategic)	One (water industry)	Focuses on standardisation of data architecture (ontology, not direct data use), acts as strategic consultant for digital infrastructure.	1. Private Sector (Consultant)	A, D
12. PHD Student	Both	Planetary Boundary Layers. Uses Python to process/combine data (NetCDF, HDF5, text files) for technical implementation. Data harmonisation.	**2. Science/Research**	B, A, C (uses APIs for quick validation, not analysis)
13. Senior Hydroclimate Scientist, UKCEH	One (Water Resources and Drought Group)	Leads data production. Manages institutional data outputs (Org level). Works with collaborators on non-public data (bridges data gaps). Analysis for technical implementation.	**2. Science/Research**	**B, D. C (uses Python for rapid visualisation)**
14. Machine Learning Datasets Team, Met Office	One (climate modelling)	Data harmonisation and custodian. Focuses on making data "ML-ready" (Zarr, STAC, CF-NetCDF). Not a direct "technical implementer" (delegates coding), drives data infrastructure.	**2. Science/Research**	A, D

Science & Research users



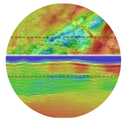
Climate, hydro-climate, or hydrological scientist
Integrate multiple datasets into workflows (modelling). Need interoperable, harmonised datasets, with good metadata. **Hydrological need near-real-time data (gauges, rainfall, tides),** whereas **hydro-climate often requirel historical records as well** as projections /forecasts.



Atmospheric Researcher (incl. Air pollution modelling)
Forecast/Data validation: needs historic datasets, metadata completeness, and comparability to validate back-trajectory. Needs cross-border access; suffers from fragmentation across countries (similar to global ocean modelling). Formats: NetCDF, HDF5; needs compression (dataset size is a challenge).



Epidemiologist (combining geography, epidemiology, and data science)
Sources NDVI (Landsat), DEFRA air pollution, Met Office heat data, biodiversity, flood zones, blue space (all to link with anonymised health data). Uses Python, Google Earth Engine, Parquet files, machine learning (neural networks) to model urban heat islands and indoor air temperature. Builds custom code to link environmental + health data (not off-the-shelf tools).



Seismic data (acquisition and processing) Specialist
Offshore renewables, oil and gas eg: NDR, Met Office, Hydro Portals, Crown Estate etc.) Locate, assess, compare and download large seismic datasets, avoid manual detective work to locate missing datasets. Need: SEG-Y file formats, field data (raw tape or even better, digitised).



Geospatial Scientist
Sources aerial imagery, third-party data (e.g., address, change intelligence), and aggregates it into digital products. Works in cloud environments (Databricks, Snowflake). Uses machine learning (semantic segmentation, classification) to automate feature extraction from aerial imagery. Builds pipelines to transform raw data into cloud-native formats (Parquet).



ML/AI specialists
Remote sensing, model emulation. Seasonal/sub-seasonal prediction once long datasets exist. Needs high-quality (data) inputs, high quality metadata, for comparison of data sets.



Data Scientists / Consultants (part of Data capture team)
Customer facing: working with customers, helping them use (our) data, solve problems (works at Government, Institutional, Environmental/Private Agency)



Science fellow(s), operational “support” user
Not managing a team but connecting cross-cutting themes and projects.
Uses environmental or geospatial data to support real-world decisions (land use, siting, planning, risks), verification of models. Eg: *“Help identify where land use could be changed, where it can't be changed because of sort of specific designations, how you might change it if you wanted to achieve certain objectives.“*



Central government department scientists / researchers
Supports / collaborates with Government policy makers & others needing to understand climate impacts, hazards, and trade-offs in planning contexts. **Need high spatial resolution climate data** (eg: end use - home insulation, need resolution more granular than 5m - *“resolution limits direct use for detailed decisions.”*). Relies on trusted data sources (e.g., government shapefiles, ECMWF’s ERA5).

Wider ecosystem users



Government data services and data owners
(NDR, Met Office, Hydro Portals, Crown Estate etc.)
Intermediate stakeholders: manage and store national datasets, respond to data upload requests, provide consistent metadata and formats.



Institutional data service organisations
Some of these are end-users of each other's data.
Eg: MetOffice uses NOC's tidal data, ONS uses OS data etc



Environment agencies, National regulators (eg: UKRI)
Require high-accuracy data inputs and model outputs, used to issue warnings. Concern about over- or under-prediction.



Local Authorities / City Councils (flood risk, planning, emergency planning)
Shared view of environmental risk data. Avoid misinterpretation or miscommunication - sometimes without specialist skillsets in science. Dependency on data that's difficult to obtain.



Offshore wind farm site planners
Reliable, up to date subsurface information for turbine foundations, spatial information. Need: Up to date shape files, license-free datasets (open source), subsurface datasets



Government / policy makers (National and local ministers, civil services etc.)
Stakeholders responsible for national planning, net-zero strategy, resilience, regulation and land-use policy. Translating scientific outputs into actionable policies. Need high-level, aggregated insights (not hyper-local resolution). Interested in trade-offs (e.g., carbon vs. flood risk) and scenario planning.



Academic partners
Universities, Institutes: includes academics producing datasets, collaborating on modelling, or contributing expert review.



Environmental consultants (Private sector, internal and external)
Consultants hired by clients (asset owners, government, forestry, infrastructure) to provide assessments, often contracted via collaboration networks.



Development / architecture / engineering (teams and individuals)
Pipelines, formats, interoperability, ability to automate workflows.



University students (eg: PHD)
Needs free/open source or subsidised access; faces paywalls; expects open data; struggles to obtain restricted datasets. Uses dashboards for quick verification of datasets.



Emergency services (Fire, rescue)
Need tools which can support emergency services before, after and during events.
"something that they can actually deal with at that time."



Public (end users of blue/green infrastructure)
Affected by outcomes of infrastructure decisions, but often not included in the decision making process.

Identified user archetypes

Users typically fit into **only one of the Level 1** archetypes, but can fit across **one or multiple of Level 2** archetype categories.

Level 1

1 OR 2 OR 3

Private Agent
(Private sector or Government, typically commercial and/or public service)

Research Professional
(generally non commercial, academic researcher or scientist)

High-Stakes Decision Maker
Policy decision makers, institutional leader or public authority

AND

One user can perform multiple roles, and thereby fit into more than one sub-archetypes at different times:

Level 2

A1 **Data harmoniser**

A2 **Data integrator**

B **Technical implementer**

C **Visual explorer**

D **Data custodian (Org level)**

L1 in detail

Note: Users typically fit into **only one of the Level 1** archetypes

1

Private Agent
(Private sector or Government, typically commercial and/or public service)

Core behaviour: niche consultancy within limited time frame, commercially driven

- Cannot freely use APIs (security/firewall policies prohibit external calls).
- Must download data manually or use “approved endpoints only”.
- Data ingestion more complex, often 4 or more datasets used in conjunction to arrive at results required - need: data harmonisation
- Would like more open-source data to be made available across different commercial entities

2

Research Professional
(generally non commercial, academic researcher or scientist)

Core behaviour: mission / hypothesis / remit driven

- Uses open-source tools (Python, Jupyter, NetCDF, HDF5).
- Prefers downloadable bulk datasets vs API calls.
- Actively merges, cleans, re-formats datasets.
- Expects full metadata transparency.
- Cannot afford paywalled or commercial datasets.
- Strong belief that data should be openly accessible.

3

High-Stakes Decision Maker
Policy decision makers, institutional leader or public authority

Core behaviour: high-stakes decisions

- Only acts on data with clear provenance, peer-reviewed methods, and authoritative sources (e.g., Met Office, ECMWF, DEFRA).
- They use data from custodians, but don’t own the national data infrastructure.
- Needs dashboards, summaries, or scenario scores
- Time constrained: can’t wait for custom analysis; needs rapid, reliable insights.
- Accountable: decisions may affect public safety, budgets, or legal compliance — so traceability and accuracy (of data) is critical.

L2 in detail

Note: One user can perform multiple roles, and thereby fit into more than one sub-archetypes at different times on a project.

A1

Data harmoniser

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

- Deadline-driven: Works under short timelines and within budgets (e.g., 2 months), needs tools that reduce ingestion time.
- Applies quality standards (e.g., QMF, FAIR) even though they don't own the data.
- Suffers when datasets lack consistent schema/format.
- Often creates/ensures harmonised metadata.
- Prioritises data consistency, transparency, and interoperability.
- Frustrated by fragmentation - Wants unified discovery platforms "one unified source"
- Has to process 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.
- Data cleaning focus: cleanup/standardisation of data itself

A2

Data integrator

Works across multiple domains (eg: atmos, ocean, oil + seismic, agriculture, social, digital twin/modelling, ML etc.)

- Combines environmental, geospatial, atmospheric, meteorological datasets eg: integrating seismic data (from oil/gas, government, NDR) into renewable energy workflows (bridging industries).
- Needs harmonised metadata.
- Bridges scientific modelling and policy/ decision-making (often serves as the connector between scientists, policymakers, and technical teams)
- Frustrated by fragmentation - Wants unified discovery platforms "one unified source"
- Often requires 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.
- More technical focus: standardisation of data infrastructure / architecture

B

Technical implementer

Technical implementer, expert user ("first users"). Core behaviour: balances performance, cost, and environmental impact (e.g., minimising data storage/transfers)

- Data Engineers, Python/ML Developers
- Utilises machine learning and/or data modelling. Uses AI/ML deliberately for specific, interpretable tasks (e.g., clustering, uncertainty estimation).
- Creates/builds/implements standardised data formats and APIs.
- Pre-processed datasets to reduce coding overhead.
- Needs metadata / measurement metadata, calibration information, QA/QC.
- Cannot trust or validate data without context.
- Requires domain clarity to interpret columns, variables, or classifications.
- Suffers most when metadata is missing, inconsistent, or vague.
- Works within constraints to define design: resolution, cost, licensing

C

Visual explorer

Quick verification, visual pattern analysis and understanding

- This archetype contrasts sharply with Scientists using datasets
- Uses tools for rapid prototyping (e.g., "plot a map" without scripting), web maps, tile services)
- Would like a single, trusted portal to find, download, or API-access data (e.g., "data.com", "ArcGIS Online APIs for proximity tools", "Geopandas.explore")
- Looking for patterns in data (using visualisation)
- Often not a subject matter expert (eg: Policy maker, "end end user)"
- Sometimes has to review maps, documents (eg: PDFs, PowerPoints) for quick client-facing comparisons (consultancy end use)

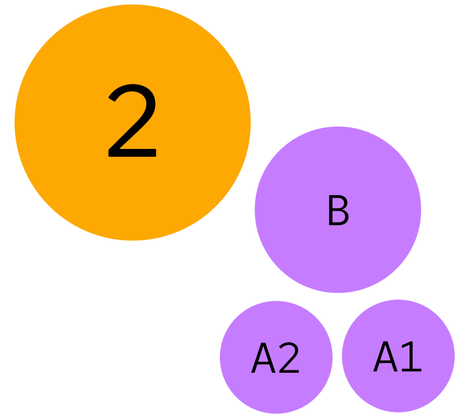
D

Data custodian (Org level)

Data custodian or steward, Organisational data provider

- Core behaviour: curate, validate, and translate data, ensure it's fit for purpose, reproducible, and defensible (even when data is fragmented, inconsistent, or politically volatile).
- Technically fluent: Uses Python, R, SQL, Databricks, APIs, and often builds custom tools.
- Data detective: Spends significant time sourcing, validating, and reconciling data from multiple providers (Copernicus, ERA5, government sources, universities).
- Trust-focused: Prioritises authoritative, curated sources (e.g., government, peer-reviewed). Validates with 2–3 independent sources.
- Global need: Works across regions (EMEA, Africa, US), adapts to data availability gaps (e.g., Africa's sparse rain gauge data).
- Needs a single, trusted portal to find, download, or API-access data (e.g., "data.com").
- Standardised formats, units, and metadata (FAIR principles).
- Tools for rapid validation (e.g., cross-source comparison, bias correction).
- Owns and (often) is funded for the platforms data is provided through
- Advises policy but doesn't make it!

High5 for AI focus user



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

Archetype Level 1
2. Research Professional - generally non commercial, academic researcher or scientist)

Archetype Level 2
Technical implementer, potentially: data integrator or harmoniser

This is because they are the individuals responsible for creating, building or implementing standardised data formats and APIs.

These users are most likely to want to build new connections or pathways to different datasets from across different services, and are most likely to use an API to do this.

INSIGHTS: DATA CONTEXT

How data use context differs

We've found that depending on the type of science, the type of data (standard or legacy) and the User Archetype, individuals use different methods and processes to approach their tasks. This will be useful to be more focused in any future user research or digital product design, as being specific as to which kind of data and which user archetype the tool will be aimed at, will increase the chances it will be useful, and useable for the target users.

Standard vs Legacy data approaches

Standard Data	Legacy Data
Sources data from APIs, portals, or public repositories	Hunts for data from 70s–90s, stored on tapes, paper, or lost in corporate archives
Compares metadata, formats, dates	Reads acquisition reports, contacts former employees, submits formal requests
Uses scripts, tools, or APIs to clean/ transform	Manually downloads, transfers, and restructures data. Often with no automation
Works with modern, machine-readable data	Works with non-standard, poorly documented, or "unreleased" data
Focuses on interoperability	Focuses on archival data, making lost data usable again
Time-consuming but somewhat automated	Extremely manual, slow, and dependent on human networks
Hydroclimate - need recent datasets	Hydrological - needs legacy data

User Archetype approaches

By User Archetype	Sub Archetype A1 Harmonising	Sub Archetype A2 Integrating
Scope	Within a single domain (e.g., atmospheric data)	Across domains (e.g., climate + infrastructure)
Complexity	Medium (format/structure alignment)	High (semantic, thematic, system bridging)
Output	Clean, aligned dataset (columns, variable naming etc)	Unified analytical framework or pipeline
User Type	Analysts, modellers	Interdisciplinary researchers, AI/ML teams
Tooling	Python, CDO, xarray	STAC, Zarr, GeoParquet, AI frameworks. Manual comparison, spatial mapping, metadata sleuthing.

DEEP DIVE

Moderated Research Interviews Findings

Finding 1 - Foundation standardisation and interoperability needs are not being met Pg 18

Finding 2 - Context is king: without foundational metadata, we're not "AI data ready" Pg 20

Finding 3 - A shift towards cloud-optimised and machine-readable formats was identified Pg 21

Finding 4 - The desire for a "single source of truth" Pg 22

Finding 5 - Trust, transparency, and the "human-in-the-loop" requirement Pg 23

Finding 6 - API limitations Pg 24

Finding 7 - Gatekeeping and sensitivity of critical environmental datasets Pg 25

Finding 8 - Regional data dependencies and inequity Pg 26

Finding 9 - Overlapping knowledge and community sourcing Pg 27

Finding 10 - Visualise and understand what's *inside* the dataset *before use* Pg 28

FINDING 1

Foundation data harmonisation and interoperability needs are not being met.

Standardisation is seen as a pre-requisite for the delivery of public services and scientific advancement - but datasets often require laborious manual "cleaning" before they can be integrated or used (because of the differences between individual contributors, organisational differences, and even industry-wide differences).

Inconsistent data structures, grids, naming, and units were the most cited barriers to effective data use. However, experts mentioned this focuses on the wrong level, noting that "data architecture and infrastructure" matter more than standardising formats, which "will change all the time."

"Particularly in critical infrastructure... the thing that's become increasingly clear is that lack of standardisation is really getting in the way of effective and good value delivery of public services."

"Doing that on a sector by sector basis with a view to joining those sectors together, the challenge that you have is that a lot of people just don't understand - people are not understanding what the problem is, because they don't understand how the infrastructure is or what it does..."

Tech people understand really well what the problem is, but they don't see it necessarily as a problem because they know how to fix it. They know that you need to integrate the. They know how to do it, and they're not really listened to (by managers)...."

The manual labor currently required to link metadata as a barrier: "sometimes linking in the mesh [metadata] data, having the correct vocabularies... having some of those automated workflows in place just to speed up the data availability would be great."

"Even something as simple as trying to connect those two [first name and surname] columns together across hundreds or thousands of data sets [is] really complicated and time consuming."

"You would think they'd have some standard process for capturing their data. They don't."

"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."

“I’m talking about standardisation, not standards. What you find is that a lot of people are emotionally committed to certain ways of describing the data than others. They fight over it. So then the process of standardising, gets delayed because people can't agree, that’s why I’ve started to use AI. Because AI is a really good way of developing a standardised set of ways of describing the data that is based itself on...

What a good thing to feedback to the team is to make sure that they are standardising across their data model design and development and that they're synchronising because they're they're in a good place to provide proof of concept best practise - and to deliberately introduce (standardisation) into their process...”

Jessica Symons - visioninglab.com

FINDING 2

Context is king: without foundational metadata, we’re not AI ready

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 is a prerequisite for AI and machine learning readiness.

Data isn’t currently ready for API operations due to the gaps & lack of ‘depth’ of metadata information. APIs don’t exist - which is what High 5 for AI is trying to address - but users have also shared there’s a lack of confidence in using them - it is something that’s desirable, but difficult to achieve).

“So looking at things like data spaces where you could have some secure way or sharing data with trusted partners or making certain other views of it more publicly available. Sounds like it would be the the right way to go, but I don't foresee it happening in the near future anyway”

“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's just a faff. 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... [they] go through either our NWP standardisation or our climate standardisation and that makes them CF compliant and puts them into net CDF format. So these are better, **but they're still not ideal, certainly for the certainly for training on large data sets, very large data sets.”**

“all the metadata has to be quite ideally it's quite well formatted and and populated... it has a massive impact on the, on how the data will behave when we start doing our work on it.”

“Cultural change... *I'm worried about the please, can you run this at the end? Once you've run your model, can you just run this [metadata] so that we everybody knows what it's about?”*

“...calibration parameters and everything should be given as a metadata. Then only we can process the data and come with the final output.”

“double checking, triple checking metadata helps" [when resolving issues with incorrect data copies]”.

“If everybody was using Medin discovery metadata and then pushing their data to a central environmental data portal, that would be incredibly useful."

FINDING 3

A shift towards cloud-optimised and machine-readable formats was identified

Although traditional formats like CSV and NetCDF are still widely used, there is a strong pivot toward cloud-native, analysis-ready formats such as Parquet and Zarr. These formats are preferred for their scalability and ability to handle large, multi-dimensional data sets without the friction of downloading and re-uploading.

Experts noted that "AI-readiness" requires these formats to allow for efficient streaming and processing in machine learning frameworks.

"I'm trying very hard to push for usage of sort of the equivalent format GeoParquet... a pandas data frame would fit nicely into a GeoParquet."

"Currently we don't publish or make available our data in cloud native formats, but we're we're sort of on that path and that's **in response to user demands for that."**

"For machine learning the most useful thing would be being able to stream data and it being in a format that you can stream it in."

"A lot of our N dimensional data would pivoted towards our primarily sort of as a ideal is a cloud storage format and analysis ready cloud storage format... Zarr is much better [than NetCDF]."

"Probably most like the the well, so for machine learning the **most useful thing would be being able to stream data and it being in a format that you can stream it in. So so stuff that can be read by**

X-Array basically and that will integrate well with **FS spec**. Net CDF isn't too bad, although also ZARR is much better, yeah."

"Parquet is very popular cloud native format because it's very scalable and it's very flexible... it's optimised for distributed processing."

ECMWF Anemoi... Anemoi is a framework developed by ECMWF in partnership with pretty much older member states, in fact, and it is. It's what they, you know, primarily developed for them to run their AI operational forecasts, so AIFS....

FINDING 4

The desire for a "single source of truth."

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.

Datasets are sought in an ad-hoc, individual manner. Everyone we spoke to had their own ‘favourites’, and would spend time networking within their community in the search of authoritative data sources - because each project requires different, specific datasets. An open-source centralised, curated portal or aggregator with a national or global overview of available data was cited as desirable although was also noted to be a “naive” consideration (due to the complexity of creating such a portal). This isn't necessarily about hosting all environmental data **in one place**, but about having a searchable catalogue that directs users to authoritative sources with clear provenance. **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).**

> “No, I just want one place. I just want heredata.com... I just want one one place just to go there, just be happy."

> “They are like scattered, so even within UK if we will be able to integrate this data set and form a common platform... placing them at one location... that will be very much helpful.”

> “In an ideal world, I think geopolitics aside, **we should have data [open] all the time**. I know it’s a pipe dream... U.S. government shut down and basically we have lost all U.S. data...”

> So from our side of it, the challenges in Ireland are that there's **no consolidated architecture for all marine data.** There's talks of it in the pipelines for many, many years of having a solution for this, but nothing has been funded.”
>
> "Something that’s missing in this for me is the single sources of the truth for an API... it’s very easy to pick up a different source... I would probably like to see a limited sources... [a] bank of organisations that we're harvesting data from."

> "Naively, what you really want is everything in one place... you almost need some kind of trusted third party... a portal that collates the trusted data that you can then access from the trusted source."

FINDING 5

Trust, transparency, and the "human-in-the-loop" requirement.

When asked if AI results are trusted, a resounding “No” was often the answer (even by those who build their own machine learning models). Despite the the excitement around potential, there is a deep-seated reserve regarding the accuracy and reliability of AI-generated outputs. Interviewees consistently argued for a "trust but verify" approach, where automated recommendations are checked by human experts.

Verification through multi-model comparisons: trust is built through the ability to replicate across multiple models, if different models or data sources converge on the same pattern, confidence in the result increases. **Concerns about AI "hallucinations" and the "black box" nature of some models mean that transparency (eg: auditable trails, sources, ensuring replicability) are vital for scientific *and* legal assurance.**

“Because you don't trust one model, you always have to have multiple models... in order to get for instance uncertainty in certain ranges.”

“We were very careful to... create a pipeline that basically for every data set we know exactly what we did to it and we've got... signposts to the code that we used for doing that"

”It would be really easy to just kind of chuck a load of data at an algorithm... and it looked kind of realistic, but it may have been complete rubbish.”

“In systems like ArcGIS Online, data published can be downloaded and republished by anyone (which can be misleading) - it may appear legitimate but it’s not controlled or maintained by the original source. **AI currently lacks this verification capability unless explicitly programmed to do so.”**

"If you can have the provenance carried through, it's not so much a problem if someone republishes your data as long as it's marked as the original collector"

“Something that’s missing in this for me is the single sources of the truth for an API... it’s very easy to pick up a different source... how can you ensure an AI does the same thing?”

“The way that you can use AI effectively in these contexts is to get it to analyse the data and make recommendations which you, the human then decide. That’s the human in the loop.”

FINDING 6

API limitations

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 (especially in the private sector) still prefer bulk downloads or direct "in-place" cloud access where they can run proprietary models securely.

In addition, the "fragility" of APIs (code breaks when a provider updates their service) is a major concern for long-term project continuity and stability. Some Senior data scientists noted that the industry is moving away from traditional APIs toward "cloud-native" direct access.

However it was noted that in ***“some cases APIs are viable. It's a viable option depending on the scale and complexity of that data set.”*** It therefore stands to reason that clarifying and explicitly naming who we are targeting the APIs to (who the intended end-users are, which datasets it’s intended for) will be beneficial to the uptake of the tool.

“We are not relying on an API for highly detailed bathymetry data because it's just... it can't handle it. It's it's not a realistic way of doing it.” **[Big Data]**

“not APIs, I don't think that that would work because that we're talking terabytes worth of data and... you need to pull the whole data set down into our onto our data centre to, **to load it into our specific software... i**t’s just not realistic for them to utilise an API for this stuff. They need to have have basically offline access for it to be able to run this stuff." **[Access]**

“getting some even one degree resolution data is too heavy... I don't know whether it will be applicable with APIs.” **[Big Data]**

“I think there's less justification now to build API traditional APIs and I think our I feel like users prefer download or the ability to sort of like just access in place in the cloud." **[User preference]**

"for me to download data or use a local copy on Jasmine... is much easier because I know that method, than doing the API type of stuff." **[User preference]**

“We haven't done much with APIs. Again, firewalls aren't great for us... getting stuff unblocked is occasionally a pain.” **[Security]**

“We're trying to keep the data secure that we have, but by this building a firewall which isn't always working with all the APIs that are available.” **[Security]**

“the reality [is] really quite different. There are multiple different APIs, different formats. Some of the formats changes during the project...” **[Ongoing support / maintenance]**

“spending a week trying to figure out how to use the key [for the API]... which is nuts. Waste of time, yeah.”

“There is some checking that we still need to do in an APIs. In case APIs change over time, they go broken or an updated one gets... a new API instead... That’s a big risk." **[Access]**

FINDING 7

Gatekeeping and sensitivity of critical environmental datasets

Commercial sensitivity (e.g., from oil companies, water utilities), geopolitical shifts (e.g., changes in US data availability) and cost to transfer/host (eg: NDR field tape data) hinder and prevent access to critical data.

The "gatekeeping" **by private utilities or commercial entities** and "stonewalling" of requests for specific datasets (like marine activity data) - **even when it is essential for public good or research projects -** creates significant blind spots in scientific research and planning.

Researchers also highlighted how changes in international policy and government decisions can lead to the sudden loss of vital datasets (additional quote on open-data on page 8).

On water utility data: “One of our main source of data we would have liked was from was from the local water utility, OK and... No chance. OK, there’s just 0 chance of of get getting hold of any of any of that, any of that data... Or [data] that is commercially sensitive..."

“We're being stopped by being able to share it at a useful resolution because of confidentiality concerns"

“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"

“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”

“For certainly for us, the main challenge is always around data confidentiality. So because we're often dealing with commercial fisheries data... there’s issues around personal data and also the business confidentiality around the data.”

“We were working on a project... across the US on flood plains... but we couldn't access the data anymore because it was just just shut shut down”

“governments don't want to publish the pollution data sets or water quality data sets because they're worried that people will get angry”

FINDING 8

Regional data dependencies and inequity

Environmental research using regional datasets from Africa and India faces specific hurdles, **ranging from a physical lack of observation infrastructure to restrictive government paywalls and a precarious reliance on international systems.**

Researchers noted that in Africa, data is often "scattered" or inconsistent, with a heavy dependence on American satellite data that can be arbitrarily shut off, causing local operational services to fail. Especially in the area of climate and strategic environmental science - these issues are going to be a rising concern going forward because the field is increasingly dependent on global observation systems that are vulnerable to political shifts and funding instability**. Geopolitical data interruptions are particularly damaging to AI workflows, which require consistent data streams to maintain reliable predictive environmental models.**

"We're slowly working on projects in Africa and I noticed there's a big lack of data sources... especially when it comes to what do because we want to assess for instance like a drought metric within Africa."

“Operational systems that were giving climate services and like these rural parts of Africa have gone down because they were based off Some American satellite data that they just shut off for no apparent reason.”
[Geopolitical challenges]

“The agriculture data is so suspect. We’re getting low [confidence]... it’s the input data... I have seen those maps. It doesn't look right."
[Data reliability]

"Indian government IMD data set we will not get for free... some archive data sets are free but not the current ones."
[Paywalls]

Limitation: it’s not possible to address this finding on the High5 for AI project (out of scope).

FINDING 9

The real engine of data discovery: overlapping knowledge and community sourcing

The interviews reveal that personal networks and institutional relationships are one of the **primary mechanisms through which scientists actually locate, trust, and successfully integrate environmental data.** Practitioners often act as "*data detectives,*" following a network trail with colleagues and former partners to find historical or fragmented datasets that are not listed in official catalogues. The practitioners also reported benefit when overlapping expertise on projects allowed for a shared language across different vertical structures. **Reframing data sharing as a social and collaborative act, and building a tool which could support this collaborative flow could be beneficial.**

"You are heavily reliant on individuals within certain research spaces and topic experts to be able to go to you and say...

I know that the Sea Watch Foundation in North Anglesey has a particularly good bit for that area of the Irish Sea and actually you need to go and talk to these people... **there is observation data that otherwise we would simply have no idea they existed."**

"The thing that we found really worked was having overlapping sort of expertise... building that kind of shared understanding despite using different terminology and different ways of thinking"

"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..."

“The things that are particularly problematic at the moment is that there are decisions that may need to be made by humans and we do not and and we collectively aren't... able to make them very well.”

“I personally also ask people... you would know people around you who have worked on certain projects... That is one of actually very good sources of like just asking members around.”

“That was basically through community knowledge... it was if we could find what's the standard data set that people use in this community... that's what we would use.”

"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

NERC Environmental Data Service

FINDING 10

Users want to visualise and understand what's *inside* the dataset *before use*

"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 to visualise the data very quickly to see whether it's what you want".***"

This is desirable because it would enable basic operations without having to move data around a lot, and allows for quick, visual pattern comparison before a 'data deep dive' or statistical quantification.

"in some cases APIs are viable. It's a viable option depending on the scale and complexity of that data set. And that can be really useful sometimes because obviously it's a really cool and you're just adding a layout - makes it so much easier to to bring this data in"

"We want the European view, so **we need the downscale versions of the regional climate models.**"

"something like that that would enable you to do some basic operations. Without having to move the data around a lot would be would also be very useful I think"

"...probably looking at distributions. So how do the patterns vary spatially and visualising that? **Then I would probably go and have a look and try and quantify that statistically anyway.**"

"The thing I found difficulty in them, I cannot use multiple datasets...

Only one data set at a time, which we can analyse.... **so if we need to combine 6 months of datasets, and do some arithmetic on them and visualise them.** So it was quite difficult because it's a heavy software..."

"We will initially visualise the data quickly."

"Whatever we imagine, **we're able to visualise with the data...**" [Need]