

DataLabs and Discoverability

A Summary Report on the DataLabs Enhancements Project

David Philip Green, Carolynne Lord, Kelly Widdicks, Jennifer Roebuck, Lily Gouldsbrough, Mike Hollaway, and Gordon Blair

Introduction

The NERC Digital Strategy 2021-2030 outlines a vision for a “*culture that places data and digital technologies at the heart of current and future UK environmental science,*” ensuring “*all areas of environmental science are well positioned to capitalise upon the transformative potential of data and digital technologies*” (NERC Digital Strategy, 2022). This vision depends on the advancement of Digital Research Infrastructure (DRI), and its ability to support the collaborative, integrative and systemic environmental science we need to collectively meet UK environmental targets. In aid of this, **DataLabs** is a core element of DRI development in UKCEH and across the NERC Environmental Data Service (EDS).

DataLabs is a cloud-based digital platform designed to support collaborative, open, transparent, and data-driven science (Hollaway et al, 2020). It allows users to flexibly combine multiple elements within one integrated ‘environment’, and for this reason it is sometimes classified as a Virtual Research Environment (VRE). For the avoidance of confusion, *DataLabs* is the name of both the *platform* and individual *projects* (*DataLabs* within *DataLabs*). Each project integrates a unique combination of digital assets – data, models, methods, code, visualisations, etc. – which are then used to answer environmental research questions and support decision making. In this report, we refer to the *platform* as *DataLabs*, and *projects* using the shorthand term, ‘*Labs*’.

DataLabs has been maintained by the UKCEH Environmental Data Science team since its release in 2017. In 2023, a co-design and agile software development methodology was adopted to facilitate ongoing, iterative updates to DataLabs. The agile process is overseen by two ‘product owners’, whose shared role is to prioritise development requirements, to “*maximize the value of the product*” (Scrum.org). The aim of the move to an agile methodology is to support the team to develop and improve the user experience (UX) of DataLabs and the various digital assets it supports, so that – in turn – it can better support collaborative, integrative and systemic science.

A key challenge facing DataLabs is that the kinds of digital assets it works with tend to be fragmented, heterogeneous, and inconsistent, and they can be challenging to locate, interpret, and integrate. While some datasets and digital research outputs have a unique domain object identifier (DOI), making them easier to find and work with, most digital assets do not. Asset fragmentation not only obstructs systemic science, it also contributes to a duplication of efforts, as pre-existing research outputs remain undiscovered. We characterise this as a **discoverability** challenge, one that is slowing the advancement of environmental science, wasting resources, and increasing DRI’s environmental impact (Bird et al. 2023).

In this report, we draw on a two-part study that explores future directions for enhancing discoverability *within* and *through* DataLabs: The first part of the study is based on recent engagement with stakeholders in environmental science at UKCEH, conducted as part of a wider co-design strategy to better understand how DRI might be developed in the future. Taking a ‘bottom-up’ approach, this part of the study explored users’ perspectives on DRI and how these infrastructures could be enhanced in the future, including improvements to DataLabs and discoverability. The second part of the study, which reflects a more ‘top-down’ approach presents insights from the development of an

experimental Large Language Model (LLM) designed to improve the discoverability of digital assets in DataLabs.

‘Discoverability’

The term ‘discoverability’ is often used to describe how easy it is to locate new information within a system. The Interaction Design Foundation (IxDF) operationalise the term as “*the ability for users to encounter new content in a product, that they weren’t aware of previously*” (IxDF, 2024). Improving discoverability has tremendous potential value for open and systemic science (Paic, 2021) in facilitating new connections between digital assets, users, and other stakeholders such as scientists, decision-makers, and wider publics.

The current gold standard for open scientific data is encapsulated by the **F**indable, **A**ccessible, **I**nteroperable, and **R**eusable (FAIR) principles (Wilkinson et al, 2016); these principles easily extend from data, to include ‘digital assets’ more broadly. Discoverability draws from some of these principles, in relating to concepts of ‘**f**indability’ (the extent to which users can find what they are already looking for within a system) and ‘**a**ccessibility’ (the extent to which a system is usable anywhere, anytime, and by anyone). However, whilst ‘accessibility’ and ‘findability’ are well-defined terms and benefit from common standards (e.g. WAI-ARIA for web accessibility (W3C-WAI, 2024)) and plentiful examples of best-practice, discoverability has proven to be more difficult to define (McElvey & Hunt, 2019). This raises important questions, including:

- *What are the impacts on science if development efforts focus on FAIR, but not on discoverability?*
- *Must scientists already know what they want to find before they set out to find it?*

The fragmentation of digital assets has led to a “discoverability crisis” with numerous practical challenges (Kraker, 2021). Many of these amount to findability and accessibility shortcomings, but discoverability implies broader, contextual challenges, such as navigating unfamiliar disciplines (with unfamiliar vocabularies, ontologies, and publication cultures), keeping up to date with new data and new publications, and filtering out digital assets that are *not* relevant to a given context. Our aim here is not to resolve this crisis. Rather, we respond to the challenge of improving the discoverability of digital assets within DataLabs, as well as the discoverability of the platform itself. Through this process, our aim is to generate actionable insights that can help drive the open science movement forward through enhancements to DataLabs.

Methods

We draw from a recent case study of UKCEH to uncover current and future visions of DRI for environmental science. The study engaged UKCEH environmental scientists, data managers, software developers, and other staff, in semi-structured interviews, workshops about ‘FAIR’ assets, and a survey. The dataset we are drawing from includes 28 interviews, 81 survey data responses, and 36 workshop participants across 4 workshops. Given the overall size of the dataset, we offer a preliminary and illustrative (rather than representative) analysis of the key issues relating to digital assets in DRI (broadly), as well as DataLabs (specifically), concerning discoverability. These insights are derived from a deductive read-through (Bingham & Witkowsky 2022), supplemented with a simple search across the data for relevant terms (e.g. ‘discover’, ‘find’, ‘access’, ‘locat-’, ‘search’, ‘DataLab’, ‘lab’).

Additionally, to explore whether there is a technological solution to discoverability challenges that could be incorporated into DataLabs, we also carried out an experiment exploring the capability of a Large Language Model (LLM) to improve the discoverability of datasets in the Environmental

Information Data Center (EIDC). This involved comparing search results from traditional metadata catalogue searches with the LLM search to explore which returns the most relevant and useful results.

In the remainder of this report, we first share our findings from the qualitative research, and then follow with the results from our experimental LLM prototype. We, then, draw from these to present a typology of discoverability and some initial recommendations for enhancing the discoverability of DataLabs, both through its user experience and in relation to the broader contexts of digital research infrastructure and scientific research culture. We conclude with a list of opportunities for DataLabs enhancements, and actionable insights derived from our findings and discussion.

Findings

In our survey, we drew upon the FAIR principles to ask respondents to reflect on the **findability** (Figures 1+2: top) and **accessibility** (Figures 1+2: bottom) of digital assets across DRI.

Overall, 52.7% of survey respondents ‘agreed’ or ‘somewhat agreed’ that environmental science ‘data,’ ‘methods,’ and ‘models’ are currently findable, and 39.2% ‘agreed’ or ‘somewhat agreed’ that they were accessible (Figure 1). However, nearly all of participants agreed that they should be made more so, with 86.5% agreeing or somewhat agreeing that they should be more findable and 83.7% agreeing or somewhat agreeing that they should be more accessible (Figure 2).

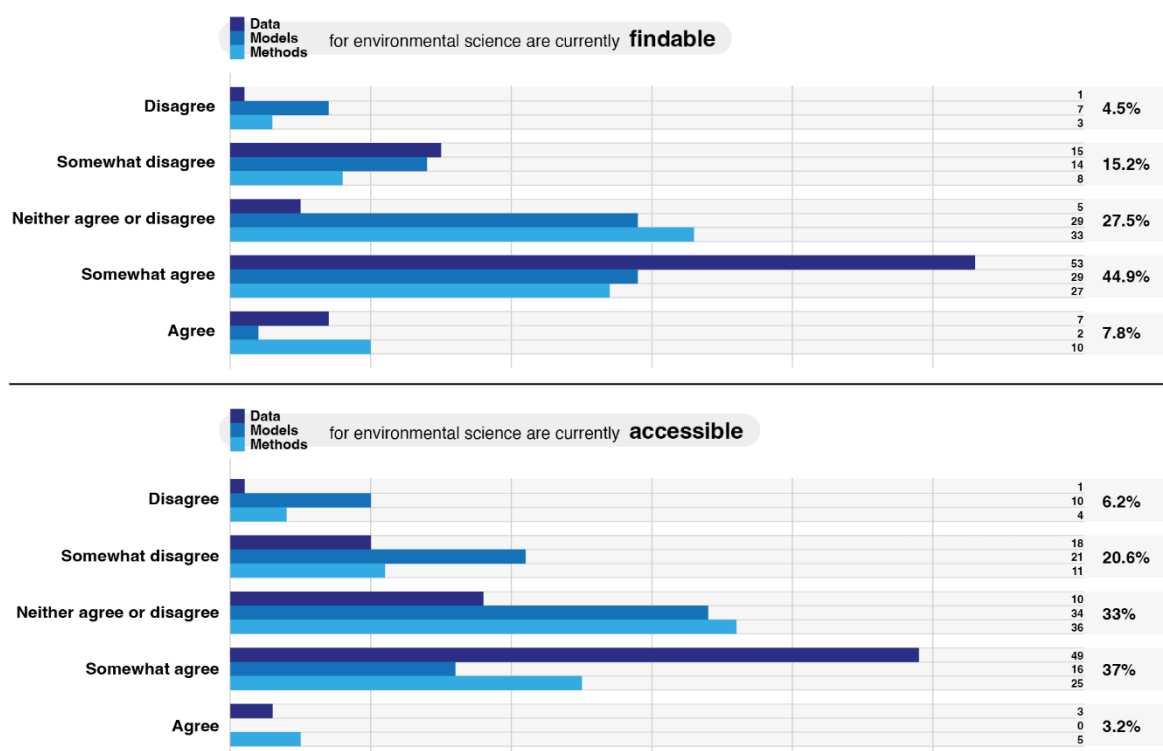


Figure 1: A horizontal bar chart showing responses from our survey regarding the current **findability** and **accessibility** of three kinds of digital assets: *data*, *models*, and *methods*. Respondents were asked to record their response to each statement using a five-point Likert scale from ‘Agree’ to ‘Disagree.’

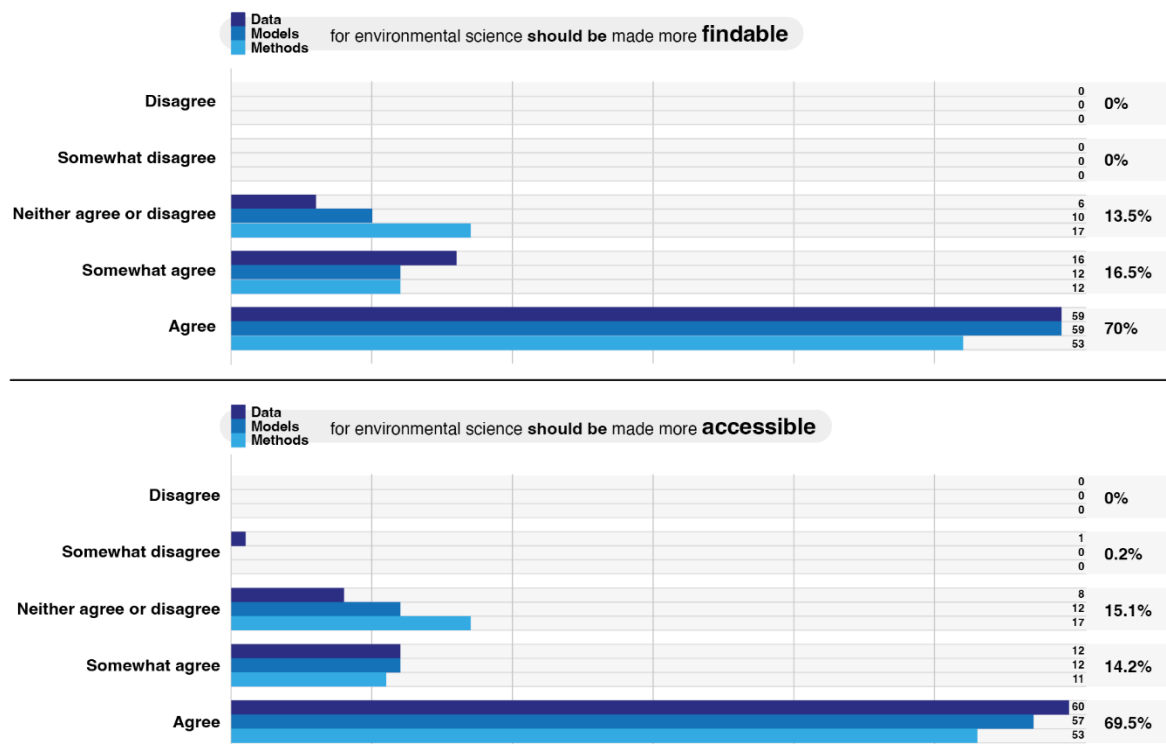


Figure 2: A horizontal bar chart showing responses from our survey showing whether the three categories of digital assets - *data*, *models*, and *methods* - **should be more** findable and accessible. Once again, respondents were asked to record their response to each statement on a five-point Likert scale from 'Agree' to 'Disagree.'

These findings suggest that existing DRI is working *to an extent*, but it also demonstrates that there are improvements to be made, and hints at areas where future development of DRI - and DataLabs specifically - could be targeted. For instance, it is notable that *data* were considered more findable and accessible than *methods* or *models* (none of the respondents firmly 'agreed' that models were accessible, for instance), so extra effort may be needed to improve the findability and accessibility of methods and models.

The idea that there is scope for improvement was also reflected in the workshop data. Participants discussed "*the joy of finding data that you need*", the opportunities of the principles for "*finding novel things*", having "*more time for science*" as well as the effect they could have for "*opening up science*". At the same time, participants noted that "*we are at the start of the [FAIR] journey*", and that environmental science still "*has a way to go*". Various tools were noted as advantageous in the process of reusing data like "*metadata*" and "*catalogues*", especially when these were "*standardised*". The EIDC catalogue was specifically mentioned, with the importance of the relationship between "*reputation and discoverability*" being highlighted due to the prominence of the catalogue in Google searches (i.e., at the top); this suggests that UKCEH has an important role in promoting and managing the discoverability of their datasets because of their role in supporting the research community in environmental science.

This report, however, is not just interested in FAIR principles, but in discoverability. In what follows we reveal the major themes that emerged for discoverability and DataLabs in the interviews. As discoverability is contextual, in that it fits within a broader chain of action that makes up research (i.e., before you can use something, you need to be able to discover it), we will focus on uses of DataLabs as a lens to understand the barriers for its future discoverability and the discoverability of assets within it.

Discoverability of DataLabs (Platforms, and Assets)

The interview data revealed the increasing importance of DataLabs for some in their ways of working. Participants spoke highly of its potentials and advantages, for example:

“In DataLabs, the data is all supposed to be there for you [...] the idea, I think is that someone else has done that step for you and they put the data files in the DataLab space and then a new user can just log in and all of that's there. So that's one big saving of time. And then it's supposed to be easier to run. They're trying to make it so that the models can be run through these things called Jupyter scripts.” (P14)

Currently, however, the broader availability of DRI means that there are many **different systems that can be used to achieve similar results**. Where teams, or individual researchers, found barriers in their use of DataLabs, or tensions that they struggled to overcome, it was common to switch over to a different system – or to split their use across different systems.

“I don't usually use [DataLabs] because I already have the data on Polar [high performance computing (HPC) service] and I have scripts doing exactly that. So you just use that because it's just makes it easier, and I can just give the script a couple of locations I want, and I don't need to put it in manually into a lot of clicking.” (P1)

The wide variety of tools available for the task at hand meant that ease of use could influence which system would be used. Participants highlighted some of the areas that they felt created **tensions in their use of DataLabs**. This was not always about the design of the tool itself, but about how users might best leverage its potential for the science they were conducting. P19, for instance, describes the challenges they had in knowing when to set up a new notebook, and when not, as well as how versioning GitHub worked in DataLabs.

“I think once it's sort of explained to you it is quite easy to do in DataLabs. It's just that it wasn't obvious, you know how to sort of access it, but to me without doing a stepping me through it ‘Ohh you go to this bit’ and you know yeah. So it's probably, yeah, protocols and best practise really.” (P19)

Whilst participants discussed the **need for better training and protocols for use**, certain elements of the system that could be automated currently **require the help of others** – with little documentation existing.

“[E]specially if it's not something that you've used before or it can be a bit, I'd probably be a bit intimidated by having to ask someone to set up a project for me. [...] I think it wasn't very clear, that the documentation wasn't clear [...] how you did things, who did you e-mail to get this product set up and that sort of thing. I don't think that was there [...] it might be there now, but it wasn't there.” (P27)

Where this support was not easily accessible, or took a certain amount of time, impacts could be felt in the work itself, or for the rest of the team.

“[There's not] much support for it. It's the fact that people are relying on their DataLab to produce a report or an analysis or report that at the end of the day, so you know if [that isn't] working then there is a knock on for a lot of people.” (P3)

Some researchers highlighted that alongside computational infrastructure more generally, **DataLabs was not yet as discoverable as it could be** within the organisation, posing questions for its

discoverability beyond UKCEH to the wider NERC community. P13 explains their perceptions of DataLabs awareness and findability through the lack of information about the platform on UKCEH's internal knowledge 'Hub':

"And I was thinking [...] there's no information about it on like the Hub or anything about how to get onto it but then there's also nothing about how to get onto JASMIN either." (P13)

While the Hub does not include materials to make users aware of the system itself, nor of how to use it, this may be somewhat intentional due to the **lack of resource capacity to support a wider community in using the platform**:

"[N]obody really knows about it. And it seems like there would [be a] capacity issue with everybody wanted to use it, so um, I guess [... UKCEH needs to] make sure that we've got enough resource for before [DataLabs is] really freely available to everyone." (P13)

There is a tension between the potential benefits of DataLabs in research processes and the organisational capacity to maintain and support widespread use of the tool. This highlights the fact that **infrastructure platforms, like DataLabs, are more than simply tools and technologies**. Their use is underpinned and supported directly by support systems, funding, documentation, training and resources; and this relationship goes both ways. Technological development of DataLabs must be accompanied by developments in these other elements too, to ensure the effectiveness of the DataLabs ecosystem. This was indicated by P7, who referred to the 'hidden' costs of infrastructure to be operationally maintained:

"How do projects run fund things because I feel that so often [UKCEH is] hosting things almost for free or the cost is hidden from the projects and they expect things to be running in perpetuity, so [...] support of these things gets really difficult as they get older and older and you know, you would like the project to be able to fund some cloud architecture that you could just deploy things too." (P7)

Whilst the financial and environmental cost of digital tools and infrastructures are often hidden from their users, organisationally there is a **cost to hosting and maintaining infrastructures** which enhance the discoverability of assets. A balance must be struck between the benefits of discoverable assets with the burden of upscaling, supporting, and financing discoverability.

Cultural Barriers

Other barriers to discoverability were discussed which are less easily attended to through the design of DataLabs, or through the potentiality of LLMs. Participants noted that the **current research culture, and ways of working, limit the discoverability of digital assets** in environmental science; especially in the desire of researchers to contribute towards improving this. Discoverability, in short, requires not just technological tools, but also requires resourcing and motivation on the part of researchers to make their research outputs – and the digital assets encompassing these outputs – discoverable.

"It would just be a little bit more open [...] there are people who are used to working a certain way, they've written a model. It's their model. They don't want necessarily it to be Open Access to everybody to see. And I can I understand a bit of that. That's fair." (P21)

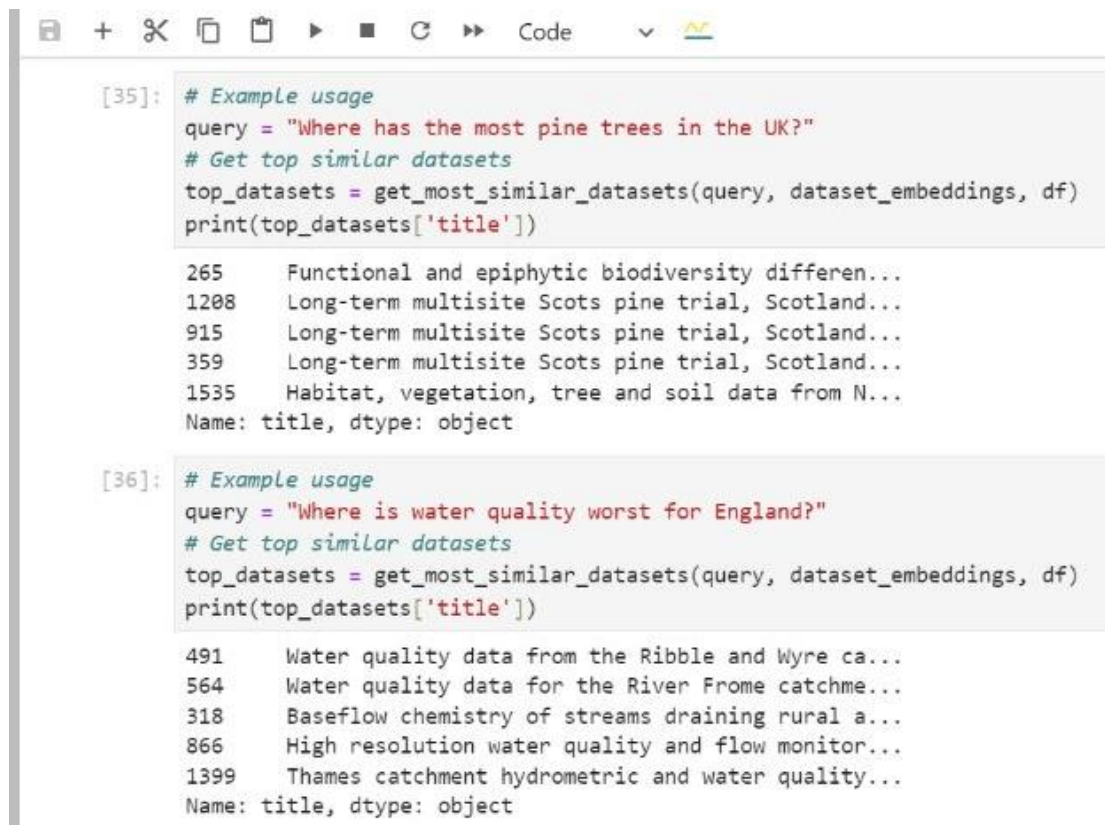
Despite the potentialities of collaboration and discoverability, the ‘publish or perish’ paradigm means that **researchers (especially early career researchers) do not always necessarily want to make their work more discoverable**, particularly where findings were preliminary or partial, or the work was publishable-yet-incomplete. Whilst tools like DataLabs could contribute to challenging some of the competitive aspects of scientific culture, it is likely that **organisational and broader cultural policies will have to accompany technological solutions to create truly collaborative spaces for environmental research**.

Despite the focus of this report on discoverability and DataLabs, some participants did suggest that not everything needs, or should, be discoverable. Some DataLabs were so specific that they were unlikely to be of interest to larger audiences (e.g., P3 said they, at times, were developing DataLabs which would only be of interest to a small group), or the data that they depended on required a level of security or privacy that meant that they could not be shared (e.g., P13 dealt with datasets which included locations making them sensitive). Whilst DataLabs aims to make scientific work more discoverable, it **must also allow for the privacy and security of assets – and data in particular – that require it**. Tools of this type need to have areas that are ringfenced, and users need to be made aware of how to keep their data secure and non-discoverable in new working environments.

Exploration of LLMs for enhancing asset discoverability

Alongside our qualitative understandings of DataLabs and discoverability, we also explored a prototype Large Language Model (LLM) as a promising and emerging technological solution that could improve the discoverability of assets (see Figure 3). A LLM is a type of artificial intelligence model trained on a vast corpus of data, which allows it to understand and generate text that closely mirrors human language. This capability is not only revolutionizing our interaction with technology but also transforming how we access information. When applied to the field of asset discoverability, LLMs can interpret and facilitate the identification of relevant metadata assets for given user queries. They can help facilitate an intuitive environment for end-users, significantly reducing the need for specialized scientific knowledge during the search for a digital asset.

We performed an initial scoping exercise into the application of an LLM for dataset discovery within the EIDC, which could later be incorporated into the DataLabs platform. We developed a semantic-based LLM prototype which allows users to pose broad scientific questions within the search, such as *“is the river Ribble clean?”*. In response, the LLM sifts through the catalogue and returns datasets that are most semantically similar to the query, providing a user-friendly and intuitive approach to data discovery.



```
[35]: # Example usage
query = "Where has the most pine trees in the UK?"
# Get top similar datasets
top_datasets = get_most_similar_datasets(query, dataset_embeddings, df)
print(top_datasets['title'])

265    Functional and epiphytic biodiversity differen...
1208   Long-term multisite Scots pine trial, Scotland...
915    Long-term multisite Scots pine trial, Scotland...
359    Long-term multisite Scots pine trial, Scotland...
1535   Habitat, vegetation, tree and soil data from N...
Name: title, dtype: object

[36]: # Example usage
query = "Where is water quality worst for England?"
# Get top similar datasets
top_datasets = get_most_similar_datasets(query, dataset_embeddings, df)
print(top_datasets['title'])

491    Water quality data from the Ribble and Wyre ca...
564    Water quality data for the River Frome catchme...
318    Baseflow chemistry of streams draining rural a...
866    High resolution water quality and flow monitor...
1399   Thames catchment hydrometric and water quality...
Name: title, dtype: object
```

Figure 3 – A screenshot of the code used to develop the LLM.

Our initial exploration into the use of LLMs to enhance data discoverability suggests they do have the potential to significantly improve the discoverability of digital assets. The LLM-based search approach proved successful in returning relevant datasets for certain search queries, such as “*which regions of the UK have the driest soil?*” and “*are there any data on heavy metal pollution in the UK?*”. Even with minimal metadata, the LLM was able to return one reasonable answer for almost all of the questions tested. This highlights the power of using a small LLM and a single component of metadata to help users discover data for answering broad or targeted questions. It also hints at the future potential for using LLMs at different levels of abstraction, for example, to discover more about individual datasets and as a result enhance their reusability. Fundamentally, rich and high-quality metadata will be an essential part of realising this potential in a sustainable way.

The development of the LLM-based search approach was **subject to several constraints, primarily due to the computing resources available in a standard Lab**. For instance, our current approach only considered the title of the datasets, rather than all available metadata (note: existing EIDC metadata structures are currently under review, to identify and address potential gaps; this includes the experimental development of a supporting documentation template). We also limited the tokenization of the dataset titles to 250 characters and utilized a relatively small pre-trained LLM.

Future work will involve further development of an LLM-based search approach, leveraging more computing resources. This will enable a more comprehensive study into the benefits and drawbacks of this approach, versus the existing metadata-based search approach within the catalogue. Additionally, we will need to consider the environmental implications of employing an LLM-based search method, striking a balance between potentially improved discoverability of assets and environmental sustainability – explicitly asking whether the LLM’s functionality is ‘worth’ its environmental cost. Given that LLMs have a large carbon footprint, it is crucial to question whether

the leveraging of increasing computing resources will give the intended improvements to discoverability to justify its costs.

Discussion: A preliminary typology of discoverability barriers

Drawing upon insights from both studies, we discerned four ‘levels’ of discoverability relevant to DataLabs (see Figure 4). The preliminary typology below articulates these levels and offers **twelve key barriers** to discoverability (three at each level) that are suggested by the findings. In this discussion, we reflect on these barriers, and include additional observations based on the current state of the platform based on our expertise in Human-Computer Interaction. The first two levels – (i) and (ii), in blue on the left side of Figure 4 – refer to barriers within DataLabs itself. The latter two levels – (iii) and (iv), in black on the right side of Figure 4 – refer to barriers in the wider contexts of DRI and scientific research more broadly, which have a trickle-down impact on discoverability through DataLabs.

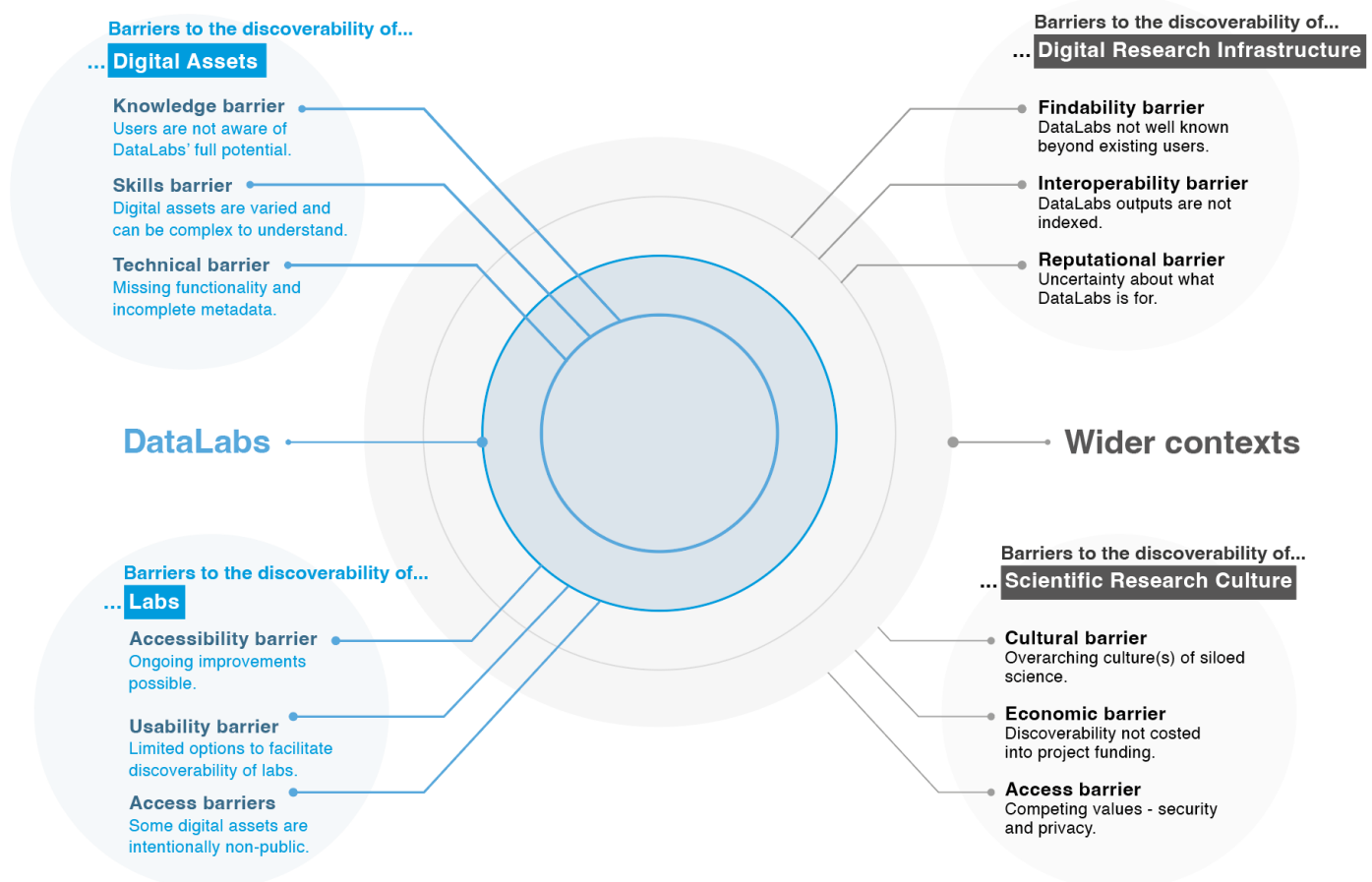


Figure 4 – A typology of barriers to discoverability, framed around four levels - (i) Digital Assets, (ii) Labs, (iii) Digital Research Infrastructure, and (iv) Scientific Research Culture. (i) and (ii) are elements within the *DataLabs* environment, whereas (iii) and (iv) related to the wider contexts of *DRI* and *Science* (in the broadest sense).

(i) Barriers to the Discoverability of Digital Assets in DataLabs

DataLabs is a versatile platform that supports the integration of a wide variety of digital assets. However, our findings reveal that many users are unaware of the full breadth of its capabilities, and the range of digital assets it can support. This **knowledge barrier** limits discoverability by reducing the scope and potential of the platform from a user’s perspective. For example, where users are unaware that they can query a dataset in a certain way, they may be unaware of the potential relevance of that

dataset to their research. This barrier can be overcome through clearer messaging and signposting that communicates the full range of functionality to users. It can also be tackled by clarifying and expanding the documentation to include examples of common user journeys, or example projects, to highlight key functionalities. To maximise the impact, this should be presented in ways that can be easily accessed, shared, and understood by both experienced and less experienced users, and ideally tailored to individual users based on their level of experience and existing knowledge.

There are also **skills barriers** that limit the discoverability of digital assets in DataLabs, particularly for new and less experienced users. Many datasets, for instance, are in formats that require specific scientific knowledge and/or technical literacies to access or analyse. These barriers could be overcome through contextual support. For instance, where users are uploading or browsing files of a particular type, the system could draw upon embedded vocabularies to provide relevant information about that file type, and links to examples of code that integrates those files in ways that are tailored to the user or context. Identifying, developing, and integrating appropriate vocabularies would facilitate this functionality, perhaps in ways that could be leveraged by LLMs. As an interim measure, static contextual support could be implemented to cover common processes, such as setting up a new notebook, or importing datasets from external sources.

While it is increasingly common for scientists to have some technical skills, and for technical professionals to have some scientific knowledge, using DataLabs effectively requires *both* technical expertise *and* scientific knowledge. Our findings suggest that more support is required for users to ensure they can overcome this skill barrier. This could be achieved through training, or through the provision of external reference materials (e.g. relevant YouTube videos).

Another solution could be to provide high-level perspectives on technical and scientific information. This could take the form of data visualisations, or annotations drawing on available metadata. For example, if a data asset refers to water quality in a specific location over a certain period of time (e.g. 10 years), a clearly labelled visualisation could provide a more accessible representation of the relevant insights than raw data. This could improve discoverability for expert and non-expert users alike, making data easier to parse without the need to open (potentially complex) datasets or code.

Technical barriers to the discoverability of assets also exist in the form of missing functionality to discover non-data assets (e.g. models and methods), as well as incomplete metadata. One solution to this would be clearer labelling, potentially using ‘meta-metadata’ (i.e. information about the state of metadata) that advises users when metadata are incomplete or missing. Our findings reveal that metadata shortcomings are commonly caused by limited financial and staff resources, so to address the cause of this problem there is a need for better governance and data management, not just ‘*better technologies*’. Building in extra capacity within projects to ensure discoverability ‘compliance’ (and defining what that means) could help to ensure that data and metadata are complete and of a high standard. However, there is a risk that increasing the administrative burden could lead to further barriers (e.g. compliance barriers). There is also a potential operational risk that normalising incomplete metadata could validate ‘incompleteness’ – especially if technologies such as LLMs can be shown to function well with minimal metadata (as in our experiment). Further work is needed to explore this challenge, to identify and weigh up the risks, and to propose a balanced way forward.

(ii) Barriers to the Discoverability of Labs within DataLabs

Our findings revealed difficulties in finding and interacting with other Labs in DataLabs. At the time of writing, there are around 170 Labs in DataLabs (although this number is likely to grow). By default, Labs are presented in a simple list, ordered by title and accompanied by a short description. The title is often an acronym, and the description is usually just a few words, so these categories rarely provide

users with much insight into the nature or content of the Labs. Furthermore, users cannot access other users' Labs by default. To request access, they must contact the (unlisted) Lab owner (outside of DataLabs). Development work is currently underway to incorporate access requests, but given the paucity of information available, there is little to motivate access requests. A lack of information about the content and context of other users' Labs is a key barrier to discoverability.

Several **accessibility** and **usability barriers** at the interface level combine to form this barrier. In some cases, development work is already underway to resolve them. Others (like the examples above) can be actioned directly in response to this report. Others are not yet known, and can only be integrated into future agile development sprints once they have been identified. This raises the question: *How are accessibility and usability barriers identified and integrated into the agile development process?* The current product owners host *ad hoc* 'community conversations' to collect feedback directly from users and help troubleshoot specific issues. These events provide a valuable communication channel with users, but a clearer strategy is needed to integrate input from users and expertise from the design and development teams in a way that feeds into the agile process.

The 'semantic similarity' functionality of LLMs could also be used to improve the discoverability of Labs. Implementing the LLM prototype into DataLabs would enable users to find relevant Labs without needing to know which specific terms are used in Lab titles/descriptions. However, this functionality needs to be carefully balanced privacy and security needs and sustainability considerations due to the high energy footprint of LLMs. Future work is needed to understand the cost-balance ratios of integrating LLMs.

An important consideration is that many of the **access barriers** to DataLabs are *intentional*. For instance, to register for DataLabs users need to have a relevant project that is funded by NERC. This limitation ensures DataLabs can scale at the right pace. Although being more open could lead to better discoverability, it could also create unsustainable demand for the service in how it is currently funded. There is also a need to balance discoverability and competing values such as privacy and security. User restrictions prevent unauthorised access to assets, such as those that may be sensitive, or privileged, but this raises practical and ethical questions such as *who grants access*, and *how is access managed?* It also presents additional **technical barriers**. Presently, APIs and apps can be used to make certain data and assets visible, but these outputs are not indexed or searchable – either within DataLabs or through other infrastructures (e.g. via DOIs or similar unique identifiers). Future work should consider ways to index DataLabs' outputs (e.g. APIs and web apps) to improve their discoverability, whilst also supporting controlled discoverability of limited-access areas *within* DataLabs to ensure that users can harness the potentials of the tool whilst protecting the assets within it.

(iii) Barriers to Discoverability in Digital Research Infrastructure (DRI)

DataLabs is only one of a growing number of digital tools, services, and platforms providing digital infrastructure for scientific research. For it to be useful, users first need to know that it exists and be able to find it; this is a marketing challenge. DataLabs is competing in a crowded online space with powerful tech companies (e.g. Google) with attention-based business models, and there are other VREs (e.g. CoCalc, Code Ocean, Ocean DataLab) offering similar functionalities. This is a discoverability barrier built atop of a **findability barrier**. Participants noted that there is minimal information about the platform shared internally at UKCEH – this might be easily addressed through clearer communications and messaging, however further research is needed to explore how potential users (especially those external to UKCEH) find and engage with DRI (broadly) and DataLabs (specifically). Simple improvements, such as search engine optimisation (SEO)¹ will help improve the visibility of

¹ The topic indexed Google search for 'datalabs' links to ukmaps-ecomapsdemo.datalabs.ceh.ac.uk

DataLabs, but it is also important to gain a better understanding of users' 'user journeys' from 'first contact' to regular usage – to maximise the user experience.

To facilitate this aim, it is important that users – and potential users – are supported to understand how DataLabs fits within the overall landscape of digital tools and DRI. For example, tools like *Google Scholar*, *ArXiv*, *Web of Science* also facilitate discoverability of research outputs, but they focus on indexed research outputs with DOIs. Here, an **interoperability barrier** arises since DataLabs' outputs (e.g., web apps and APIs) do not have DOIs. Conversely, tools like *Google Docs*, *GitHub*, etc. which do support collaboration, do not offer discovery as a key feature, although they do have various other features – and limitations. Users need to know what DataLabs is designed for, and how its features compare to other systems; knowledge that can take time to build up. This manifests as a **reputational barrier**, where users reported feeling confused about what DataLabs is *for* and *how it works*. To address this risk, key features should be communicated more clearly. The DataLabs home page could be re-designed to include clearer messaging about how the system works, and improvements made to the readability of the content, which is currently written in quite specialised language. It would also be useful to address the confusion between 'DataLabs' (the platform) and 'DataLabs' (the project files, which we have been referring to as 'Labs'), perhaps by re-naming the latter. More work is needed to address these barriers in the wider context of DRI.

(iv)- Barriers to Discoverability in Scientific Research Culture

Another barrier is the complexity of the field, which presents an **overarching barrier** to discoverability. Environmental science encompasses many overlapping disciplines, communities, institutions, and a complex, dynamic web of connections to other domains (e.g., biological sciences, management sciences, etc) and wider society (e.g., funding bodies, government, media). It can be challenging for individuals – even experienced professionals – to navigate the field, and to discover knowledge gaps within it, particularly when knowledge is fragmented across many digital platforms. Many of the discoverability barriers arise from this complexity, and some can only be addressed at this complex, socio-cultural, political, and economic level.

For instance, **economic barriers** to supporting and maintaining discoverability (e.g., open access fees), are not always adequately accounted for in research funding models but could be overcome by costing-in time and expertise to individual projects to ensure outputs are discoverable. This relates to another **access barrier** that arises due to scientists currently being fearful of sharing datasets and assets before they have had the chance to disseminate their own findings from them. To overcome these fears, more flexible citation and attribution systems with licencing protections could be explored to make scientists feel more comfortable about sharing data and other assets at earlier stages of their research.

Lower-level barriers to discoverability of digital assets (i) and DataLabs (ii) can be addressed through the design and development of DataLabs; both technical development of the platform, and socio-technical development of the wider infrastructures that support it. However, it is important to acknowledge the inter-relationships between these barriers and the higher-level barriers of DRI (iii) and research culture (iv). To overcome the higher-level barriers (and the systemic issues they represent), strategic interventions may be required at institution and policy levels. It is beyond the immediate remit of this preliminary study to propose specific changes at this level. However, by developing DataLabs to be an exemplar of discoverability 'best practice', it can contribute to the gradual process of prioritising discoverability. In this respect, the barriers become opportunities; to create immediate impact in the form of better infrastructure for integrative science, but also for longer-term impact.

Conclusion and future work

In this report, we have uncovered opportunities for the enhancement of discoverability of DataLabs as well as the discoverability of assets within the platform. We have drawn on users' experiences of DataLabs and experimented with the use of LLMs as a technological solution to the discoverability of assets. We have described a typology of barriers to discoverability, both within the DataLabs platform and beyond.

To address these barriers, there is a need for more research. Specifically, we see three key areas of work to explore moving forward:

1. Improving the **User Experience (UX)** of the DataLabs platform through user-centred design;
2. Enhancing our **co-design** approach to address different levels of barriers to DataLabs and discoverability;
3. **Mapping the stakeholder landscape** to ensure all stakeholders are considered in our approach.

We explore these each in turn below.

User Experience (UX) Improvements

Our findings point to some initial steps that could improve discoverability *within* DataLabs through UX improvements. These include, for example, enhanced contextual support and documentation in the platform alongside more visual representations of assets that are available to explore and use within Labs. Barriers affecting the discoverability of Labs within DataLabs could also be overcome through the automation of certain interactions, such as setting up a Lab, and developing features that allow users to discover assets *beyond* data, such as other Labs, models, and methods. There is a need for a renewed focus on user-centred design for DataLabs, and for this, we suggest an industry standard usability evaluation should be conducted in dialogue with DataLabs' product owners and the development team. By drawing on established heuristics (e.g., Nielsen, 1994), UX enhancements can be easily adopted in the platform and offer immediate usability improvements through intuitiveness, consistency and intelligibility. We also suggest that this should be done in a way which maps across the DataLabs system and its underlying infrastructure, ensuring a dynamic blueprint of the DataLabs platform can be established. This will ensure that future iterations to DataLabs' UX will be easier to update, with additional user journeys identified in future co-design activities with stakeholders.

Co-Design Approach

As previously noted, an agile, co-design methodology has been adopted within the Environmental Data Science team at UKCEH to support enhanced engagement with stakeholders within the design and development of DRI – delivering value quickly and iteratively to users of DRI. However, improvements to the methodology are required to truly integrate stakeholders needs within the design and development of DataLabs and to overcome the variety of barriers highlighted by this report – which extend from within DataLabs itself, to the research culture that surrounds it. We envision three levels of stakeholder engagement in our co-design approach that are needed to address these barriers (see Figure 5).

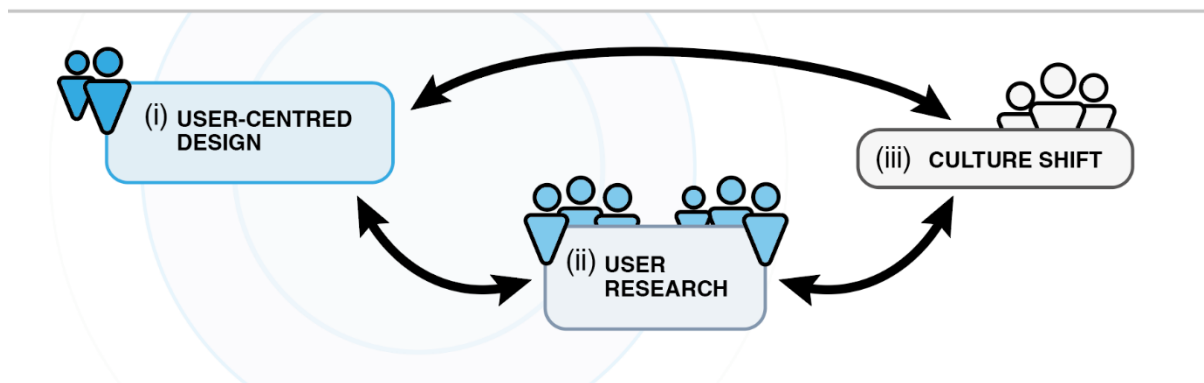


Figure 5 – A diagram showing how the three aspects of the co-design approach inform one another; (i) user-centred agile development, (ii) user studies, and (iii) culture shift.

(i) User-Centred Design

Firstly, we need to bring user-centred design into the agile software development process, showcasing specific DataLabs features to the most relevant stakeholders and engaging them in the design of user interface (UI) elements. For instance, testing with users will help bring UX improvements (such as those noted above) continuously into the development process, and support product owners in answering specific questions. For example, questions such as ‘*does visualisation A or visualisation B work best for discoverability of dataset Z?*’ could be tested with users using A/B testing, with the results of these tests feeding directly into development sprints.

(ii) User Research

Secondly, we need to continue conducting focused user studies with stakeholders to drive forward new possibilities for DataLabs and discoverability that will inform and support larger pieces of software development. This will draw upon established HCI/design methods such as workshops and probes. These will help us to answer broader questions, such as:

What scientific questions do stakeholders want to answer that need discoverable and integrative assets and how can DataLabs support this?

How should different assets (models, methods, Labs) be made discoverable in DataLabs, and what privacy and security mechanisms are required for these?

What are the environmental costs of LLMs, and do LLMs add enough value to users and environmental science beyond traditional metadata searches to make these environmental costs worthwhile?

(iii) Culture Shift

Finally, we need even higher-level ways of engaging with stakeholders that can, over time, support efforts to overcome the cultural barriers to DataLabs and discoverability that currently exist. As it stands, improving the discoverability of DataLabs implies a further burden on individual researchers and teams to make their work more discoverable to others – with perhaps no clear benefits being obvious to those individuals. Moreover, there are economic barriers that prevent long term support for discoverability across science projects, requiring funds and resources to upscale the support systems and infrastructural capacity of DataLabs and asset discoverability. Through new policies and governance of DRI, as well as enhanced training around the importance of discoverability, it may be possible to shift this culture to one which is more community focused and collaborative – and, specifically, a culture which truly supports the integrative and systemic environmental science needed

today. Changing culture will take time, but we suggest community impact and stakeholder engagement activities should be conducted to support this. These may include, for example, exploring the use of different media and storytelling techniques to get across why sharing assets and their discoverability is important, or case studies of how discoverability has supported impactful, systemic science to encourage other stakeholders to engage.

Stakeholder Landscape

Enhancing discoverability of DataLabs and the assets within it requires a better understanding of its users – and potential users. Our findings show DataLabs can be difficult to find, and it is relatively unknown in the community, suggesting it is not reaching the stakeholders for whom it is created. Additionally, there will be further stakeholders across UKCEH and in the wider community that are currently not being considered at all. Given the focus of DataLabs is to support collaborative, integrative and systemic science, there will be a wide variety of stakeholders for whom the platform could be useful; we need to engage with these stakeholders more deeply to map their requirements and the science questions that they may wish to answer through enhanced discoverability of assets (via an LLM or traditional metadata searches). These will include scientists, academics, and decision makers such as policymakers (e.g., *“what policy should be introduced to control river pollution?”*) as well as individual members of the general public (e.g., *“is it safe to swim in the River Ribble?”*). We need to better understand the stakeholder landscape in environmental science and uncover which tensions and barriers they face in their work that could be supported through DataLabs or other DRI, and in turn, how we may best engage them through our suggested approach of stakeholder engagement and co-design. Future work should map the landscape of stakeholders across environmental science and its connecting domains, acknowledging that this is a dynamic category; as the science changes, so too do the stakeholders, and the potential application areas for the kinds of science DataLabs supports.

Summary

The enhanced, future version of DataLabs will connect siloed scientific knowledge by supporting inclusive collaboration across disciplines and uniting heterogeneous digital assets and services via an intuitive, accessible interface. It will be easy to find and access, and clearly positioned in relation to other platforms. It will balance privacy and openness, adapt to users’ changing requirements and support new and experienced users alike with relevant, well-structured documentation and messaging. It will build on the strengths of the existing system and, by overcoming the barriers presented in this report, DataLabs will significantly enhance the discoverability of environmental science whilst simultaneously acting as a key case study for addressing the wider, cultural aspects affecting the discoverability of DRI and scientific knowledge. These opportunities are summarised in Figure 6.

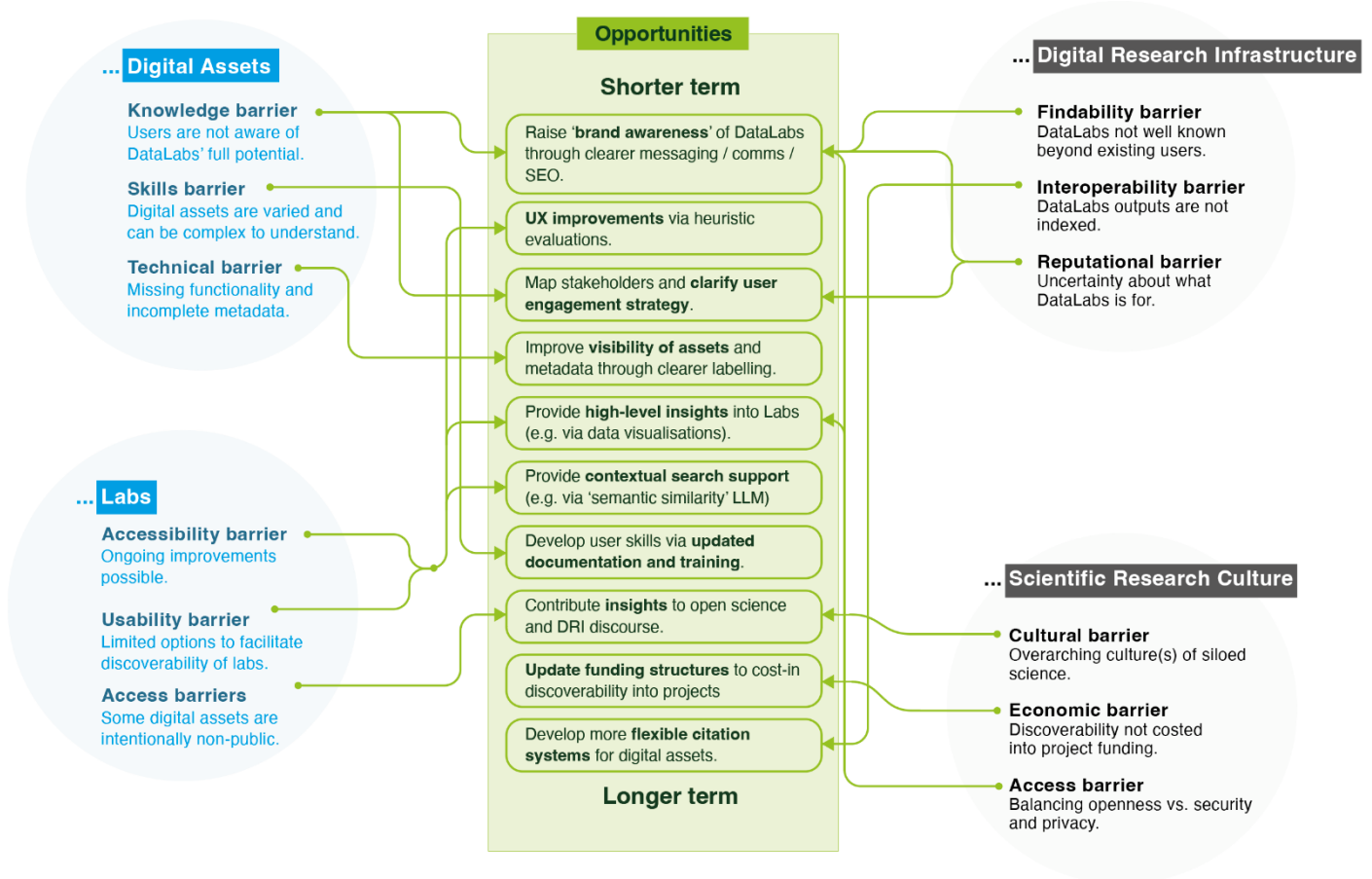


Figure 6 - A breakdown of the actionable opportunities presented in this report, and how they relate to the discoverability barriers (originally shown in Figure 4).

ACKNOWLEDGEMENTS

This work is funded under a NERC grant 'Enhancing DataLabs in Response to Changing User Needs and New Opportunities' as part of the 2024 opportunity 'UKCEH DRI Programme Funding Expectations'. We also thank our participants within the DRI Futures study (survey, interviews, and workshops) for their time, insights and vision for the future of DRI.

REFERENCES

1. Bingham, A.J., & Witkowsky, P. (2022). Deductive and inductive approaches to qualitative data analysis. In C. Vanover, P. Mihas, & J. Saldaña (Eds.), *Analyzing and interpreting qualitative data: After the interview* (pp. 133-146). SAGE Publications.
2. Bird, C., Priest, C., Lord, C., Friday, A., Widdicks, K., Kayumbi, G., Lambert, S., Jackson, A. (2023). *Learning from the Big Picture: Applying Responsible Innovation to the Net Zero Research Infrastructure Transformation* (ARINZRIT). <https://zenodo.org/records/7966424/files/ARINZRIT-Final%20Report-23May2023.pdf?download=1>, accessed April 2024.
3. Interaction Design Foundation (IxDF) (2024). What is Discoverability? <https://www.interaction-design.org/literature/topics/discoverability>, accessed April 2024.

4. Kraker, P., Schramm, M. and Kittel, C., 2021. Discoverability in (a) Crisis. *ABI Technik*, 41(1), pp.3-12. <https://doi.org/10.1515/abitech-2021-0003>
5. McKelvey, F. and Hunt, R., 2019. Discoverability: Toward a definition of content discovery through platforms. *Social Media+ Society*, 5(1). p.2056305118819188. <https://doi.org/10.1177/2056305118819188>
6. Hollaway, M.J., Dean, G., Blair, G.S., Brown, M., Henrys, P.A. and Watkins, J., 2020. Tackling the challenges of 21st-century open science and beyond: A data science lab approach. *Patterns*, 1(7). <https://doi.org/10.1016/j.patter.2020.100103>.
7. NERC (2022). Digitally Enabled Environmental Science – NERC Digital Strategy 2021-2030. <https://www.ukri.org/wp-content/uploads/2022/05/NERC-170522-NERCDigitalStrategy-FINAL-WEB.pdf>, accessed April 2024.
8. Nielsen, J. (1994). Enhancing the explanatory power of usability heuristics. *Proc. ACM CHI'94 Conf.* (Boston, MA, April 24-28), 152-158
9. Paic, A. (2021), "Open Science - Enabling Discovery in the Digital Age", *Going Digital Toolkit Note*, No. 13, https://goingdigital.oecd.org/data/notes/No13_ToolkitNote_OpenScience.pdf
10. Scrum.org - 'What is a Product Owner' <https://www.scrum.org/resources/what-is-a-product-owner>, accessed April 2024.
11. Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016). <https://doi.org/10.1038/sdata.2016.18>
12. W3C WAI (Web Accessibility Initiative). WAI-ARIA. <https://www.w3.org/WAI/standards-guidelines/aria/>, accessed April 2024.