INDUSTRIAL STRATEGY | UK Research and Innovation

HDRUK Health Data Research UK

DIGITAL INNOVATION HUB PROGRAMME PROSPECTUS APPENDIX:

HEALTH DATA RESEARCH INNOVATION GATEWAY OVERVIEW

May 2019

Contents

Purpose of document
This document is an appendix to the Digital Innovation Hub Programme (DIH) Prospectus and provides a description of the tools and functions that will be developed as part of the UK Health Data Research Innovation Gateway. This document is **not** a Request for Proposals nor intended to be a detailed specification of requirements. More information on the Gateway development process will be given at a future stage.

1. Introduction

The major barriers to the use of data for research and innovation are the inability to quickly access data (mentioned by 70% of respondents across direct industry engagement[1]) and the inability to identify the location of data and understand data quality (mentioned by 55% of respondents).

The UK Health Data Research Innovation Gateway (the 'Gateway') is being established to address this and will act as a common access point to UK health research data for industry, researchers and innovators.

The data, acquired by members of the UK Health Data Research Alliance (the 'Alliance') and Digital Innovation Hubs (the 'Hubs'), will be held in Safe Havens (Trusted Research Environments) to provide a safe location for data storage and access, facilitate interoperability, and provide analytical capability[2]. **Only the metadata (i.e. no personal identifiable information) will be available in the Gateway.**

The Gateway is designed to support the whole user journey, from identifying the data that is available, coordinating an access request across multiple data custodians, requesting a linkage of multiple datasets and facilitating access to these through Safe Havens.

The Gateway will be available to industry users from all sectors, as well as to researchers and innovators across the NHS and academia.

[1] Industry Engagement conducted as part of the DIH Design and Dialogue phase, which involved 32 in-depth interviews with industry representatives

[2] As part of the UK Health Data Research Alliance, a workstream has been established to define the functions and best practice of Safe Havens across the UK, with a view to agreeing on a standard/accreditation process.

The Gateway will build on tools and best practices that already exist in the UK and make use of existing assets and previous investments, such as the NIHR Health Data Finder and associated Metadata Catalogue tool.

The Gateway will provide the following specific functions for users:

- Discover – identify data through a common point of access to provide a view of available datasets
- Understand – provide information on the data quality, terms of use and timescales for accessing the available data
- Access – coordinate data requests to individual data controllers, providing a harmonised way for uses to request access to data
- Link – support the ability to request linkage of datasets within Safe Havens, in line with information governance requirements, to provide enhanced datasets for research and innovation
- Share / Learn (repository) – collect information from the user community on descriptions and definitions (phenotype library), analysis codes, annotations and methodologies to support further research.

The underlying data does not flow through the Gateway nor will there be storage or computing services within the Gateway. The data will continue to be held by data custodians in Safe Havens. The Gateway provides the above functions by linking to the Safe Havens.

The Gateway will be available to access by Hubs, as well as directly by users from across the UK. While the focus of the Gateway is to support research and development across industry, academia and the NHS, HDR UK will work in partnership with NHSX to align this work with mainstream NHS endeavours, including the development of clear standards for the use of technology in the NHS.

2. Benefits

The Gateway will greatly improve the experience for patients and users.

Patients: Patients will experience more targeted treatments and interventions as a result of data linked across previously unconnected and inaccessible datasets, supported by research and innovation made possible on a much larger scale than previously. Patients and the public will also have greater visibility into how data is being used for research and innovation.

Industry: The Gateway will provide a common shop window for organisations wanting to access the wide range of high quality healthcare data across the UK. It will be easier to identify and faster and more cost-effective to discover and access the information available.

Academic researchers: The ease of finding the right datasets and streamlined access to datasets, will save researchers much needed time and effort allowing them to focus on information value rather than the pursuit of data.

Data custodians: Data owners will be able to manage data access requests using a single workflow and the number of clarification cycles will be reduced by providing consistent, structured and system validated requests. They will be able to see how their data is being used, how often and for what purposes.

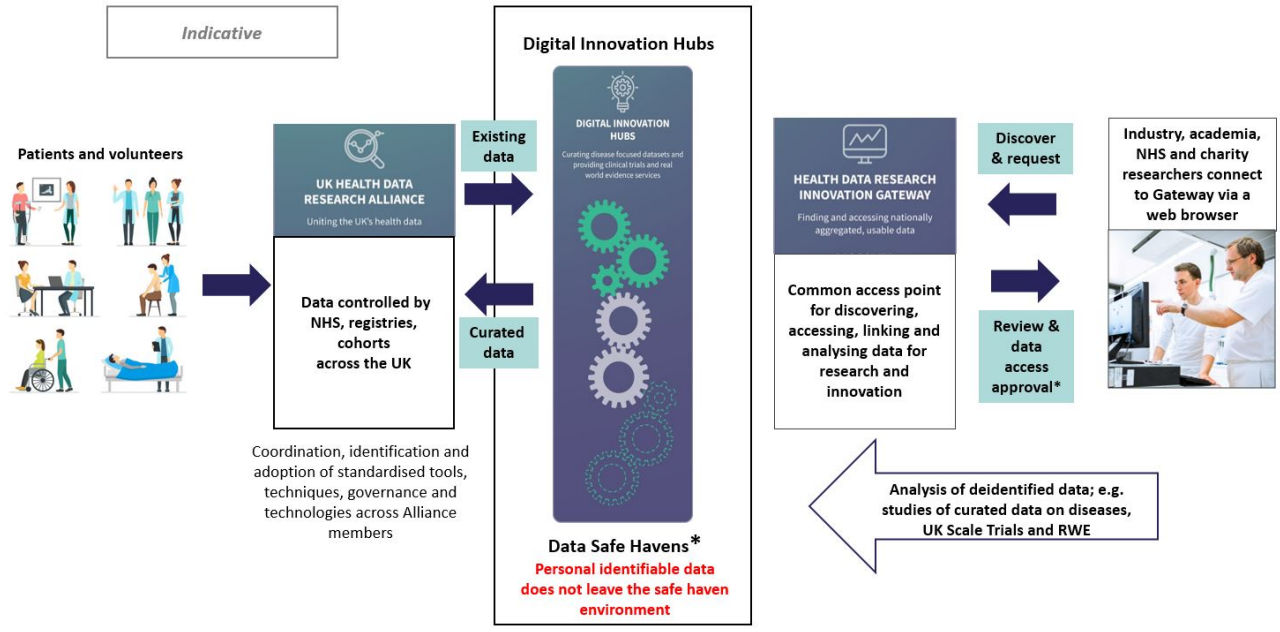
Research and healthcare funders: Improved systems for discovering and accessing data will improve the efficiency of research funding, allowing for more targeted research which achieves a greater impact.

3. Approach

The Gateway will build on tools and best practices that already exist in the UK, including open source technologies. Where new solutions are needed, existing 'best of breed' solutions will be sought from the market and custom development will only be required where a suitable product or application does not already exist. Harnessing capability across the UK, the majority of the elements of the Gateway will be developed by industry experts in partnership with NHS and academia. Once developed, HDR UK will work with partners to evaluate the appropriate long-term approach to running a sustainable, reliable and scalable service.

The Gateway will not take on the role of data storage. Data security is paramount, and data will remain held by existing data custodians in secure environments. It will operate under principles of federated access, with existing data custodians retaining robust governance procedures to ensure that any access to data remains appropriate, proportionate and necessary. The Gateway will provide the ability to operate at scale, as further health data research activity takes place. As the Gateway stimulates further activity in industry, the reliability of this will allow for significant increase in industrial capability.

The Gateway will be developed in phases through a process which is aligned with, but distinct from, the competition for hubs, with Phase 1 commencing summer 2019. It will be developed in a modular fashion, with all components able to operate independently and integrate through fully documented open APIs. Components of the Gateway will be developed in the open, using open-source principles and with focus on providing a services that incorporates best practice in security, reliability, availability and scalability such that it can meet the business critical requirements of industry as well as researchers from academia and the NHS.



*All Hubs must use a Data Safe Haven - a secure place, controlled by Alliance data controllers, used to store research data, with access governed by the "Five Safes" = Safe projects (GDPR compliant and approved to be in the public interest); Safe people (accredited users); Safe data (de-identified); Safe places (secure analytical environments with no data travel); Safe outputs (statistical disclosure control to ensure privacy);

Figure 1: Schematic showing data and information flows for Gateway Minimal Viable Product

Phase 1 (commencing summer 2019)
The initial delivery phase will focus on developing a functional approach (Minimum Viable Product) to:

- Allow users to **discover data** through a common point of access, which provides a view of available datasets, enabled by federated metadata management and search tools to provide improved visibility, comparability and navigation of existing datasets.
- Provide a means of **accessing data** that retains data controllers' requirements whilst harmonising processes, reducing transaction costs and improving access to more isolated, important datasets.
- Embed analytics capability to **assess use** in order to understand how to meet the needs of users and support further access to data.

This will be brought together under an integrated user interface.

Phase 2 (2019-2020)
Following the completion of the first phase, further functionality will be developed to:

- **Assess** and pre-validate data requests based on workflow management to improve response times for decisions, whilst allowing data controllers to stay in control.
- Extend the existing **access** functionality to provide a business process engine, collecting information on requests, turnaround times and use, allowing for seamless and rapid processing of requests
- Link with best in class de-identification and encryption software across the Alliance in **privacy and interoperability,** in line with information governance requirements, to enable secure, novel data linkages across a wide range of datasets to meet researcher needs while ensuring privacy of sensitive health data.
- Harness **interoperability** to support access to linked datasets across the data custodians in the Alliance. The interoperability support will also extend to support mapping between datasets. This will focus on enabling re-use of mapping carried out as part of planned research projects, as well as codifying interoperability issues between different datasets, for example through the calculation of an interoperability score which would be available to other researchers considering working with the same datasets.
- Provide a **library** to act as a single storage area supporting reusable or shareable analytics, linkage maps and phenotypes.
- Support closer integration with **Safe Havens / Trusted Research Environments** to allow for direct provisioning of workspaces.

4. Interaction between Hubs and the Gateway

Hubs will be required to store data within a Safe Haven and to join the Alliance. Hubs will curate and improve data as part of their service, and they will be required to make the underlying metadata searchable through the Gateway to increase use for research and innovation. In the first instance, Hubs will be expected to provide the metadata to be used in the Gateway, the ability to keep this metadata current, and to support the user request access link between the controller and Gateway. By the time Hubs are operational, the Discovery and simple Access functions of the Gateway will be in place. Linkage functionality will be supported by the Gateway and the Alliance and will be developed in 2020.

5. Phase 1 – Functions

This first phase will deliver a robust and functional, Minimum Viable Product with features to satisfy early user requirements and to provide feedback for future product development. This phase of the Gateway will support a user's information gathering journey from the point of posing a question through to requesting access to data.

The Gateway will transform how users from industry and academia interact with health data. It will provide a seamless user experience, developed through a 'Design Thinking' led process to ensure that it meets the needs of all users. The single user interface will be consistent across all the services provided by the Gateway. The tool will support self-service, and will collect information on queries, access and turnaround times. This will allow further refinement of the tool, and will support work across the Alliance to prioritise process improvements, data quality and availability. Feedback through the Gateway will inform curation of data, targeting these services on the areas that would have greatest impact for research and innovation across all sectors.

Phase 1 of the Gateway will provide:

Discovery
The Gateway will provide a searchable database of metadata for users to identify available information from the Alliance. Already, over 270 datasets have been identified across the Alliance that could be made discoverable through the Gateway. This number will continue to grow as the Hubs begin to increase the availability of data, and more organisations join the Alliance.

The metadata catalogue that underpins the Discovery function will support the development of a more integrated data culture across the UK and benefit from knowledge across many sectors, as the community will both use and contribute to the data catalogue.

The Discovery function will support users with varying levels of experience and background knowledge. It will contain open search tools suitable for users less familiar with the datasets to refine their queries, and advanced search functions for experienced users. The approaches taken will be support users ranging across research and development in industry, innovation for medtech and AI as well as academic researchers.

A conceptual example of the discovery user interface is given below in Figure 2, detailing search and discover functionality.

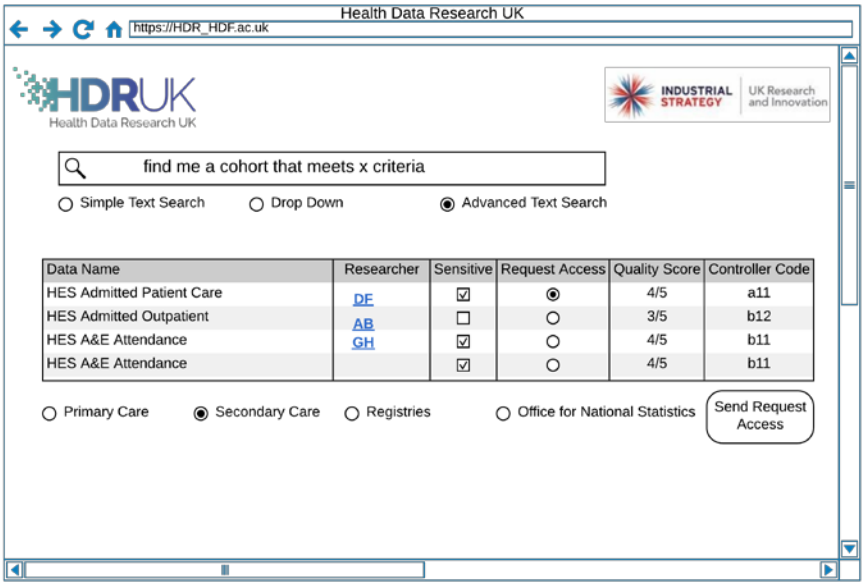


Figure 2: Conceptual Example of User Interface for Discovery

When reviewing the data available to support a user's request, the Gateway will provide information about the dataset, including coverage, terms of access and expected turnaround times.

The technology base for the Phase 1 Discovery function will be through developing and enhancing the existing National Institute for Health Research (NIHR) Health Data Finder and Metadata Catalogue. These will be extended to fully support the capability and user experience required for the user community.

Access

Phase 1 of the Gateway will facilitate the access of datasets from multiple data custodians. Currently, researchers from academia and industry must directly contact many different organisations to access the right dataset.

Access request will incorporate processing requests directly to organisations on behalf of the user. Taking information from the search outputs and the user profile, the tool will coordinate requests to multiple custodians in a seamless manner, requesting additional information as necessary and in the required format. This will significantly reduce the administrative burden on users and data custodians, without reducing the security supporting ethical information governance. Storage of information from previous requests will allow information to be pre-populated, where appropriate, further reducing administrative duplication.

Requests will be monitored through the Gateway, supporting automated chasing and escalation and providing the user with information on the current stage and next steps.

The Access function will be developed further in Phase 2, including creating an underlying business process engine to manage workflow for access requests and a rules engine to enable pre-validation of submission,

this will make things even more straightforward from a users' perspective. It will include an open API that will allow direct integration with the systems of data custodians.

6. Consultation and development

The outline of the Gateway has been developed through specific engagement with a broad user community, representing industry, academia and the NHS. Over the DIH Programme Design and Dialogue phase, over 2,700 people have been engaged to understand their specific needs and challenges. A series of workshops were convened to review the major challenges that could be addressed by the Gateway, including metadata management and interoperability, and to make sure that these are fit for purpose.

The development of Phase 1 will be broken down into a series of modules to deliver the functions listed above. These will be developed in an agile manner, in two-week sprints that will each result in a functional update to the Gateway. This means that users will be able to interact with beta versions and provide feedback on an ongoing basis through the development process. Further information on the development process, specification and where appropriate procurement will be released in subsequent months.

7. Phase 2

Following the completion of the first phase of development, usage will be monitored, and user feedback obtained from across industry, healthcare and academia. This will be used to inform the second phase of development, which will be co-developed with the user community and will likely provide many of the functions outlined in Section 1.

Example use case:
Alison is a researcher working for a UK-based SME which provides lifestyle management interventions for people with heart conditions. She would like to understand the factors which lead up to hospital admission, so that, with earlier support and intervention, people will not have to be admitted to hospital. In order to identify the information that is available to support her request, Alison accesses the data catalogue in the Gateway. The ability of the tool to handle unstructured queries means that Alison can explore a range of information, without knowing in advance what information is available and in which formats. The tool produces a list of datasets that will help her answer her question, as well as information about the suitability, information governance restrictions and time taken to access these. Alison sees that some of the datasets have very high requirements for approval, so she selects two datasets that contain less sensitive information but would still meet her research needs and requests access through the Gateway. She submits information on her query, as well as specific supporting detail on information governance, through a web form. This supporting information is then sent to each of the data custodians through an automated process. As her request progresses, Alison can to track the approval process through the tool and communicate with each of the data custodians through a messaging system. Following approval, she can access the research outputs and run queries against the linked dataset through the tool. The linked dataset itself is now accessible to other researchers through the Gateway, reducing the time required for approval in the future.