**Findability of UK e-cohorts available for health research: mixed methods study Supplementary File 4**

**Views of established research data centres on data discoverability**

The information was generated through telephone interviews with representatives of each data centres

|  |  |
| --- | --- |
| **Health Data Finder**  |  |
| **Overview:** HDF is a metadata catalogue aiming to inform users what health datasets are available for reuse: initial focus on Public Health England, NHS Digital, and CPRD datasets which are familiar to researchers. Interested in including additional datasets that are less well known. Also working with data custodians to try to improve (streamline) the data governance and access process as this is currently inefficient. | **Perceived barriers to discoverability:** Major challenge to solve is the difficulty in access: discoverability is first step in a long process. The potential user networks are in place so people tend to know what is out there. In England, multiple data custodians and all have different data access procedures & approaches to solving governance challenges. Additional challenges have been introduced with GDPR (e.g. opt-out). Process is long and inefficient: aiming to be as transparent as possible about the current process, but would also like to improve it. Will having local hubs potentially make this situation worse? Challenge in making linked datasets easier to access: even when same linkage is performed frequently, each request must be treated independently and the linkage repeated. Also re: metadata, lots of organisations, and use lots of formats to curate their metadata (pdf, excel, word); information not complete and not up to date. The catalogue can therefore be helpful for the data custodians because it is streamlined. A different challenge is finding out what other people are doing in this same space – there is no community, people tend to stumble across each other’s work by accident. |
| **Current discoverability:** the Health Data Finder catalogue: <http://www.hdf.nihr.ac.uk/>  | **Future plans regarding discoverability:** Add functionality, e.g. automated forms/queries for applying for access. Would like to include more datasets (e.g. common linkages) but challenges in this. Besides these, two major aims: to involve and engage with researchers (in a manageable way) to make sure the catalogue is as useful as possible; the ultimate aim is to allow the catalogue to grow into a helpful standard that helps people explore and investigate metadata. |
| **UK Data Service**  |  |
| **Overview:** Enables access to a range of data resources. Data are classified as either open, safeguarded, or controlled. Mainly hold collections of social and economic data but are increasingly including health and clinical data, and data from relevant trials and cohort studies (e.g. brain scans, blood test results). Increasingly receiving data from MRC-funded research, and acting as a data store for data papers (e.g. Scientific Data). | **Perceived barriers to discoverability:** Getting complete metadata from data owners is a challenge as it is not a priority for them. The quality of metadata can be poor, and there can be issues with file naming and adopting controlled vocabularies. The metadata received can need a lot of enhancing or editing, which is time-intensive. There can be a difference between what data owners report in the metadata and how the data were actually generated – this is difficult to verify. In terms of making data available for reuse, data owners tend to be conservative in terms of locking down their data. |
| **Current discoverability:** Datasets are represented in a catalogue: <https://discover.ukdataservice.ac.uk/> Knowledge of the catalogue will depend on the person, for example students may not be aware of the catalogue, but researchers using social or economic data probably will as this is the primary data store for such data. Individual datasets are discoverable through web searches so if a researcher is searching for a particular dataset they will be able to find them. | **Future plans regarding discoverability:** Exploring different software options, e.g. DataVerse |
| **Consumer Data Research Centre (CDRC)**  |  |
| **Overview:** Enables access to routinely-collected consumer data (only a limited amount of health data). Data are classified as either open, safeguarded, or controlled. | **Perceived barriers to discoverability:** It takes a lot of time to get the word out, and self-promotion requires a lot of work. Hoping to capitalise on publications from data users; these are starting to appear after the first wave of data access. The CDRC ask for 10 days’ notice of submission and add case studies to their website. One source of delay was researchers waiting for their funding to begin. |
| **Current discoverability:** The website (<https://www.cdrc.ac.uk/>) was launched in 2015 (use Google Analytics to track engagement). The program has been promoted in the media and via tweets, newsletters, and word of mouth through the social science community. Much of the data is openly available which attracts users. | **Future plans regarding discoverability:** Reviewing the portal for improvements in e.g. searchability and layout. Plans to upgrade profiles. |
| **Urban Big Data Centre (UBDC)** |  |
| **Overview:** Enables access to urban-related data. Data are classified as either open, safeguarded, or controlled. Include arrange of datasets e.g. datasets purchased from Zoopla and Strava. | **Perceived barriers to discoverability:** Staffing, and appreciation of the importance of data curators: originally planned a more organic approach without curation. Development (e.g. web development) is a lengthy process. Challenge to find ways to promote the safeguarded data. |
| **Current discoverability:** Open datasets are represented in a catalogue: <http://ubdc.gla.ac.uk/dataset> and there are plans to include metadata for the safeguarded datasets. The service is promoted at conferences and other events, through social media (including LinkedIn, Facebook and Twitter), and through the website. | **Future plans regarding discoverability:** plan to include metadata for the safeguarded datasets in the catalogue. The website is under development and there are plans to include better integration with the catalogue. Would like to improve the catalogue, e.g. embedding better/more relevant and selective metadata. |
| **Administrative Data Research Network (ADRN)**  |  |
| **Overview:** Acted as a secure access service, without holding a store of data. Originally planned to act as a top-tier access mechanism, putting together datasets and linkages on demand, but this model was difficult to resource. The service moved towards keeping and reusing datasets, and proactively building linked datasets. The ADRN was a network of five projects – the four UK Nations and a UK-wide centre.  | **Perceived barriers to discoverability:** In many cases, metadata is not accessible because it does not exit. This results from the fact the data were not collected for research, so the same need to describe the dataset did not apply. In addition, for the datasets that are fully documented (e.g. some of the NHS Digital datasets), the challenge comes from the fact the existing data is poorly structured so it is difficult to interpret, see clearly what data are available, and what items are required. |
| **Current discoverability:** Data were described on the website and in a catalogue (<https://www.adrn.ac.uk/get-data/catalogue/> - no longer live). Each of the five projects had their own resources. Where possible, the information was published in the catalogue. Over time the catalogue changed scope, from listing possible datasets to listing only those known to be accessible. Some of the content was similar to information in existing catalogues. Researchers could also contact the ADRN directly, or perform web searches (e.g. “UK administrative data dentists”).  | **Future plans regarding discoverability:** The ADRN programme has ended and will be replaced |
| **eDRIS** |  |
| **Overview:** Service coordinating and advising on access to National Scottish health datasets and linkges. National datasets are available for reuse and must be analysed in the data safe haven. | **Perceived barriers to discoverability:** not discussed |
| **Current discoverability:** Datasets are shown in the National Data Catalogue (<http://www.ndc.scot.nhs.uk/>) which is maintained by ISD Scotland. In first instance, potential users are directed to the NDC. eDRIS offer one-to-one support to help potential users understand & access the data. | **Future plans regarding discoverability:** There have been discussions regarding creation of a catalogue combining info from all four of the Scottish data safe havens. |
| **Health Informatics Centre - Trusted Research Environment (University of Dundee)** |  |
| **Overview:** The data safe haven is run as part of the University of Dundee, affiliated with NHS Tayside and NHS Fife, and the TASK trial unit. Have many datasets (>100) covering approximately 20% of the Scottish population. Some go back ~60 years, though majority are from mid-90s. All datasets are linkable using the CHI. They are stored within a clinical repository accessible to the three team members, who are responsible for linkage and dataset extraction. Researchers are never given any identifiable data. Using a research data management platform (RDMP), datasets are prepared and extracted, meaning that (a subset of) the same fields are given to all researchers. Extracted datasets are transferred to the safe haven: projects are always independent within the safe haven although multiple users can collaborate on one project. There is an output checking process (all outputs are disclosure checked). | **Perceived barriers to discoverability:** In terms of how familiar researchers are with the data, some experienced users will be well versed due to prior experience. Many of the research assistants and PhD students have limited knowledge at first. In terms of publishing the datasets, there is a lack of tech support, and decisions required for who would be responsible (e.g. national vs local). Requires engineering expertise. There is a lack of standardisation across the datasets, other than those from the ISD. Each extracted dataset is bespoke. This issue is not necessarily taken seriously enough, potentially due to lack of time or lack of interest. Lack of consistency – easier to utilise the datasets in research if they are standardised. If a researcher wanted to run a project across multiple health boards, the time spent processing/ preparing the dataset could take up a significant portion of the study time. Could put users off. Room to improve training – embed bioinformatics and data engineering training within relevant courses. |
| **Current discoverability:** The website (<https://www.dundee.ac.uk/hic/datalinkageservice/datasetinventory/>) gives an idea of the scope and coverage of the datasets. Potential users will get most information when they discuss their projects with the team. The team present to external groups, particularly the RDMP, which can help show the efficiency of their system. The team will spend time helping the researchers to create a complete list of requirements, and also give a lot of advice re: appropriate requests. | **Future plans regarding discoverability:** Have previously discussed plans for making a public facing web front end based on the RDMP. Also have been discussions about making a Scotland-wide catalogue. With NHS partners, engage with other large projects, such as INSITE - Belgian group funded through a European technical innovation kickstart looking to improve efficiency in clinical trial feasibility assessment/scoping and recruitment. This is running within the safe haven – users are able to easily identify cohorts, and see active pharma studies.  |
| **NHS Greater Glasgow & Clyde Safe Haven** |  |
| **Overview:** Service coordinating access to local health datasets for research. ISD Scotland provide monthly updates of (cleaned) national datasets. Also, Glasgow Safe Haven have the most up-to-date copies of these datasets (not cleaned). Further datasets also available – Imaging safe haven. In addition, new data can be brought into the save haveb for linkage. If “deposited”, these datasets are then available for further sharing and are listed on website. | **Perceived barriers to discoverability:** The team deal with a vast amount of data, including more than are on the online list. Challenge to establish how much information to share, what information, and how to ensure all of the requirements/expectations are covered? E.g. those who are not experienced with health datasets are often surprised with how messy the data are. Others are not experienced with working with such large datasets and need more support. Team have considered sharing samples of synthetic data. |
| **Current discoverability:** Through the website or word of mouth. There is a list of available datasets on the website: <http://www.nhsggc.org.uk/about-us/professional-support-sites/nhsggc-safe-haven/your-research/available-datasets/> | **Future plans regarding discoverability:** The website is being developed / updated – this service will become West of Scotland later this year. Working out details of what will be presented on the site & how. Currently the data dictionaries are excel files – plans to integrate with datasets in the future. |
| **CALIBER (University College London)** |  |
| **Overview:** CALIBER is a platform for sharing data and methodologies. Data are from CPRD linked to HES, ONS mortality, and Cancer registry. Availability of data is based on collaborative model - needs to involve someone at UCL | **Perceived barriers to discoverability:** reported none |
| **Current discoverability:** Word of mouth. Launch event 18 months ago and publicised within UCL. Give presentations at conferences, have website, and have newsletter. The CALIBER portal gives details of the available research variables: <https://www.caliberresearch.org/portal>  | **Future plans regarding discoverability:** Want to improve discoverability, particularly among international collaborators - more talks. Feel well known within the relevant research networks. A database, such as those for clinical trials, would be beneficial. |
| **HMRC Data Lab (via email only)** |  |
| **Overview: “**The HMRC Datalab allows approved researchers to access de-identified HMRC data in a government accredited secure environment” <https://www.gov.uk/government/organisations/hm-revenue-customs/about/research#the-hmrc-datalab>  | **Perceived barriers to discoverability:** reported none |
| **Current discoverability:** There is a list of available datasets on the HMRC datalab website (<https://www.gov.uk/government/organisations/hm-revenue-customs/about/research#the-hmrc-datalab>). There is also a wider HMRC data catalogue (<https://www.gov.uk/government/publications/hmrc-data-catalogue>). | **Future plans regarding discoverability:** not asked |
| **Connected Health City North East and North Cumbria** |  |
| **Overview:** CHC North East and North Cumbria are developing infrastructure to connect local hospitals with their trustworthy research environment – this will include development of a metadata catalogue. Future data sharing/reuse is planned. | **Perceived barriers to discoverability:** None yet – still in design phase. Experience from past work (UK Birth Cohort Studies) – can be a challenge to determine what should be included in the catalogue. Specifically, some argue that any information about potentially sensitive data (e.g. telling people that data exists) is in itself sensitive and should be avoided. Also, retrospectively marking up datasets is a difficult task & so should be performed as the data are collected. |
| **Current discoverability:** n/a – in development | **Future plans regarding discoverability:** Currently planning to develop a catalogue. Envision two levels, an ‘internal’ catalogue with full metadata & covering all data available data items, to be viewable by appropriate clinical staff within the hospital, and an open but reduced catalogue directed at researchers. Rationale behind this is that some of the data will be identifiable, and the clinical staff already have the ‘right’ to view this, whereas the researchers will be accessing pseudonymised data so there is no need to display the full catalogue. Interviewee would favour a variable-level metadata catalogue with a sensible ontology to mark-up the datasets at the time of collection. |

|  |  |
| --- | --- |
| **Connected Health City Connected Yorkshire** |  |
| **Overview:** Connected Yorkshire work with the established Born in Bradford cohort study | **Perceived barriers to discoverability:** Key enabler is getting the data linked. |
| **Current discoverability:** Born in Bradford has a dedicated website, including a catalogue. Data dictionaries and question lists are available through this website: <https://borninbradford.nhs.uk/research/documents-data/>Lots of papers have been written about BiB over the years | **Future plans regarding discoverability:** Stick with current format. Would like to provide remote access to data – looking to move data into a data centre but also discussing with IT at hospital where currently held. |