Metadata Onboarding

Ben Gordon, Health Data Research UK
Jim Davies and James Welch, University of Oxford
Mike Jones, Parity, and Adam Milward, MetadataWorks

November 2019
Agenda

• Overall Aims (Ben)
  • Metadata Catalogue and architecture (Jim and James)
  • Metadata Standards (Mike and Adam)
  • Onboarding Walkthrough (Mike and Adam)
  • Support available (Mike and Adam)
  • Questions
Our mission is to unite the UK’s health data to enable discoveries that improve people’s lives
Infrastructure – uniting the UK’s health data

Ambition: We will create a world-leading data infrastructure and UK-wide approach to secure data services to accelerate scientific research and digital innovation

Priorities:

• Establish an ecosystem of Health Data Research Hubs
• Develop an Innovation Gateway to provide safe and secure access to UK’s health data
• Establish the UK Health Data Research Alliance
• Maximise the impact and learning from the Sprint Exemplar projects
• Enable proportionate, rigorous, efficient and transparent information governance across the UK Health Data Research Alliance
• Facilitate fair capture of value by the NHS and government for the benefit of patients and UK tax payer
UNITING THE UK’S HEALTH DATA TO MAKE DISCOVERIES THAT IMPROVE PEOPLE’S LIVES

TRUSTWORTHY USE OF DATA

UK HEALTH DATA RESEARCH ALLIANCE
Uniting the UK’s health data

- NHS DATA
- COHORT DATA
- REGISTRY DATA
- TRIALS DATA

Convened by

HEALTH DATA RESEARCH HUBS
Curating disease focused datasets and providing clinical trials and realworld evidence services

HEALTH DATA RESEARCH INNOVATION GATEWAY
Finding and accessing nationally aggregated, usable data

- DISCOVER
- DE-IDENTIFY
- LINK
- ACCESS
- ANALYSE
- PHENOTYPE LIBRARY

LINK TO TRUSTED RESEARCH ENVIRONMENTS (SAFE HAVENS)

Operated by

USERS:
Researchers, Industry, NHS, Charities and Government

PATIENTS

PATIENT AND PUBLIC BENEFIT
Health Data Research Innovation Gateway

The Gateway is a common access point to UK health research data for accredited researchers and innovators.

Developed in 2 phases:

**Phase 1 (Minimum Viable Product) – Sep 2019 to Jan 2020**
- Portal (Discovery and Access) – developed by IBM
- Metadata catalogue – delivered by NHS Digital and University of Oxford
- Metadata onboarding – delivered by Parity and Metadata Works

**Phase 2 – Mar 2020 to Aug 2022**
- Technology Partnership
Onboarding metadata for discovery

Metadata onboarding options

- Option 1: CSV table – bulk upload
- Option 2: Web-based metadata input
- Option 3: API

Responsibilities
- Gateway User
- Data Custodian
- Gateway User

Gateway Portal MVP

Metadata Catalogue (MDC)

Search and browse

Query API

API

Web-based metadata input

Metadata Input

Metadata Output

Dataset

Business metadata

Data Dictionary

Tooling?

Innovation Gateway Portal MVP

Metadata onboarding options

IBM

Oxford / NHS Digital

Parity/MetadataWorks & HIC Co-ordinating Centre

Business metadata

Data Dictionary

Tooling?
Agenda

• Overall Aims (Ben)
• Metadata Catalogue and architecture (Jim and James)
• Metadata Standards (Mike and Adam)
• Onboarding Walkthrough (Mike and Adam)
• Support available (Mike and Adam)
• Questions
Metadata Languages

- **essential for data management at scale**
  and for scalable, accountable, reproducible ‘big data’ research

- **describing properties of datasets, ...**
  supporting high-level discovery, selection, engagement

- **... data elements (fields, variables, columns), and data classes (forms, tables)**
  supporting detailed review of provenance, interpretation, and compatibility
  and data re-use

  with a high degree of automation
Metadata Catalogue

- a metadata database containing link-able descriptions - or models - of
  - datasets, populated databases
  - database schemas, XML schemas, forms
  - data standards
  - data transformations
  - phenotypes
  - ...
- with ‘importers’ for generating models of datasets and schemas
- and ‘exporters’ for generating schemas, forms, and scripts
- and role-based access control, and ...
Mental Health & Learning Disability Data Set

Author: NHS Digital
Organisation: NHS Digital

Description:
The Mental Health & Learning Disability Data Set (MHLDDS) contains record-level data about the care of adults and older people using secondary mental health, learning disabilities or autism spectrum disorder services at:
- NHS hospitals
- community clinics
- NHS-funded activity in the independent sector.

Type: Data Standard

Classifications:
- NHR Health Data Finder

<table>
<thead>
<tr>
<th>Data Classes</th>
<th>Types</th>
<th>Properties</th>
<th>Summary</th>
<th>Comments</th>
<th>History</th>
<th>Diagram</th>
<th>Links</th>
<th>Attachments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type</td>
<td>Name</td>
<td>Description</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>-------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| DataElement  | Communal establishment        | 5n or 1a *(Primitive)*
The communal establishment code is a five-digit code derived from the place of death as supplied on the death certificate. close |
| DataElement  | Date of registration          | Date *(Primitive)*
Date on which the death was registered                                                                                                    |
| DataElement  | NHS indicator                 | nhs_indicator *(Enumeration)*
Indicates whether the communal establishment code refers to an NHS establishment, referring to the physical building rather than the service close |
| DataElement  | Original underlying cause of death | 5an *(Primitive)*
This is an ICD* code that identifies the medical condition judged to be the underlying cause of death. Underlying cause of death can be defined as: a) the disease or injury which initiated the train of morbid events leading directly to death; or b) the circumstances of the accident or violence which produced the fatal injury. The underlying cause may be a long-standing, chronic disease or disorder that predisposed the patient to later fatal complications. *ICD - Deaths registered (Date of registration) before 01-Jan-2001 have the Original underlying cause of death and all cause of death mentions coded in ICD-9 and those registered since 01-Jan-2001 are coded in ICD-10 close |
<table>
<thead>
<tr>
<th>Description</th>
<th>Data Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type of prescribing exemption the patient has currently (e.g. medical or</td>
<td>prescf</td>
</tr>
<tr>
<td>maternity)</td>
<td>(Enumeration)</td>
</tr>
<tr>
<td></td>
<td>Key Value</td>
</tr>
<tr>
<td>0 Data Not Entered</td>
<td>0</td>
</tr>
<tr>
<td>1 Under 16 years of age</td>
<td>1</td>
</tr>
<tr>
<td>2 15, 17 or 18 and in full-time education</td>
<td>2</td>
</tr>
<tr>
<td>3 Woman aged 60 or over</td>
<td>3</td>
</tr>
<tr>
<td>4 Man aged 60 or over</td>
<td>4</td>
</tr>
<tr>
<td>5 Has a maternity/medical exemption certificate</td>
<td>5</td>
</tr>
<tr>
<td>6 Has a prescription prepayment certificate</td>
<td>6</td>
</tr>
<tr>
<td>7 Receives Income Support/Family credit et</td>
<td>7</td>
</tr>
<tr>
<td>8 Has a War Pension exemption certificate</td>
<td>8</td>
</tr>
<tr>
<td>9 Not Exempt</td>
<td>9</td>
</tr>
<tr>
<td>10 Gets Disability Working Allowance</td>
<td>10</td>
</tr>
<tr>
<td>11 Receives income-based Jobseeker’s Allowance</td>
<td>11</td>
</tr>
<tr>
<td>12 Is named on a current HC2 charges certificate</td>
<td>12</td>
</tr>
<tr>
<td>13 Was prescribed a free-of-charge contraceptive</td>
<td>13</td>
</tr>
<tr>
<td>14 Has a maternity exemption certificate</td>
<td>14</td>
</tr>
<tr>
<td>15 Has a medical exemption certificate</td>
<td>15</td>
</tr>
<tr>
<td>16 Receives Income Support</td>
<td>16</td>
</tr>
<tr>
<td>17 Has WFTC exemption or gets full or reduced WFTC</td>
<td>17</td>
</tr>
<tr>
<td>18 Has DPTC exemption or gets full or reduced DPTC</td>
<td>18</td>
</tr>
</tbody>
</table>
Mercury HIC Hepatitis v0.1

Data Model: Mercury
Version: 1.0.0
Last Update: 2018-04-10 14:16:08

Aliases
Author: Oliver Freeman
Organisation: Oxford BRC

Description:
This is the DataModel for the Mercury Database which holds the HIC Hepatitis Data. It is designed to match the collection XSD closely while fixing any known issues and representing the data in a relational form where XSD sometimes cannot.

Type: Data Asset

Classifications:
GSK
Oxford Hepatitis
Antimicrobial Resistance (AMR) v1.0
HIC: Acute Coronary Syndrome
HIC: ICU v3.2.2-Final
HIC: Ovarian Cancer Candidate Data Set v7.0.9.8 only
HIC: Transplantation v1.7.1
HIC HIC Cancer Model v1.1
HNR HIC Cancer Model v1.0.0
HNR HIC Cancer Model v1.0.1
HNR HIC Cancer Model v1.0.2
ePMA Pharmacy Administration
OHRI Research Data Warehouse - Archive
Remote DB-Test
StevWAP
TRUD
US Standards

Data Classes: properties, summary, comments, history, diagram, links, attachments, dataflow

Diagram: c10c242_hep_demographics
- bornoutside_the_uk_idx
- country_of_birth
- ethnicity_code
- fluency_id
- is_english_first_language
- language_id
- year_of_birth
- year_of_entry
- study_subject
- anonymised_subject
- anonymised_mrn_number
- anonymised_rns_number
- data_created
- gender_id
- last_updated
- message_id
- study_site_id
- study_subjects_id
- subject_key
- version
<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Importer</td>
<td>Oracle DB Importer</td>
</tr>
<tr>
<td>DataModel</td>
<td>Select the folder into which the DataModels should be imported.</td>
</tr>
<tr>
<td>DataModel name</td>
<td>Label of DataModel, this will override any existing name provided in the imported data. Note that if importing multiple models this will be ignored.</td>
</tr>
<tr>
<td>Finalised</td>
<td>Whether the new model is to be marked as finalised. Note that if the model is already finalised this will not be overridden.</td>
</tr>
<tr>
<td>Import as New Documentation Version</td>
<td>Should the DataModel be imported as new Documentation Version. If selected then any models with the same name will be superseded and the imported model will be given the latest documentation version of the existing DataModel. If not selected then the 'DataModel Name' field should be used to ensure the imported DataModel is uniquely named, otherwise you could get an error.</td>
</tr>
<tr>
<td>Database Name/Owner</td>
<td>The name of the database/owner which is to be imported.</td>
</tr>
<tr>
<td>Database Server</td>
<td>The name of the database server to connect to.</td>
</tr>
<tr>
<td>Database Host</td>
<td>The hostname of the server that is running the database.</td>
</tr>
<tr>
<td>Username</td>
<td>The username used to connect to the database.</td>
</tr>
<tr>
<td>Password</td>
<td>The password used to connect to the database.</td>
</tr>
<tr>
<td>Database Port</td>
<td>The port that the database is accessed through. If not supplied then the default port for the specified type will be used.</td>
</tr>
<tr>
<td>SSL</td>
<td>Whether SSL should be used to connect to the database.</td>
</tr>
</tbody>
</table>
Metadata Standards

• for describing properties of datasets
  DCAT-AP, W3C HCLS, schema.org
  vocabularies used by various organisations or communities of interest, with varying degrees of consistency and applicability
  need to agree a vocabulary for the HDR UK community

• for describing properties of data elements and data classes
  ISO/IEC 11179 +
  core language with well-defined semantics
  open API now, open source release next year
Agenda

• Overall Aims (Ben)
• Metadata Catalogue and architecture (Jim and James)
• Metadata Standards (Mike and Adam)
• Onboarding Walkthrough (Mike and Adam)
• Support available (Mike and Adam)
• Questions
What are the Metadata Standards - Background

As part of the development of the Innovation Gateway MVP, HDR UK must complete an initial onboarding activity to collect high-level information that describes the datasets that are available for research and innovation across members of the Alliance and Hubs.

• Information must be human and machine readable — and allowing the Gateway to index the information to return relevant search results.

• HDR UK’s initial requirements have been used as a baseline specification and iterated based on feedback and suggestions from IBM, Oxford and HDR UK. This is part of an MVP.

• We have tried to provide coverage for organisation who are already collecting metadata and would like to provide it to the gateway.

• We expect as the project progresses there will be further iterations of the specification – but this version applies for the current period.
Principles

The following principles have governed the work and decisions made:

- Specification that will be distributed in first iteration will be “good enough MVP”.
- Metadata should be based on a user need and a user story.
- Metadata should use standard terminologies whenever possible.
- Naming convention: Metadata "titles" will be lowercase, underscore separated.
  - (note that in the human readable specification we are using Capitalised for ease of reading)
- Prioritisation of Metadata has been based on Impact vs Effort. However, this will be easier to quantify as the project progresses.
## Metadata Specification Overview

<table>
<thead>
<tr>
<th>Summary</th>
<th>Business</th>
<th>Coverage and Detail</th>
<th>Format and Structure</th>
<th>Attributes</th>
<th>Technical Metadata</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identifier</td>
<td>Description</td>
<td>Geographic Coverage</td>
<td>Conforms to</td>
<td>Creator</td>
<td>Table Name</td>
</tr>
<tr>
<td>Title</td>
<td>Release Date</td>
<td>Periodicity</td>
<td>Controlled Vocabulary</td>
<td>Citations</td>
<td>Table Description</td>
</tr>
<tr>
<td>Abstract</td>
<td>Access Request Cost</td>
<td>Dataset End Date</td>
<td>Language</td>
<td>DOI</td>
<td>Column Name</td>
</tr>
<tr>
<td>Publisher</td>
<td>Access Request Duration</td>
<td>Dataset Start Date</td>
<td>Format and Structure</td>
<td></td>
<td>Column Description</td>
</tr>
<tr>
<td>Contact Point</td>
<td>Data Controller</td>
<td>Jurisdiction</td>
<td>File Size</td>
<td></td>
<td>Data Type</td>
</tr>
<tr>
<td>Access Rights</td>
<td>Data Processor</td>
<td>Population Type</td>
<td></td>
<td></td>
<td>Sensitive</td>
</tr>
<tr>
<td>Group</td>
<td>License</td>
<td>Statistical Population</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Derived Datasets</td>
<td>Age Band</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Linked Dataset</td>
<td>Physical Sample Availability</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Keywords</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Summary Metadata (Mandatory for MVP)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| **Identifier** | Please provide a local identifier used to identify the dataset (if available) | As A [user] - Researcher  
I Want To - know the name of identifier for the dataset  
So That - so I can identify, refer to it |
| **Title** | Completion Guidance: Name of the dataset  
Definition: A name given to the item. | As A [user] - Researcher  
I Want To - know the name of the dataset  
So That - so I can identify and find it |
| **Abstract** | Completion Guidance: Provide a short summary of the dataset (Max. 256 characters)  
Definition: A summary of the resource. | As A [user] - Researcher  
I Want To - read a plain English summary of the dataset  
So That - so that I can decide if I want to find out more |
| **Publisher** | Completion Notes: This is the organisation that is the data custodian.  
They are responsible for the data access request process / publishing / maintaining the metadata  
Definition: The entity responsible for making the item available. | As A [user] - Researcher  
I Want To - know the publisher  
So That - so that I can understand the context of the dataset |
| **Contact Point** | Completion Guidance: A URL with the information for and/or the email address of, the person or role within the data custodian organisation, who is responsible for the data access process and metadata maintenance  
Definition: Relevant contact information for the cataloged resource | As A [user] - Researcher  
I Want To - contact person  
So That - so that I can understand the datasets and request access |
| **Access Rights** | Completion Guidance: Text or link to website/documentation where data access request process and/or guidance is provided  
Definition: Information about who can access the resource or an indication of its security status. | As A [user] - Researcher  
I Want To - know what I can do with the data  
So That - so that I have a legal and ethical basis for analysing it in my research |
| **Group** | Completion Guidance: complete if the dataset is part of a group family or collection i.e. Hospital Episode Statistics has several constituents  
Definition: A collection is an entity that provides a structure to some constituents that must themselves be entities. These constituents are said to be member of the collections. | As A [user] - Researcher  
I Want To - know whether the dataset is part of a wider group  
So That - so that I can find related datasets |
Metadata Specification to support Innovation Gateway Minimum Viable Product (MVP)

November 2019
Agenda

- Overall Aims (Ben)
- Metadata Catalogue and architecture (Jim and James)
- Metadata Standards (Mike and Adam)
  - Onboarding Walkthrough (Mike and Adam)
- Support available (Mike and Adam)
- Questions
MetaData onboarding – who and what is involved?

Gateway MVP development team

Meta data catalogue development team

(Meta) data set onboarding team

IBM

NHS Digital & Oxford University

Parity & MetadataWorks

Future Alliance members x ??

Current Alliance members x 25 (Including HIC)

Priority for support and funding

Funded through DIH monies

Hubs x 7

Health Data Finder

Transition
Onboarding metadata for discovery

Metadata onboarding options

- Option 1: CSV table – bulk upload
- Option 2: Web-based metadata input
- Option 3: Tooling?

Responsibilities
- IBM
- Oxford / NHS Digital
- Parity/MetadataWorks & HIC Co-ordinating Centre

Metadata Catalogue (MDC)

Search and browse

Innovation Gateway Portal MVP

Gateway User

Query API

API

Data Custodian

Metadata Output

Dataset

Business metadata

Data Dictionary

Innovation Gateway Portal MVP

Metadata onboarding options

Gateway User

Search and browse

Query API

API

Metadata Catalogue (MDC)
Metadata Onboarding team – scope of work

- **Rapid evaluation of the current landscape of Alliance and Health Data Research Hubs datasets**, the metadata status, the data custodian’s readiness for onboarding, and potential investment required to develop a sustainable process for onboarding and updating metadata in a timely fashion.

- **Co-ordinate the metadata onboarding process for Alliance members and Health Data Research Hubs**, working closely with the metadata catalogue supplier and providing advice and targeted support to Alliance members to complete the onboarding process.

- **Demonstrate ‘proof of concept’ for onboarding the metadata** from at least two Alliance members that have not previously been involved with Health Data Finder.

- Identify and, where possible, address **opportunities to improve the efficiency and sustainability of the process** for the benefit of future datasets and ensuring metadata available through the Gateway remains current.
Hubs

- Funded through DIH monies
- Expected to onboard metadata for current datasets
- Minimal support available from MDW / Parity
Two phase approach to investment in Alliance members

1. **Rapid evaluation** of ‘within scope datasets’ across all members of the UK Health Data Research Alliance.

2. **Targeted support** for preparation and onboarding of metadata based on prioritisation following the rapid evaluation work.

- Both phases will be co-ordinated by Parity & MetadataWorks and form part of the development of the Innovation Gateway Minimum Viable Product (MVP).
- The MVP is scheduled to be launched on 10 January 2020.
- It is expected that funded onboarding of metadata will continue until the end of March 2020 and demonstrate how metadata can continue to be added and updated.
Phase 1: Rapid evaluation work

• **Organisational engagement fee.** Same across all Alliance members who agree to engage. It is proposed this should be £5,000. Payment of this engagement fee would be the return within 2 weeks of an information request covering:
  
  • “Within scope dataset return”. List of datasets/data assets that are in scope for the rapid assessment and potential loading onto Innovation Gateway Metadata Catalogue.
  
  • Provide basic access request process or documentation (or links to it) – this could be achieved by confirming that the information provided on ukhealthdata.org is correct

  • Provide their data access register and any other transparency info (or links to them)

• **Within Scope dataset fee.** This would cover the costs associated with providing the basic dataset level information requested above for each dataset in scope for the rapid assessment and supporting the Parity/MetadataWorks team to carry out an assessment of effort required to fulfil the metadata requirements of the MVP.

  • Differential fee between distinct datasets and derived or related datasets (where the business metadata is essentially the same as another in scope dataset).
Proposed Conditions of Investment

- To be eligible for funding Alliance members must provide rapid access to relevant staff and facilitate rapid assessment and work to prepare metadata onboarding.

- This proposal is about metadata to support discovery as set out in the Alliance Letter of Intent. There would be no changes to rights or control of the underlying dataset/data asset.

- The data custodian would grant HDR UK royalty-free, unrestricted and non-exclusive rights to the current and future metadata for use in MVP and any subsequent Innovation Gateway development. These rights would be perpetual and cover use of the metadata by public, academic, commercial, not-for-profit, charitable and voluntary sector entities whether based or accessing the metadata in the UK or overseas. (i.e., HDR UK would hold a metadata asset at the end of the process).

- The data custodian commits to keep the metadata subsequently uploaded to the catalogue up to date.

- The decision on what represents a distinct dataset would rest with HDR UK and any attempts to ‘game’ the approach would be considered contrary to the Alliance Letter of Intent and Principles for Participation.

- This is a one-time investment to support development and testing of MVP. It does not set a precedent for any future decisions or funding of metadata provision. No further investment should be expected, and organisations will be required to provide a sustainability plan to keep the metadata up to date and onboard metadata of future datasets that does not assume any further financial support from HDR UK.

- No payment would be available for datasets already in Health Data Finder unless there is a clear and agreed need for improvement.
Overlapping areas of Interest

Alliance Members

Hubs

MDW / Parity

HIC Co-ordination Centre

HIC

HDR UK
Thank you

Find out more:

Health Data Research UK
Web: hdruk.ac.uk

UK Health Data Research Alliance
Web: ukhealthdata.org

Social: @HDR_UK
Email: enquiries@hdruk.ac.uk