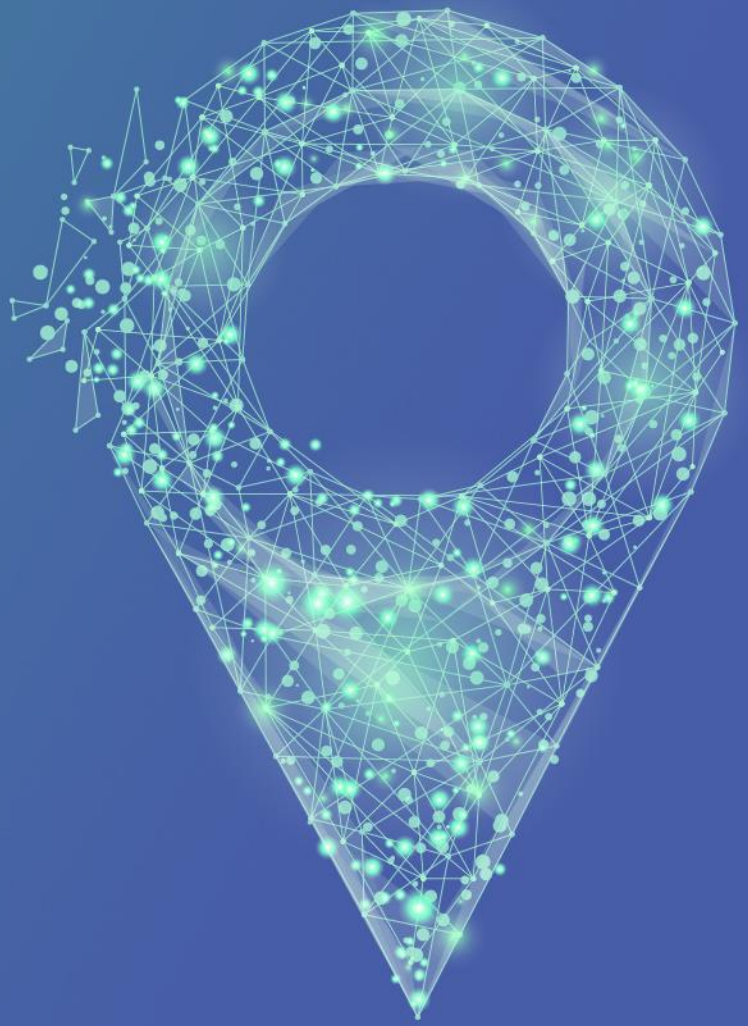
HDRUK
Health Data Research UK

INDUSTRIAL STRATEGY | UK Research and Innovation



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**

Option

HDRUK
Health Data Research UK

INDUSTRIAL STRATEGY | UK Research and Innovation

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

INDUSTRIAL STRATEGY | UK Research and Innovation

UNITING THE UK'S HEALTH DATA TO MAKE DISCOVERIES THAT IMPROVE PEOPLE'S LIVES

HDRUK Health Data Research UK

TRUSTWORTHY USE OF DATA

PATIENTS

UK HEALTH DATA RESEARCH ALLIANCE

Uniting the UK's health data

UK Health Data Research Alliance

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

Convened by HDRUK Health Data Research UK

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 HDRUK Health Data Research UK

USERS:
Researchers, Industry, NHS, Charities and Government

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:

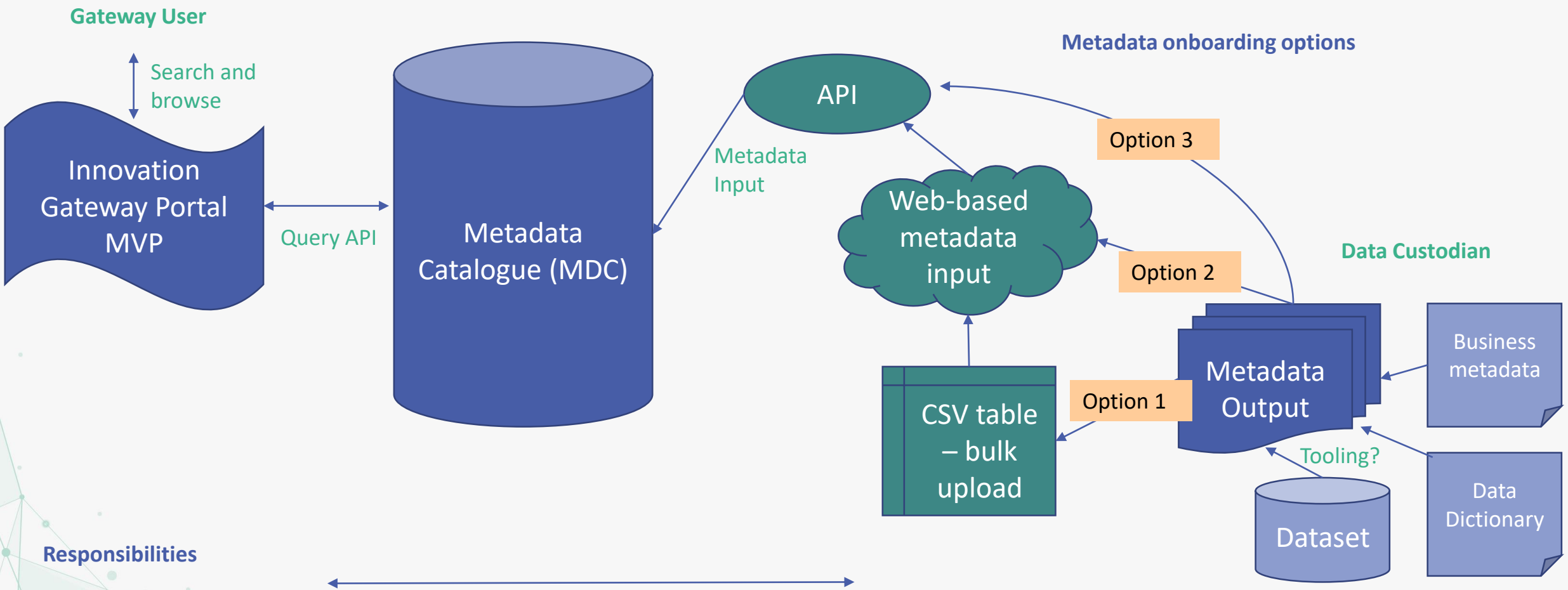
Phase 1 (Minimum Viable Product) – Sep 2019 to Jan 2020

- 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



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 …**

Models | Classifications | Favourites

Search for...

- Clinical Audits
- Genomics England Schemas
- Global Research on Antimicrobial Resistance (GRAM)
- HIC
- HIC Cancer Network Data Integrator
- Healthcare Reference Models
- Healthcare Terminologies
- ISO Standards
- Million Women Study
- Miscellaneous
- NHS Digital : TRUD
- NIHR Health Data Finder (HDF)
  - Clinical Practice Research Datalink (CPRD)
  - NHS Digital
    - Diagnostic Imaging Dataset
    - HES: A&E attendance
    - HES: Admitted Patient Care
    - HES: Adult Critical care
    - HES: Outpatient
    - **Mental Health & Learning Disability Data Set**
    - NNRD: National Neonatal Research Database
    - ONS: Mortality dataset
    - Patient Reported Outcome Measures
    - Secondary Uses Services
  - NIHR Health Informatics Collaborative (HIC)
  - Public Health England (PHE)
- NIHR Health Informatics Collaborative (HIC)
- OUH Research Data Warehouse - Archive
- ~~Remote DB Test~~
- SteveWIP
- TRUD
- US Standards

Mental Health & Learning Disability Data Set

Data Model Finalised

Version: 2.0.0

Last Update: 2018-05-02 12:01:25 *Finalised On: 2018-03-07 11:03:03*

Aliases	
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	NIHR Health Data Finder

Data Classes | Types | Properties | Summary | Comments | History | Diagram | Links | Attachments

Data Classes 50

	Name	Description
1..1	Leave of Absence (LOA)	
1..1	Master Patient Index (MPI)	
1..1	Psychosis Services (PSYCHOSIS)	
1..1	Employment Status (EMP)	
1..1	Accommodation Details (ACCOMM)	
1..1	Referral (REFER)	
1..1	Mental Health Team Episode (TEAMEP)	
1..1	NHS Day Care Episode (DAYEP)	
1..1	Consultant outpatient Episode (OPEP)	
1..1	Acute Home-Based Care Episode (HBCAREEP)	
1..1	Mental Health NHS Care Home Stay Episode (NHSCAREHOMEEP)	
1..1	Hospital Provider Spell (PROVSPELL)	
1..1	Inpatient Episode (INPATEP)	
1..1	Ward Stay within Hospital Provider Spell (WARDSTAYS)	
1..1	Delayed Discharge (DELAYEDDISCHARGE)	

Content | Properties | Comments | Links

Content 4

Type		Name	Description
DataElement	*0..1*	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	*0..1*	Date of registration	Date *(Primitive)* Date on which the death was registered
DataElement	*0..1*	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	*0..1*	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

Prescription Exemption Draft

Data Element

Last Update: 2017-09-25 09:06:48

Description	Type of prescribing exemption the patient has currently (e.g. medical or maternity)
Data Type	prescr *(Enumeration)*

Key	Value
0	*Data Not Entered*
1	*Under 16 years of age*
2	*16, 17 or 18 and in full-time education*
3	*Woman aged 60 or over*
4	*Man aged 60 or over*
5	*Has a maternity/medical exemption certificate*
6	*Has a prescription prepayment certificate*
7	*Receives Income Support/Family credit et*
8	*Has a War Pension exemption certificate*
9	*Not Exempt*
10	*Gets Disability Working Allowance*
11	*Receives Income-based Jobseeker's Allowance*
12	*Is named on a current HC2 charges certificate*
13	*Was prescribed a free-of-charge contraceptive*
14	*Has a maternity exemption certificate*
15	*Has a medical exemption certificate*
16	*Receives Income Support*
17	*Has WFTC exemption or gets full or reduced WFTC*
18	*Has DPTC exemption or gets full or reduced DPTC*

Models | Classifications | Favourites

Search for...

- Clinical Audits
- Genomics England Schemas
- Global Research on Antimicrobial Resistance (GRAM)
- HIC
- HIC Cancer Network Data Integrator
- Healthcare Reference Models
- Healthcare Terminologies
- ISO Standards
- Million Women Study
- Miscellaneous
- NHS Digital : TRUD
- NIHR Health Data Finder (HDF)
- NIHR Health Informatics Collaborative (HIC)
  - GSK
  - **Oxford Hepatitis**
  - Antimicrobial Resistance (AMR) v1.0
  - HIC: Acute Coronary Syndromes
  - HIC: ICU v8.3.2-Final
  - HIC: Ovarian Cancer Candidate Data Set v7.0.9 only
  - HIC: Transplantation v1.7.1
  - NIHR HIC Cancer Model v.1.1
  - NIHR HIC Cancer Model v1.0.0
  - NIHR HIC Cancer Model v1.0.1
  - NIHR HIC Cancer Model v1.0.2
  - ePMA Pharmacy Administration
- OUH Research Data Warehouse - Archive
- ~~Remote DB Test~~
- SteveWIP
- TRUD
- US Standards

Mercury HIC Hepatitis v0_1

Data Model Finalised

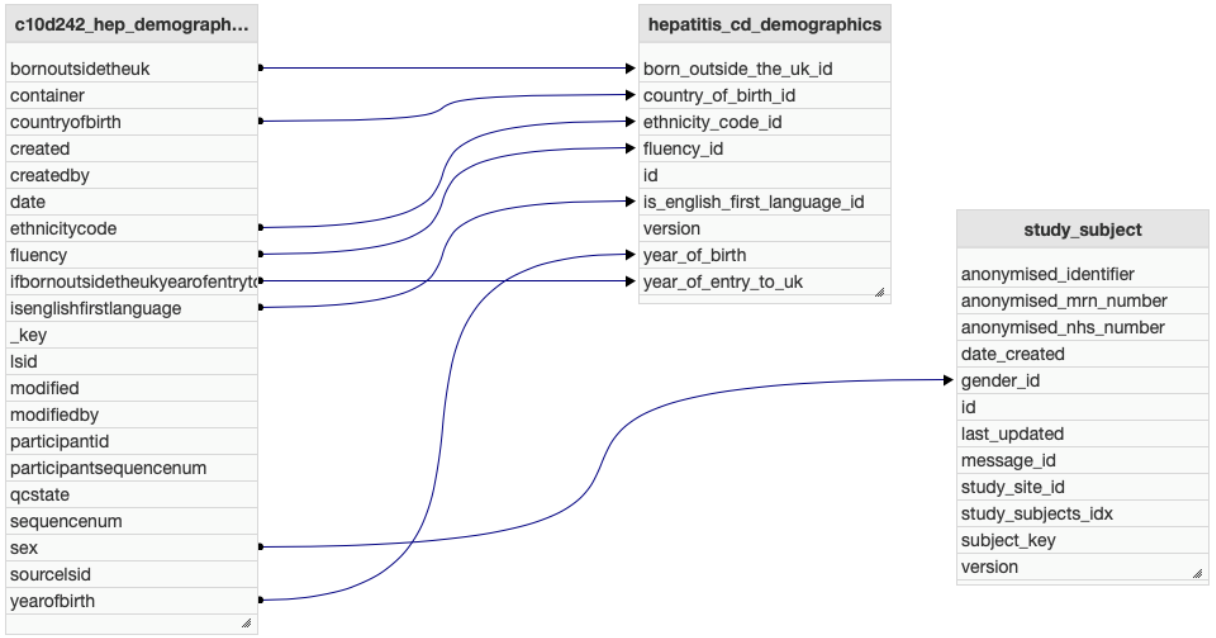
Version: 1.0.0

Last Update: 2019-04-10 14:16:09 *Finalised On: 2018-09-28 21:21:25*

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	

Data Classes | Types | Properties | Summary | Comments | History | Diagram | Links | Attachments | Dataflow

Reset — **Oxford Hepatitis to Mercury HIC Hepatitis**



Import

1 Please select an importer:

Importer *

Oracle DB Importer

2 Please fill the following form and press submit button:

DataModel

Folder*

The folder into which the DataModel/s should be imported.

Select

DataModel name

Label of DataModel, this will override any existing name provided in the imported data. Note that if importing multiple models this will be ignored.

Finalised

Whether the new model is to be marked as finalised. Note that if the model is already finalised this will not be overridden.

Import as New Documentation Version

Should the DataModel/s be imported as new Documentation Version/s. If selected then any models with the same name will be superseded and the imported models will be given the latest documentation version of the existing DataModels. 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.

Database

Database Name/Owner*

The name of the database/owner which is to be imported.

Database Server*

The name of the database server to connect to.

Database Host*

The hostname of the server that is running the database

Username*

The username used to connect to the database.

This field is required.

Password*

The password used to connect to the database.

Database Port

The port that the database is accessed through. If not supplied then the default port for the specified type will be used.

SSL

Whether SSL should be used to connect to the database.

Submit Reset

HDRUK
Health Data Research UK

DUSTRIAL RATEGY | UK Research and Innovation

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

HDRUK
Health Data Research UK

INDUSTRIAL STRATEGY
UK Research and Innovation

Summary	Business	Coverage and Detail	Format and Structure	Attributes	Technical Metadata
Identifier	Description	Geographic Coverage	Conforms to	Creator	Table Name
Title	Release Date	Periodicity	Controlled Vocabulary	Citations	Table Description
Abstract	Access Request Cost	Dataset End Date	Language	DOI	Column Name
Publisher	Access Request Duration	Dataset Start Date	Format and Structure		Column Description
Contact Point	Data Controller	Jurisdiction	File Size		Data Type
Access Rights	Data Prcoesser	Population Type			Sensitive
Group	License	Statistical Population			
	Derived Datasets	Age Band			
	Linked Dataset	Physical Sample Availability			
		Keywords			

Summary Metadata (Mandatory for MVP)

Identifer	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

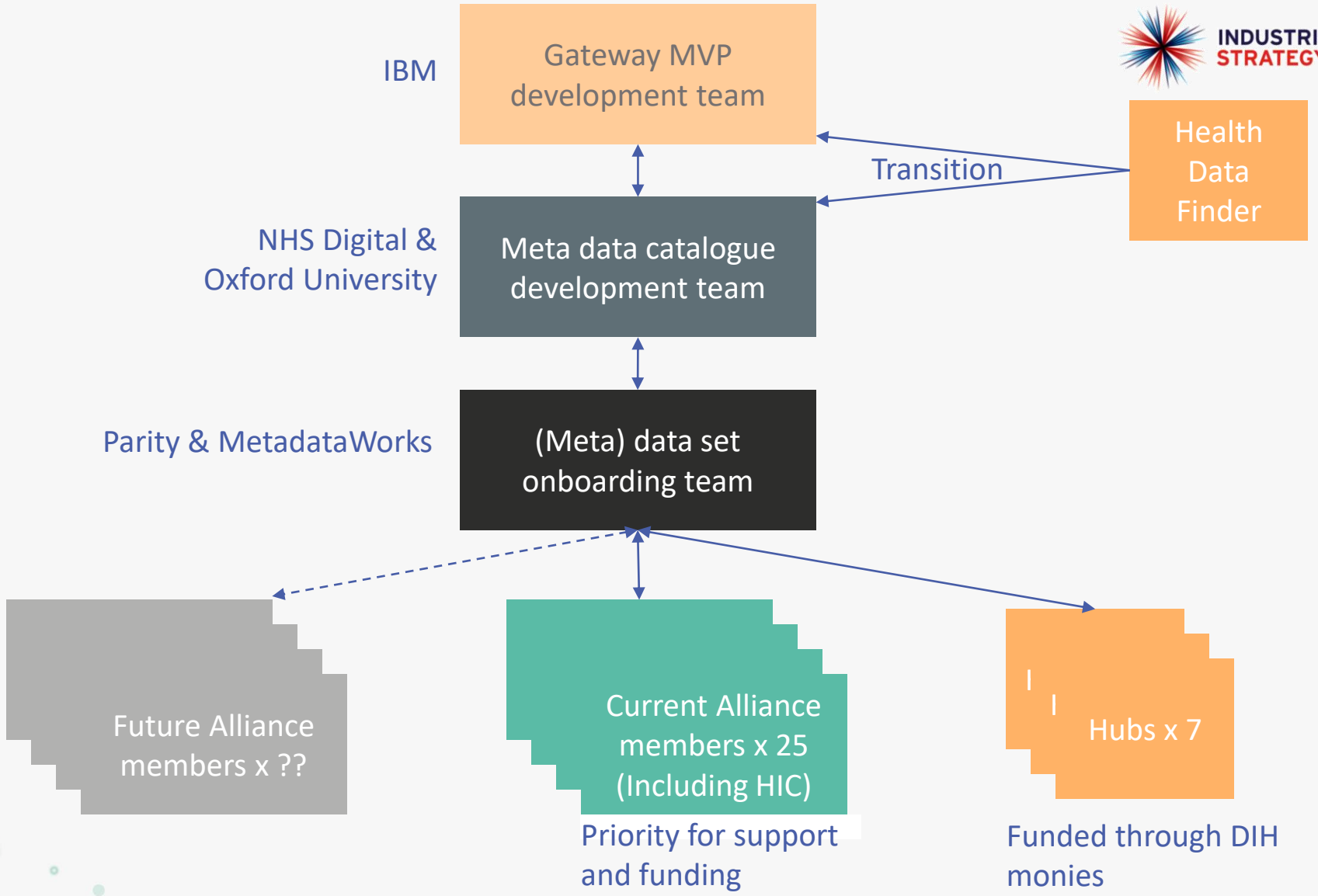
Reference

Document: MVP Metadata Specification_Final v1.1.3

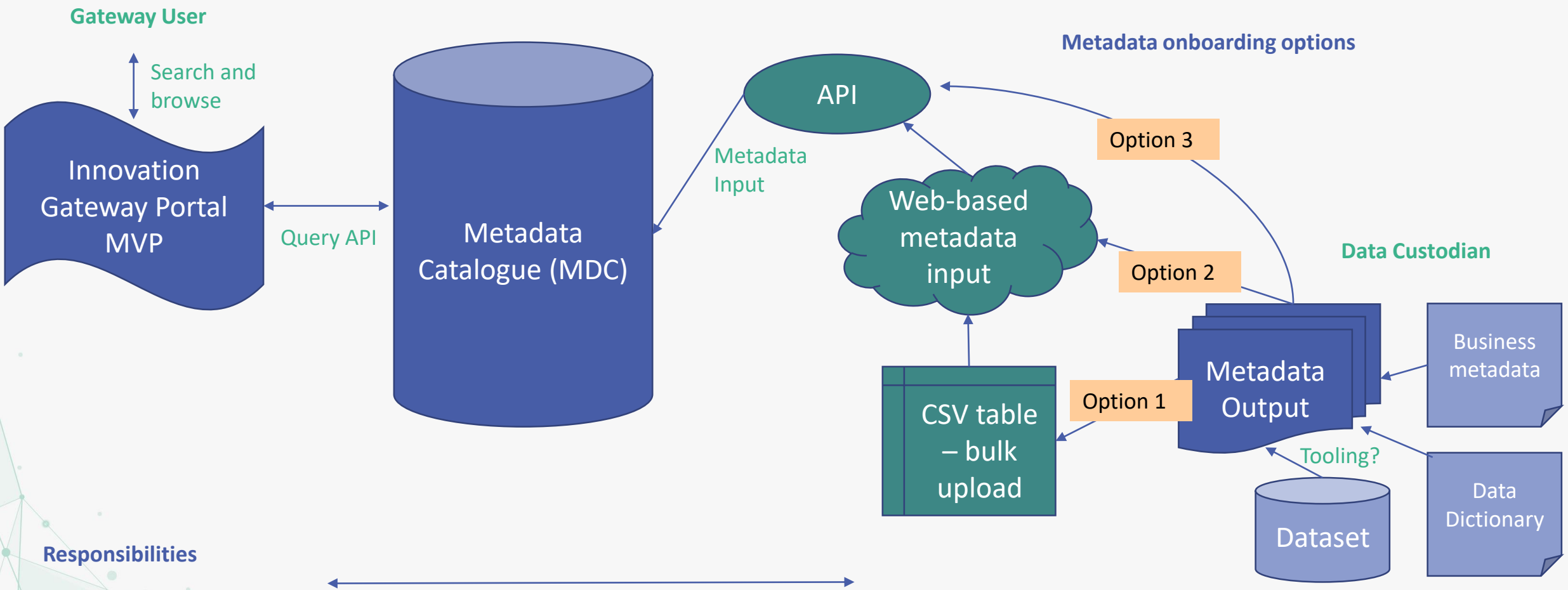
Agenda

- 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?



Onboarding metadata for discovery



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

- **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

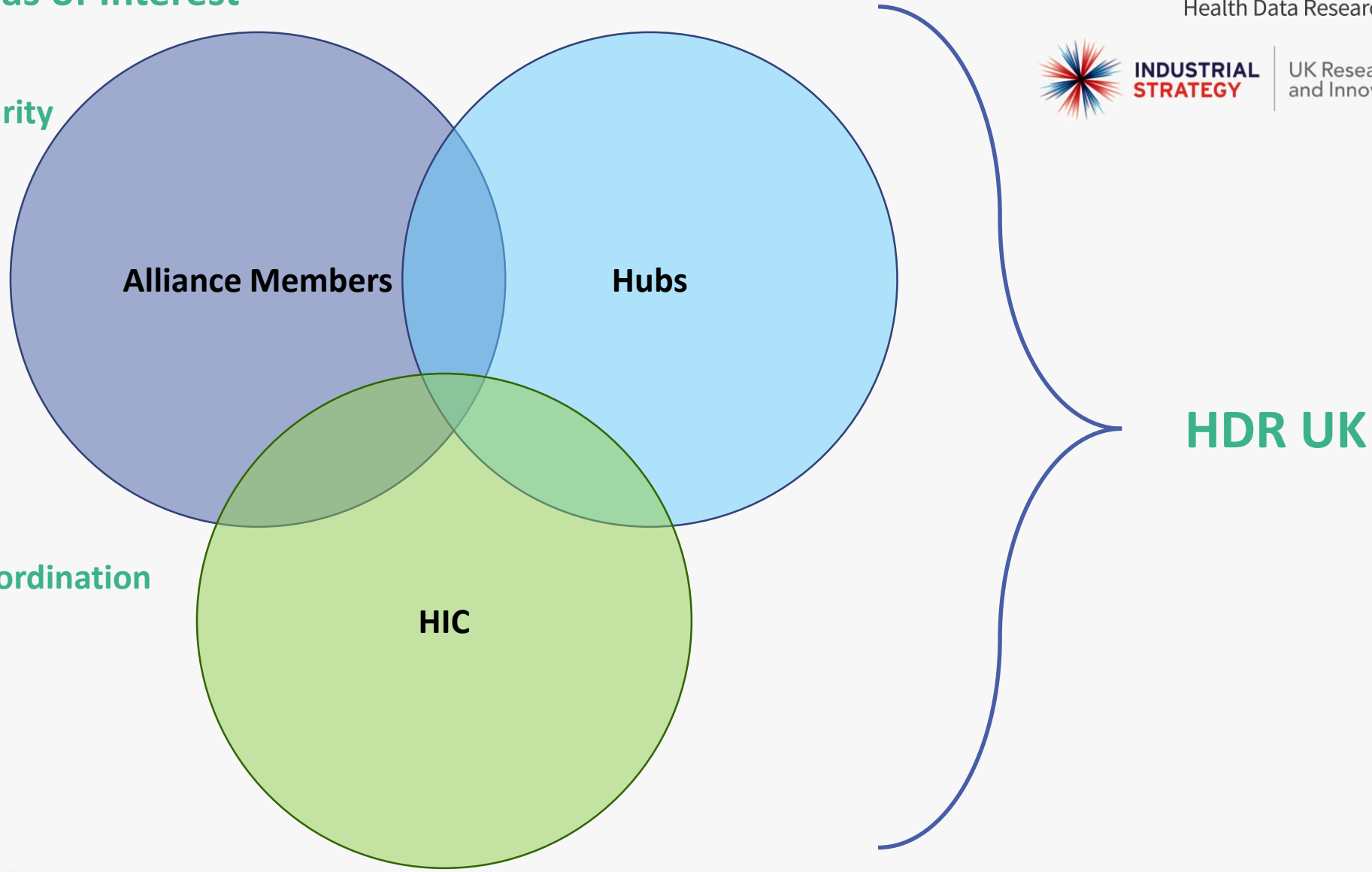
- **Organisational engagement fee.** Same across all Alliance members who agree to engage. It is proposed this should be £5,000. Payment of this engagement fee would be the return within 2 weeks of an information request covering:
  - "Within scope dataset return". List of datasets/ data assets that 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



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