



Health Data Research UK

The national institute for health data science

DOG Update - Data Design Authority (DDA)

Monica Jones

Chief Data Officer – University of Leeds

Associate Director & National Strategic Lead - HDRUK

Data Design Authority (DDA)

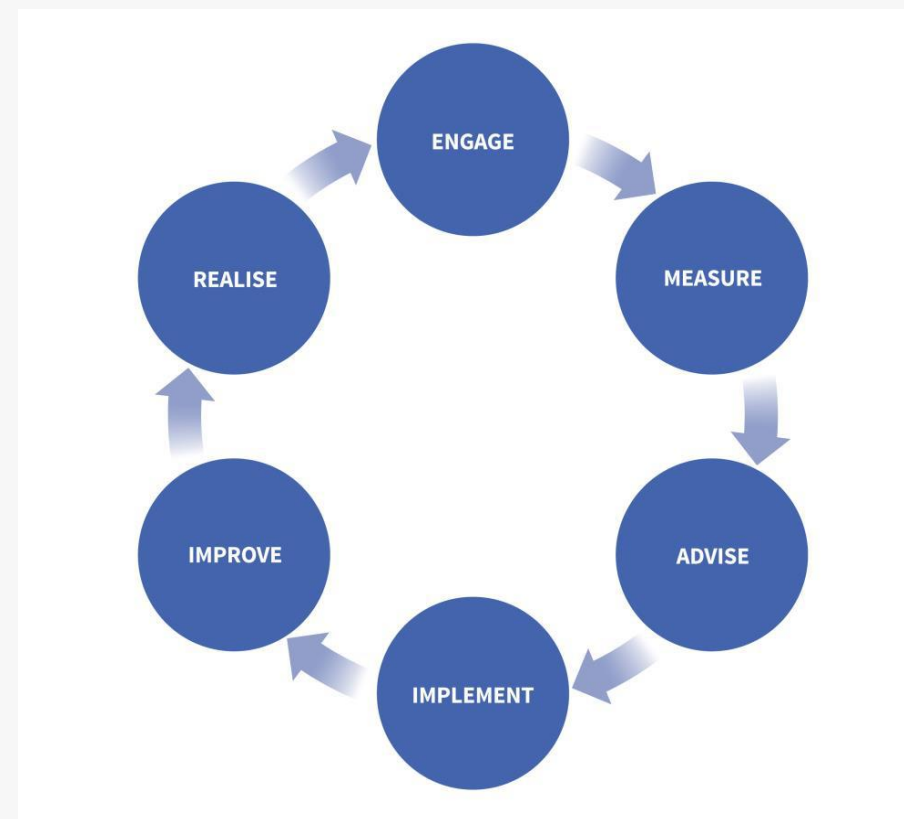
Data Design Authority has been established across the Useable Data Pillar, building on the approach trialled by HDR UK North in the first five years.

This group of experts with knowledge and experience will support projects that take place across the Institute, providing guidance and allowing for more rapid data curation than if this process had taken place in isolation.

The group will also ensure convergence in standards across the Driver Programme data foundations to enable future interoperability.

The DDA needs to engage with HDRUK partners as well as key scientific partners and PPIE community. It can then move through the life cycle as proposals come forward for datasets and projects.

Membership will include representatives to advise on behalf of HDRUK.



Data Design Authority (DDA) lifecycle

It is important to have a set of data principles to align to. These can be used by the Data Design Authority (DDA) to support and guide development of Driver Programme data products.

We need to ensure alignment with Alliance [Data Standards Principles](#) and [Data Standards Recommendations](#) and of course overarching Alliance [Principles of Participation](#)

DDA will undertake a review and update as appropriate

Item	Principle	Description
1	All data is valuable	For us the value of data is not just in the way we could monetise it, but from how it contributes to the delivery and assurance of safe, compassionate and effective education and research, and how the data we use and create supports us in delivering our activities and enabling effective decision making.
2	All data has an owner	It is important with all the data we hold or access that we ensure we understand what role we are fulfilling in order to ensure we understand the responsibilities we have and can work with it appropriately.
3	Data must be understood	Data is a representation of some aspect of reality and can only be used effectively, appropriately and reliably to contribute to valuable outcomes if it is understood.
4	Data must have a known purpose	If we cannot identify a purpose for the data we gather we should not be gathering it. It is desirable for data to have multiple purposes (and essential that all these purposes are recorded) but there will usually be one primary reason the data is valuable to us.
5	Data must have context	To be able to understand data and reuse it effectively it is important that we understand some basics about the data itself.
6	Data must have known quality	To ensure that data can be used safely to drive design and decision making it is important that the quality of the data is recorded.
7	Data should be open	Wherever possible we will work to open data and cross government standards, this will ensure the data quality is not eroded by avoidable / unnecessary transformation / translation.
8	Data use is traceable, legal and ethical	It should go without saying that the use we put data to should be bound within the legal and ethical framework we work within.
9	Use data to prompt appropriate action	Data should be used to initiate traceable action when appropriate; over time we should be able to identify patterns in the data that we have discovered often accompany problems. The nature and severity of problems associated with data patterns should prompt appropriate action within our systems whether these be technical or procedural.
10	Data should be digital by default	Although data in this strategy does not just refer to digital data and our vision of data is for data in all its forms, where possible and practical we will digitise physical data to maximise its utility.
11	Reuse data, don't recreate it	Reuse of data must be preferred over recreating or recapturing it, but to reuse it effectively it must be in media and formats that assist its reuse with minimum effort.

Update on Progress I&I Programme: Systematic Data Curation



- BREATHE/Inflammation and Immunity Driver Programme systematic data curation work on respiratory datasets [see here](#) for published paper. Proposed approach not necessarily disease agnostic – although the initial I&I Driver work has been, this is about reproducible algorithms and code that can enable better curated datasets and more reusability across regions/HDRUK programmes.
- Common data models are relevant, but this proposed work is earlier down the data curation ‘pathway’ where research teams do not necessarily want to lose any of the granularity that mapping to e.g OMOP may enable – this work fits within and is complementary to the wider HDRUK Useable Data Pillar.
- I&I Driver also aligning with HDRUK Federated Analytics programme – 2 x federated approaches being assessed, one very dependent on mapping to common data model, other one less so
- Get ‘under the hood’ of Drivers on planned and current data curation approaches. Mapping out initial scope, priorities and timelines (and resource) important first step – timelines for data access and curation will be different across Drivers.
- Building out work led by BHF Data Science Centre with NHSE important too – clear alignment. Building in checks and validation. Medicines Driver is starting out on looking at reproducibility of analyses for projects across Birmingham, Bristol, Scotland variety of data providers/TREs etc across (e.g. CPRD, BadgerNet, SAIL, PIONEER etc) – and on project-by-project basis (multiple diseases, multiple datasets). Molecules to Health Records – bring in co leads who are leading work on bowel disease, paediatric neurodevelopmental disorders
- Priority data providers – initial I&I Driver work has focused on devolved national TREs, CPRD, DataLoch (regional Scotland). Good alignment here across other Drivers, but capacity and engagement may be trickier for NHSE, CPRD. Green algorithms should also be a driver for this work.

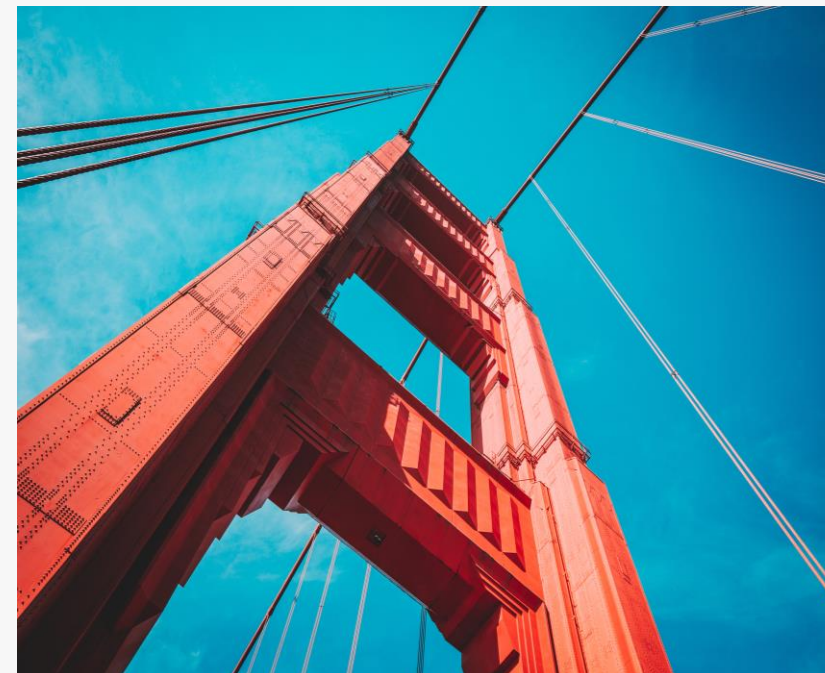
Driver Systematic Data Curation - next steps

- ALL to review map of disease/condition areas and data providers/datasets for their Driver on [Google sheet here](#) and update as far as possible
- Review once Google sheet updated – suggest initial list of priority data providers/TREs with group– then coordinate wider workshop to include Drivers and data providers/TREs, at minimum devolved nation TREs, NHSE (via BHF Data Science Centre)
- All to forward names of Driver team members (e.g. workstream leads) who we should also link into next steps – we can prioritise who we bring in and when depending on initial priority conditions/phenotypes and data providers
- Chris and Lara to start to draft project plan and timelines
- Coordinate next meeting for September 2024



Update on Medicine Driver Programme: Gateway Integration

- Updated LeHMR medicines mapping use case
- Cross referenced to Gateway development and metadata standards (including API connection)
- Further work required to take forward. DDA to work on roadmap to align with NHS England, Scotland, Wales and NI
- Context of 'National Data Library' and Government Open data standards, explicitly reference the link to DCAT <https://www.gov.uk/government/publications/recommended-open-standards-for-government/using-metadata-to-describe-data-assets-in-a-data-catalogue>



Update on Big Data for Complex Disease: Data Curation and Challenges



Notebooks

- Tutorials
- Data summaries
- Data insights
- ...

Knowledge sharing

Code repositories

- Common
 - Functions
 - Curation pipelines
 - ...
- Projects
 - CCU001_01
 - CCU002_01
 - ...

Findable reusable code

Curated tables

- Skinny
- LSOA
- COVID-19 inf/vacc
- ...

Energy efficient computing

Dataset summary dashboard

Code list comparison tool



Data dictionary

Dataset documentation

Best practice guidance

- Induction
- Efficient coding
- Tutorials
- GitLab webinar
- RStudio Server guidance
- Code list management
- Data curation specification



BHF DSC

- Common
 - Functions
 - Curation pipelines
 - ...
- Projects
 - CCU001_01
 - CCU002_01
 - ...





Monica Jones FBCS CITP MInstLM – m.c.m.jones@leeds.ac.uk