UK Research and Innovation

HDRUK Health Data Research UK

ADRUK Data-driven change

DARE UK – Early Thinking

Data and Discovery

Stakeholder workshop, Thursday 10 March 2022

Sensitive Data – Possible Definition

"Sensitive data includes data with directly or indirectly **identified information** such as names and addresses as well as directly or indirectly **identifiable data which have been de-identified** (has had all personal identifying information removed), but nevertheless **remains sensitive due to the potential for re-identification, attribution or secondary disclosure**. For the purposes of this programme, sensitive data will also be considered **to include other data associated with individual people** such as financial information, retail profiles, social media activity, and location and movement information. These data may come from data sources across academia, government, third sector and industry."

Why Data & Discovery?

- Data lifecycle management (including for associated metadata) allows Trusted Research Environment (TRE) operators to ensure the right data is shared with the right people, for the right purpose, with appropriate permissions and governance applied in the right setting
- The principles that ensure efficient data sharing practices are the Findable, Accessible, Interoperable and Reusable (FAIR) principles
- Increased visibility for data available within TREs, allows more research to be conducted
- High quality metadata allows novel linkage of datasets between different domains to be discovered for innovative research and support reuse of derived datasets

What is Metadata?

- Metadata is data about data. Metadata enriches the data with information, which makes it easier to discover, use and manage.
- There is a wide variety of metadata depending on its purpose, format, quality and volume. Some of the widely used categories of metadata are: descriptive, structural, administrative and statistical.

What is a Digital Object Identifier (DOI)?

- Digital Object Identifiers, commonly shortened to DOIs, were invented to give each electronic, or digital, item a unique, persistent identifier. Any digital object can be assigned a DOI number, for example:
  - academic journal articles
  - research reports
  - datasets
  - books and media
  - conference proceedings
  - code

First Draft Recommendations (1 of 5)

Enhance the data lifecycle for cross-domain research

- Pilot cross-council automated provisioning pipelines for sensitive and open data, with analysis being conducted in TREs
- Develop an approach for data provisioning between federated TREs to support usecases where federated analytics is not technically feasible
- Review options for a third party linkage service that can be deployed across the TREs and support different council domains
- Building upon existing best practice, design archival capability that can be integrated with TREs and support cross-council use. Develop the business case for production development and deployment in DARE UK Phase 3

Your thoughts and feedback

Enhance the data lifecycle for cross-domain research

- Scrutinise projects and approve approaches, use of data and methodology to ensure data isn't disclosed, rather than check outputs
- Need approach to link with potentially very large open datasets
- Third party linkage service might be useful, provided it is available fairly and not too expensive
- Need further study into what data to keep, and what not to keep (also related to novel / 'Internet of Things' data)
- Minimise data travel; only move data when essential
- Do we need a data steward organisation to bring skills?
- Sometimes too many approvals needs from data access to provisioning
- Need to ensure any third-party linkage service is secure. If Open Source, would it be more vulnerable to attack?

First Draft Recommendations (2 of 5)

Explore implications of new data types, models for sharing and velocity of delivery

- Review emerging data requirements for new data types (e.g. use of wearables), delivery models beyond datasets such as streaming data, and requirements for near real time access
- Develop lifecycle model to address these requirements
- MVP of service to support near real time flow of IoT data such as wearables using streaming technology

Your thoughts and feedback

Explore implications of new data types, models for sharing and velocity of delivery

- Wearable will be important, but data preparation, checking validity/accuracy, and de-identification may not be easy
- Need to consider how to stream this data
- Could this lower the environmental cost of collecting data, contributing to the NetZero agenda?
- Current legislation and governance is not ready for near real-time delivery
- Innovation here should be driven by the research use cases and not technology led
- Explore new research opportunities on streaming data, temporal characteristics
- How to do de-identification on-the-fly; what are the risks?
- Need to consider storage for new datatypes; imaging will consume more storage than is practical; some data may need to be ignored
- Is there opportunity to align PROMs/wearables and NHS virtual wards activity?

First Draft Recommendations (3 of 5)

Guidelines on Privacy Enhancing Technologies (PETs)

- In collaboration with existing initiatives (e.g, ICO, UN, Royal Society/Turing) develop guidelines on the deployment of PETs alongside TREs
- Using learning from Phase 1, develop a risk model for the linkage of cross council data and provision of linked data to support guidelines on the usage of PETs
- Focused call to demonstrate effective use of PETs for federated analysis on sensitive data
- Develop training for research and technical teams on the effective use and deployment of selected PETs

Your thoughts and feedback

Guidelines on Privacy Enhancing Technologies (PETs)

- Need to make great use of high fidelity synthetic data
- Essential to public trust, but needs standardisation and to be explainable
- Link up with the UN activity, also Department for Education work on Functional Anonymisation

First Draft Recommendations (4 of 5)

Establish a UKRI wide metadata standard working group

- Survey each UKRI-council landscape on metadata usage for different data modalities and current approaches
- Define a minimally accepted metadata standard across all UKRI councils extending existing standards to define UKRI-council minimal metadata standards. Pilot the metadata standard
- Develop or select a reference implementation of a metadata catalogue that that can support the metadata specifications and use of DOIs
- Define a federated registry to hold a list of available catalogues, standards, vocabularies/terminologies

Your thoughts and feedback

Establish a UKRI-wide metadata standard working group

- Align with existing activity, e.g. CLOSER; don't invent a new standard
- Establishing terminologies and controller vocabularies across domains will be difficult
- Any network of catalogues should be federated/peer-to-peer and not centralised
- Minimal, but sufficient for discovery
- Led with a standards body, e.g. ISO
- Work with industry and initiative like GAIA-X
- Align with existing ontologies. https://www.ebi.ac.uk/ols/index
- Existing standards not considered to be sufficient

First Draft Recommendations (5 of 5)

Leverage existing Digital Object Identifier (DOI) minting services to provide persistent identifier for all UKRI Council resources at UKRI-wide and council levels

- Provide a central UKRI-council level service and guidance of UKRI data custodians
- Federate and syndicate large data custodians that already have DOIs assigned
- Review option to mandate that all UKRI-councils have all resource metadata post-investment registered either at the central UKRI level or at the UKRI-council level

Your thoughts and feedback

Leverage existing Digital Object Identifier (DOI) minting services

- Needs a service to scale up as each release of a dataset, tool, etc. will need a unique DOI
- How do we deal with costs of doing this?
- NERC Environmental Data Service already doing this for their archived data, probably also in other councils. Can we build from existing services?
- Needs to be a production service; avoid an academic solution
- Needs to be rich enough to provide technical details and distribution
- Fund the maintenance of metadata; *"Metadata is for life and not just for posting in a registry"*
- Maybe early enough in adoption to achieve a common approach without too much legacy

Where do we need your thoughts?

- What are your current key data discovery issues? How do you expect this to change over the next 2-3 years?
- How do you envisage your data preparation and provisioning changing over the next 2-3 years?
- What novel datatypes (e.g., IoT, audio, public provided) are emerging in your research domain?
- Does the UKRI need to mandate a requirement for metadata standards and DOI assignment?
- Should UKRI councils mandate the need for metadata catalogues across all research domains? Should these be delegated to data custodians or TRE operators who may have closest knowledge of the data?
- Do you think data assets generated from bespoke data linkage be captured in the metadata to reduce wasted effort and improve transparency?
- Should we just limit ourselves to datasets? Should we cover other metadata types – software/tools, facilities, projects, people, data use?

Your thoughts and feedback

Where do we need your thoughts?

- Not enough attention being paid to reproducibility. More training on FAIR principles
- Automating production of metadata is proving very challenging
- Need approaches to managing provenance and relationships between datasets and data elements

UK Research and Innovation

HDRUK Health Data Research UK

ADRUK Data-driven change

Thank you

Find out more about DARE UK: www.dareuk.org.uk