Stakeholders feedback from pre-meeting survey: Session 3 Sustainability & Collection

Technical workshop on real-world metadata for regulatory purposes
Virtual meeting, April 12, 2021

Presented by Prof. Miriam Sturkenboom, and Dr. Rosa Gini
University Medical Centre Utrecht and ARS Toscana

Survey question 3

Please share any comments or suggestions on the preliminary options for the sustainability and process for metadata collection.

Feedback that hopefully is clarified in presentations

- **Data entry and maintenance** –
  - Automated as far as possible
  - Roles and responsibilities (who will complete/update) will be clarified in the next months
  - Time of data entry: prior or after study?
  - Maintenance depending on funding model (standard data stewardship Budget?)
- **Adoption of available standards** (CDSIC, FHIR, DCMI / DCAT, bioschemas.org, FDO..)
  - **New standards** will be suggested, when they are lacking, e.g. data banks and prompts
- **Versioning**: will be adopted
- **Use case**: finding prior to, access and re-use after study

Points for discussion

- **Quality of information:** who verifies?
- **Relation with EU PAS Register**
- **Funding for basic meta-data updates**?
  - Incentives to share initial meta-data for discoverability
- **Access options** for different stakeholders
  - Free
  - Restricted?
- **Identifiers**
  - Institutes (Persistent uniform resource locator (PURL))?
  - Data banks?
  - Persons (ORCID)
  - Studies (EU PAS?)

Survey feedback items

Theme: Process

- **Who makes initial entries? Who updates?***
- What is the role of data banks and data sources?
- most real-world data are not with the medicines agencies, but with healthcare professionals/hospitals and payers organisations.
- **The data owners and data sources will be very diverse across the EU.** data sources may be either long-established, recent, small or very large and complex and are subject to widely different access conditions. Small data holders may have very limited resources to put into improving discoverability. EMA/HMA need to think about how data holders can be incentivized to improve discoverability, what will motivate them to do the extra work.

Theme: Process

- **Catalogue tool should support version control**.
- 'DS-quant descrip' - (1) in many cases, this information will immediately out of date. Does that matter? (2) Would suggest some 'versioning' to appear in that workbook - ie, stats listed are based on data extract from DDMMMYYY so there is visibility to #1.
- Broad accessibility key for success. The aim should be to maximize its users (academia/public sector/industry) so feedback can be provided and the information to be up-to-date.
- **Specify point of time of entry of information in the beginning of a study, or at the completion (e.g. study results).**
- Any form of automated data extraction should be considered to fill in "as much as possible" of the metadata given the data banks of a data source, also when updating the metadata
- New machine learning methods may help deal with some of those issues if data are collected in the real-world setting.

Theme: Study based collection & sustainability (I)

- **First approach [study based] sounds pragmatic** + storage and maintenance at the future DARWIN EMA office?
- Proposed approach [study based] seems reasonable to ensure data quality and sustainability of the catalog. However, resolution of differences at each entry critical elements for deliverable 6a
- I agree that the study **investigators will be in charge of completing and updating the catalogue throughout the study life cycle**. Nevertheless, there should be a **quality control system in place at European level**.
- **Relevant to 'data Source' and DAPs. If DAP has been awarded EEA/EU funding for research developing/providing/using the data source it could be encouraged as part of contract to keep metadata up to date**. Similarly if EMA contracted research e.g. ACCESS, CONSIGN etc, research contractors would be obliged to maintain metadata up to date

Theme: Study based collection & sustainability

- **Proposed approach seems reasonable to ensure data quality and sustainability of the catalog. However, resolution of differences at each entry critical elements for deliverable 6a /** Users would need to be able to view and compare multiple entries of metadata for the same data banks from multiple institutions – resolution of differences not addressed
- I agree that the study investigators will be in charge of completing and updating the catalogue throughout the study life cycle. Nevertheless, there should be a quality control system in place at European level.

Theme: Centralized collection & sustainability

- **Priority centrally coordinated metadata collection and maintenance process for all meta-data that is not study- or DAP-specific**. Study- /DAP-specific catalogues may only differ with respect to metadata that is specific to the study or the DAP. Information from existing catalogues should be leveraged.
- For existing data sources, the effort is maximum the first-time data are entered into the catalogue and then only updates would be needed, then with minimum time and effort. The data originator should be responsible for this collection as it is done already in the ENCePP resources database. A control of the recorded content with (yearly) reminder for updates could be handle by one independent and trusted nominated institution.

Theme: Standards & Quality

- **In order to have a sustainable process, it would be advisable to adopt and contribute to community standards** (e.g. DCMI / DCAT, bioschemas.org, FDO..), many options are named in the report already. This would both accelerate standards adoption in the community as well as make it easier and therefore more likely to get high quality and up to date information into the catalogue. Do not assume everyone will use the catalogue to enter data, but that project and companies will develop tools to do so.
- The catalog or list will be updated by adding use cases over time or by having a general update. Moreover, one should **encourage research that explicitly delves into testing the quality attributes** of different data over time as a third means. Experimentation and scholarly discourse can serve well to improve the use of such data, as well as to receive feedback into improved data-collection.

Theme: Standards & Quality

- EHDEN, FAIRplus & other projects believe in adoption through community standards. Joining and adding to existing communities is not easy but key to long term sustainability. We have documented our search for standards communities in EHDEN in D4.5, https://zenodo.org/record/4474373.
- Recommend revising the structure of the underlying datamodel and apply agreed best practices with respect to model normalization.
- Decouple development of (meta)data standards and implementation of the catalogue. In that way, the standard could be used in other catalogues and doesn't depend on implementation.
- Will variable quality issues be flagged?

Theme: Funding

- The suggestion that part of a budget is used to update the catalogue suggest that the catalogue needs to be updated per study and this is a lot of work?
- Funding the collection of the metadata is the best way to ensure its consistency.
- EMA/HMA need to think about how data holders can be incentivized to improve discoverability, what will motivate them to do the extra work.

Theme: Overlap

- Re Fig 2, if the overall goal of this heath data catalogue is to get “access to a standard and electronic set of complete and accurate metadata information …”, it is unclear why the “data use” should be part of it, given availability via **EU PAS registry**
- Section 2.4 unclear, with data catalogue apparently including both data sources and studies performed with primary data collection in scope; **Should catalogue include all studies performed with primary data collection or secondary use. Overlap with EU PAS registry?** Expectation of the re-use of data collected for a study for another? Non-native variables created through a study as specified are not “fed back” into the data sources used, but method applied to derive/transform data published.

Theme: Use Case

- What is the use case for the catalogue?” if it is to find relevant databases for the study than this already happened prior to the study initiation. What exactly is used during the study?*
- Clarify purpose of the catalogue, currently mixture of a database catalogue, competence catalogue, network catalogue, study catalogue, catalogue of data models etc.*
- Will successful regulatory use cases be made publicly?*
- Very valuable if more information can be made available on the actual content of the data source.
- How will interested parties 'external' to this process (eg, the pharmaceutical industry) be able to access the catalogue? Similarly, how will other entities (HTAs, regulatory authorities) use this catalog?
- to be able to better explain the topic to our colleagues we would need to have small, compact "real word" example with test data and a concrete use case.

Theme: Governance & Access

- How will NCAs be included in the relevant processes. How will the governance look like and which skills and efforts are expected?
- Industry should be viewed as a legitimate stakeholder and be given access to the meta-data repository, for transparency. Consideration should also be given to giving MAHs access to the future DARWIN, e.g., in order to fulfill post-marketing commitments within a mutually trusted setting.

Thank you!

For any question on this presentation, please contact: Malgorzata.Durka-Grabowska@ema.europa.eu
Official address Domenico Scarlattilaan 6 ● 1083 HS Amsterdam ● The Netherlands
Address for visits and deliveries Refer to www.ema.europa.eu/how-to-find-us
Send us a question Go to www.ema.europa.eu/contact **Telephone** +31 (0)88 781 6000