**RDA Reproducible Health Data Services Working Group**

**Purpose:** Case statement for application to RDA Working Group

<https://www.rd-alliance.org/working-and-interest-groups/case-statements.html>

**Reproducible Health Data Services WG Charter**

The goal of the working group is to improve the reuse of health data by providing recommendations for reproducible data curation and brokerage workflow services.

A large proportion of activities in Health data services facilitate the use and reuse of data in different contexts surrounding health care and health research. The data span across biomedical domains, including clinical, genomic, and personal health data repositories with the aim of reusing data in different contexts.

Examples of health data service stakeholders include: health care data curation centers, medical data services, clinical data integration centers, biostatistics and system medicine institutes, and other data centers who assimilate, manage, and distribute health data for various uses such as research, innovation, quality assurance and improvement, and efficiency monitoring.

The actors involved in data services perform many tasks such as data curation, mapping, integration, and publishing. These interdependent tasks build upon each other to create common workflows that transform siloed data into new, curated datasets, requiring the navigation of data interoperability, data quality, and data security. Thus, understanding these health services processes is vital to support reproducibility and ensure FAIR data practices.

The case statement outlines our work and provides the focus and the boundaries for the working group activities.

The following stakeholders will potentially benefit from our contribution:

* Data curators/brokers in their daily activities
* Data consumers (e.g., clinical researcher, application developers, innovators)
* Health research data repositories or archivists
* Health research funders

The benefits may include the ability to reuse processes, gain credit for work, provide transparency, and facilitate machine readable workflows.

The RDA Reproducible Data Services Working Group (i) will generate recommendation statements to identify, capture and store curation metadata, and (ii) will develop an adoption and training guide to improve the uptakes of our outputs.

**Value Proposition**

Biomedical data are valuable resources for multiple purposes beyond the original collection context. Yet, the data reside in distributed repositories in various forms (e.g., written reports, structured data, semi-structured data such as genomic tests, imaging). Additionally, due to privacy reasons and high barriers to communication with local systems, most biomedical data curation is handled via health data services. These services receive data requests and deliver the curated data set. While there might be internal mechanisms to record data provenance, there is no explicit, standardized method to describe and document the processes for collecting and preparing secondary data for reuse within the health sciences.

Processes such as finding, selecting, and integrating the data for a given research question requires a set of data curation activities including data access, query, extraction, transformation, cleaning, aggregation, and sharing. Each of these steps impacts the scope and coverage of the resulting curated data set.

For reproducible research, the research data curation workflow should be clearly documented, if possible in a machine interpretable way, and should be accessible beyond the lifetime of the data curation process. However, current documentation practices primarily stem from the goodwill of individuals.

Implementation of the WG recommendations will improve the capture and storage of salient metadata provenance - in a machine processable way wherever possible. Data consumers (researcher, innovators, etc.) will be able to access detailed data curation metadata together with the data itself. This documented and machine actionable metadata will enable reproducible research.

**Engagement with existing work in the area**:

This work will be directly associated with the **Health Data IG**. We will also collaborate with the following IG/WG to optimize our work:

* Working Group for Data Security and Trust (WGDST)
* WDS/RDA Assessment of Data fitness for Use WG
* RDA/CODATA Legal Interoperability IG
* RDA/NISO Privacy Implications of Research Data Sets IG
* Ethics and Social Aspects of Data IG
* PID Kernel Information WG
* Reproducibility IG

Outside of RDA:

* Non-Profit/NGO’s focused on biomedical data
* Academic medical institutions
* Industry

*Final Deliverables*

1. Recommendation Statement for Reproducible Health Data-Services:   
   Reviewing and documentation of existing standards which can potentially capture data curation provenance; identifying gaps within current health data services practices producing limitations in study reproducibility and transparency; recommendations for future standard development activities.
2. Adoption and Training Guide:   
   Document state-of-the-art methods and standards for clinical data curation; best practices for capturing and storing data curation metadata for reproducible research. The final recommendation statement will demonstrate protocols for documenting the data, materials, and processes essential for reproducing the collection, cleaning, assessment, and sharing of health data as executed within health data service centers.

documenting of state-of-the-art methods for clinical data curation for reproducible data services from two perspectives: 1) Data processing; and, 2) Data governance.

*Milestones and Intermediate Documents*

Documents will be created and made public through tools such as the Open Science Framework, Google docs, and GitHub. From the start of the WG, we will complete the following:

6 months Feedback on initial workflow draft:

Feedback will be collected through presentations, meetings, and workshops (potentially during RDA plenary 13 and 14) with data brokerage teams and clinical researchers who lead or participate with such teams, in essence the primary adoption audience. In addition, use case examples and feedback will be garnered through github commits and comments, similar to the maDMP common standards WG. Key feedback concerns will include the generalizability, granularity, and comprehensiveness of the proposed metadata standard, as well as any potential risks or barriers to adoption that ought to be overcome throughout development and testing. Feedback will be documented and adjudicated by members of the Health Data Service Workflows WG, edits will be made to the existing metadata templates, and metrics based upon these concerns will be developed in preparation for gap analysis and use case tests.

6 months Gap analysis completed and test cases will be identified:

Test use cases will ingest materials and data generated through completed or ongoing health data brokerage projects.

Metrics of success will include the following:

1. Completeness of data ingest within the CEDAR metadata database;
2. Ease of usability, gathered through interviews with teams participating in test cases;
3. Cleanliness of data held in CEDAR metadata database and the feasibility of extracting, transforming, and loading data captured in CEDAR into existing metadata repositories associated with project publications.

9 months Use case presented at RDA Plenary:

Presentations will take the form of working group session interactive talks, posters, and panels. Feedback from plenary group attendees will be adjudicated by WG team members and adapted within preparation for workflow completion and adoption.

12-18 months Complete workflow and prepare for future adoption:

*Mode and Frequency of Communication*

In addition to meeting at plenaries, we will have two or more formal calls in between the plenaries. Using on-line collaborative tools (e.g. Google docs, OSF) will allow for work and comments will also serve as a form of communication. Those individuals actively working on outputs will have ad-hoc meetings as needed (e.g., Skype). Trello and Github will be used for planning and tracking group deliverables.

*Develop Consensus*

The chairs and active members will work together in a small-group to achieve the goals. When there is draft outcome, this will be presented to the larger group through a publicized call for anyone to attend. Any conflicts will

* + A description of how the WG plans to develop consensus, address conflicts, stay on track and within scope, and move forward during operation, and
  + A description of the WG’s planned approach to broader community engagement and participation.

**Broader Community Engagement and Participation**

The developed adoption guideline will be discussed in different networks at Europe, America and Australia, including GoFAIR, German Medical Informatics Initiative, .

**Planned Activities**

**Review of the workflow components and related challenges**

* Define the processes of moving data through a clinical data service center and break down into a set of possible curation activities in a workflow.
* Identify challenges for each curation activity from the perspective of reproducible research.
* Identify the possible metatypes for each curation activity to trace the data provenance.

**Perform a Gap analysis to identify the supporting metadata standards:**

* Survey and map existing standards and recommendations supporting data provenance in each curation activity step.
* Map the curation steps with reproducibility assessment frameworks (such as RepeAT).
* Identify gaps and document suggestions for future standardization efforts.

**Adoption and Training Guideline:**

* First adoption will be implemented by Stanford CEDAR project. See the adoption plan below.
* Other adoption use cases will be explored both among group members (German Medical Informatics Initiative, GoFAIR, eResearch Services ...

**Adoption Plan:** A specific plan for adoption or implementation of the WG outcomes within the organizations and institutions represented by WG members, as well as plans for adoption more broadly within the community. Such adoption or implementation should start within the 12-18 month timeframe before the WG is complete.

Documentation of workflow best practices will be shared as a data dictionary of materials to be collected, stored, and shared throughout the data brokerage process and FAIR principles for each piece of materials. This data dictionary will be developed into templates within the CEDAR metadata registry tool, which will provide an interface for data entry, storage, and export, as well as a display of the existing metadata standards and ontologies mapped to each element within the Health Data Service Workflow. These CEDAR templates for metadata collection will be shared with all CEDAR users, as well as exported as JSON and RDF schema. In addition to sharing metadata collection templates through CEDAR, these templates will be hosted and shared on a project Github, Open Science Framework, and shared Google drive. An adoption guide will be created to assist adopters in the use of the metadata collection templates, as well as best practices associated with collecting, storing, and sharing each element within the Health Data Service Workflow. This adoption guide will also be made available within a project Github, Open Science Framework, and shared Google drive, and potentially disseminated in the form of a publication.

The primary audience for community output adoption includes project managers of clinical data warehouses, health data registries, and clinical research investigators/teams who regularly interact with clinical data brokers. Metrics of successful adoption include:

* Training of clinical data warehouse staff in reproducibility best practices using the disseminated adoption guide;
* Successful collection and ingest of project metadata satisfying the elements within the reproducible health data service workflow framework;
* Implementation and adaptation of adoption guide and/or framework into existing clinical data management and research methods education curriculum for research students or staff.

**Initial Membership:** Chairs and founding members

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | Member Type | RACI | Region/Country | Contact mail |
| Oya Beyan | Chair | R/A | Germany | beyan@dbis.rwth-aachen.de |
| Anthony Juehne | Co- Chair | R/A | US | aljuehne12@gmail.com |
| Ludovica Durst | Co- Chair | R/A | Italy | l.durst@lynkeus.com |
| Kate LeMay |  |  | Australia | kate.lemay@ands.org.au |
| Leslie McIntosh | Health Data IG Liaison |  | US | leslie.mcintosh@rda-foundation.org |
| Julie Toohey | Member, Health Data Librarian |  | Australia | julie.toohey@griffith.edu.au |
| Malcolm Wolski | Member |  | Australia | m.wolski@griffith.edu.au |
| Mark Musen | Member |  | US | musen@stanford.edu |
| Matthias Löbe | Member |  | Germany | matthias.loebe@imise.uni-leipzig.de |
| Henriette Senst | Member |  | Germany | sensth@rki.de |
| Gareth Knight | Member |  | UK | gareth.knight@lshtm.ac.uk |
| Rob Hooft | Interested, distant and relaying member | C/I | The Netherlands | rob.hooft@dtls.nl |
| Austin, Claire |  |  |  | claire.austin@canada.ca |
| John Borghi |  |  | Lane Library - Stanford Medicine | jborghi@stanford.edu |
| Gary Berg-Cross | Interested party | I | DC Area US | gbergcross@gmail.com |