RDA Reproducible Health Data Services Working Group
Case Statement: https://bit.ly/2PgkvJV
Slides: https://bit.ly/2Jd1pAP

Purpose: Case statement for application to RDA Working Group
https://www.rd-alliance.org/working-and-interest-groups/case-statements.html
https://bit.ly/2PgkvJV

Reproducible Health Data Services WG Charter

The goal of the working group is to enhance the reuse of health data for research and improve the FAIRness levels of aggregated and curated data sets for secondary use by providing recommendations for reproducible data curation and brokerage workflow services.

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 patient generated health data repositories

Examples of health data service stakeholders include: health 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 primary and secondary 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 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 pertaining to the collection, cleaning, and curation of health data for analysis and sharing.

The RDA Reproducible Data Services Working Group (i) will provide recommendations to identify, capture, and store metadata documenting workflows for collecting and curating health data for secondary reuse, 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, and 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 or clinical decision support pathway 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 processes generated for the need of each lab, department, or research project, with little attention paid to the interoperability of the final data with appropriate research repositories, or the capture of the entire research workflow. For example, in the case of pathology, paper and digital reports containing the interpretation of samples and data vary widely, including what information ought to be documented, what terminology is used for interpreting pathology status, the sectioning and order of information in the document, and the final data type of textual data stored in the medical record. Adjudicating these variances in clinical interpretations and data nuances is left to the researchers and data brokers collecting and preparing the data for re-use. These same variances and nuances witnessed in textual pathology reports can be observed in almost every type of clinical observation and health data, such as the clinical ontology used for documenting a disease, the manner in which lab tests are coded and timestamped, or whether drugs in a participant record are ordered, administered in hospital, or prescribed for home use. All of these distinctions matter when attempting to study and generalize findings to particular disease types, cohorts of patient sub-groups, the efficacy of particular drugs or treatment regimens, and etc. Thus, the Reproducible Health Data Services WG aims to generate a machine-readable method for documenting these data nuances and the manner in which they are curated and adjudicated within a final research data set.

Implementation of the WG recommendations will improve the capture and storage of salient metadata elements documenting - in a machine processable way wherever possible - data provenance and curation activities, thus contributing to the overall FAIRness of the data and project as a whole. The projected use case of the final deliverable will allow Data consumers (researchers, innovators, etc.) to access detailed data curation metadata together with the data itself. This documented and machine actionable metadata will enable reproducible research and improve data quality.

Multiple ongoing research initiatives across the biomedical sciences demonstrate the need for such a metadata standard for the documentation of data curation workflows.

For instance, within the United States, the National Academies of Science and Engineering have hosted multiple expert workshops aimed to define the best-practices of transparent reporting and appropriate stakeholder support and incentives to achieve reproducible workflows. The development, implementation, and testing of our working group's deliverables would allow for increased transparency, interoperability, and reproducibility across such government-funded projects.

Members of the EU lead Fair4Health and HL7 FHIR initiatives have also expressed the need for such a schema to document the data provenance, curation activities, and associated research workflow materials (including query scripts and code) for the clinical trial data they aim to merge within a central repository.

National projects such as Germany Medical Informatics Initiative (https://www.medizininformatik-initiative.de/en/about-initiative) creates data integration centers to aggregate research and healthcare data and share. Similar projects across the social sciences, agriculture, and humanities wherein collection and secondary use of data is relevant could benefit from lessons learned throughout the development and implementation of the deliverable developed within our group's work.

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
- Metadata Standards WG
- Metadata IG
- PID IG
- PID Kernel WG
- Data Foundations Terminology IG
- Research Data Provenance IG

In addition to collaborating with existing working and interest groups, we will leverage relationships with the RDA Secretariat and OAB to increase the representation and engagement

of perspectives from the community of potential adopters, particularly targeting data brokers, curators, and clinical data warehouse managers within academia, government, and industry.
Outside of RDA:

- Non-Profit/NGO's focused on biomedical data, including epidemiology, public, and global health groups, in addition to international groups focused upon increasing the reproducibility and transparency of secondary data use in the health sciences
- Academic and governmental medical institutions. Co-chairs and members of the working group have ties to academic and governmental medical institutions with large clinical data warehouses regularly used to support clinical research.
- Industry: solution providers for health care IT

Final Deliverables

1. <u>Recommendation Statement for Reproducible Health Data-Services:</u>
   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. <u>Adoption and Training Guide:</u>
   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.

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 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.

12 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 an institutional metadata database, capturing metadata describing complete project workflows.
2) Ease of usability, gathered through interviews with teams participating in test cases; The aim of this metric would be to provide an evaluation metric to support uptake by the targeted community of use.
3) Cleanliness of data held in institutional metadata databases and the feasibility of extracting, transforming, and loading data captured in the metadata repository into existing domain and publisher metadata repositories, thus providing further linkage to additional project metadata documentation within external repositories, such as NCBI, PubMed, or ClinicalTrials.gov

18 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 a 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 deliverables will be discussed in multiple networks across Europe, North America, South America, Africa, and Australia, including GoFAIR, German Medical Informatics Initiative, and meetings of HL7 FHIR working groups. Currently, a collaborative working group within HL7 FHIR is being developed utilizing multiple aspects of the Reproducible Health Data Services concepts and will provide a method for testing adoption of the WG deliverables.

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 data service 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. The projection is that testing within the CEDAR repository will be scalable to similar institutional metadata repositories across medical informatics cores and clinical data warehouses.
- Other adoption use cases will be explored both among group members. Stakeholders who have expressed interest in participating in such adoption include German Medical Informatics Initiative, GoFAIR, eResearch Services at multiple university medical informatics cores.

Adoption Plan:

Reproducible Health Data Services Metadata Model:

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 metadata schema 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. In addition to long standing working relation between members of the CEDAR team and co-chairs of the WG, the CEDAR platform provides integration with multiple data and metadata repositories across the health and biomedical sciences. The CEDAR platform centers on the use of metadata templates, which define the data elements needed to describe particular types of biomedical experiments. The templates include controlled terms and synonyms for specific data elements. CEDAR uses a library of such templates to help scientists submit annotated datasets to appropriate online data repositories enabling.

- Community-based organizations to collaborate to create metadata templates, investigators or curators to use the templates to define the metadata for individual experiments, and
- Scientists to search the metadata to access and analyze the corresponding online datasets.

These CEDAR templates for metadata collection will be shared with all CEDAR users, as well as exported as JSON and RDF schema for scalabile implementation within similar metadata repositories. 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.

Adoption Guide:

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 metadata about workflows generated by projects 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	Co-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	Interested		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	Interested			claire.austin@canada.ca
John Borghi	Interested		Lane Library - Stanford Medicine	jborghi@stanford.edu
Gary Berg-Cross	Interested party	I	DC Area US	gbergcross@gmail.com
Mário J Silva	Interested		Portugal Europe	mjs@inesc-id.pt
Mary Uhlmansiek	Member		US	muhlmansiek@wustl.edu
S. Venkataraman	Interested		UK	s.venkataraman@ed.ac.uk
Irene	Interested		Pt	ipr@uevora.pt

Rodirgues				
Carlos Luis Parra-Calderón	Interested (potential adopter)		Spain	carlos.parra.sspa@juntadeandalucia.es
Kiko Núñez	Interested (potential adopter)		Spain	kinube@gmail.com
Celia Álvarez	Interested (potential adopter)		Spain	celia.alvarez@juntadeandalucia.es
Serena Battaglia	Interested (potential use cases)		France	serena.battaglia@ecrin.org
Thu-Mai Christian	Interested (in application to other domains)		US	thumai@email.unc.edu
Christine Jacquemot	Interested		France	marie-christine.jacquemot@inist.fr