RDA-OfR Creating a Multi-omics Metadata Schema Standard Reporting Matrix WG

Note: The WG is actively seeking co-chairs and members from different countries and continents who cover various Omics domains, including expertise in mass spectrometry and sequencing. A dedicated communications campaign via the RDA's communications and social media channels will take place to recruit co-chairs and members from geographically diverse regions.

1. Overview

Multi-omics data integration merges multiple Omics data types (e.g., genomics, proteomics, metabolomics, phenomics, etc.), leveraging a wide range of high-throughput technologies as a holistic approach to quantification and characterisation of large complex pools of biological molecules. Multi-omics data integration and analysis provides a closer look into the complex structural and functional interactions at the molecular level for improving our understanding of biological dynamics of living organism(s) across the life science landscape.

Significant advancements continue to evolve high-throughput Omics technologies (sequencing, mass spectrometry, imaging, etc.), bridging a variety of subject matter expert methodologies and applications. As a result, there has been an unprecedented increase in the volume of multi-omics data generation and storage over the past decade. However, many challenges remain regarding multi-omics data management and sharing for reuse of complex Omics datasets (Figure 1).
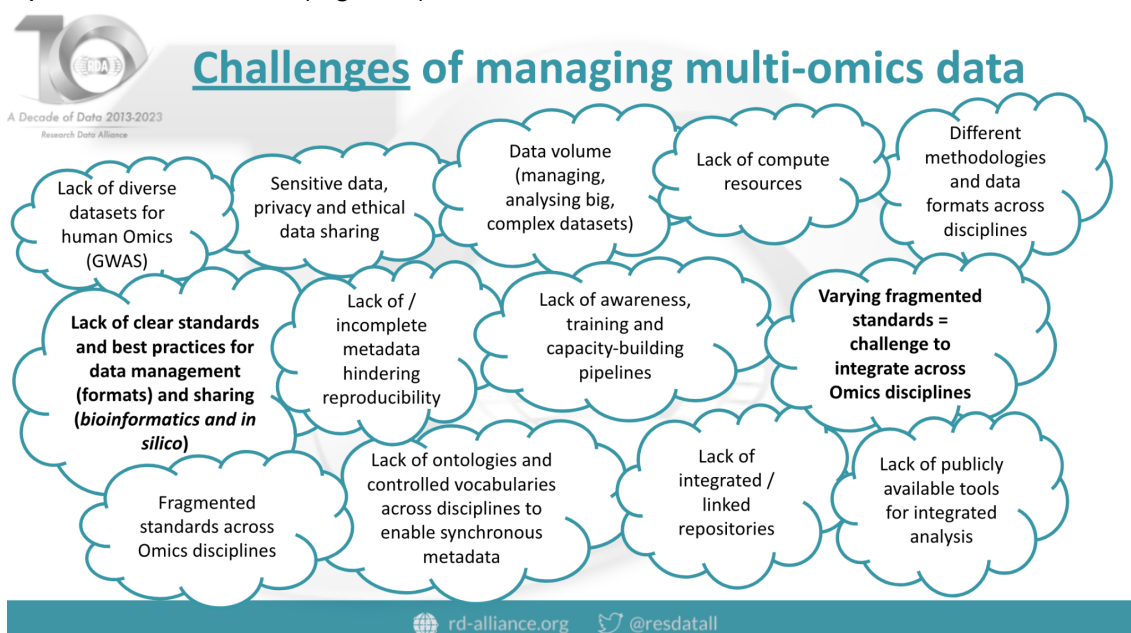

Figure 1. Challenges of effectively managing and integrating multi-omics data identified during the RDA-OfR WG brainstorming workshop in September 2023.

A persistent challenge is the vast array of fragmented experimental metadata and data formats generated across the different Omics domains, making it difficult to manage and integrate multi-omics data prior to downstream analyses.

This RDA working group (WG), supported by Oracle for Research (OfR), aims to address a few of these challenges by creating a matrix of identified reporting guidelines and standards essential for integration of multiple Omics metadata elements across the different domain technologies (Figure 2).

1. **Landscape review and collection of existing Omics community standards (Deliverable)**

The WG will research and consult existing work in the area by undertaking an in-depth landscape review to identify current Omics domain data standards (common data formats, controlled vocabularies and ontologies, metadata reporting guidelines, and identifier schemas) outlined by community accepted data management and sharing best practices within and across the different Omics domains. This work will leverage and build upon existing resource records at [FAIRsharing](#) (an [RDA WG](#) and a curated, informative and educational resource on data and metadata standards, inter-related to databases and data policies) in evaluating existing community Omics standard records in contribution to an iterative and open [Omics Domain Collection](#) at FAIRsharing as part of the landscape review recommendation output. Leveraging FAIRsharing [educational materials](#) to help guide WG curation activities, the group will collect standalone Omics domain [community standards](#) and reporting guidelines for informed downstream crosswalk activities. The Omics landscape review and analysis collection serves to benefit and encourage continuous (machine-actionable) community level standard curation beyond the lifecycle of the WG for existing and future research community stakeholder groups focused on data standards in life sciences.

2. **Omics community standard and reporting guideline crosswalk (Deliverable)**

There are currently many well-established and well-developed Omics standards, but knowing intuitively or immediately which standards are complementary across domains is not obvious. Based on the results of the Omics landscape review, this WG will provide a crosswalk that will identify common data standards and metadata reporting guidelines implemented across Omics domains to link complementary integration points. This activity aims to support common metadata elements outlined in the resulting standard reporting matrix and downstream data interpretation endeavours. This crosswalk will highlight domain metadata reporting gaps and areas where current standard implementations may not be in alignment across the various Omics standards where supplementation may be of use.

3. **Multi-omics metadata schema standards and reporting matrix plus use case collection (Deliverable)**

The multi-omics metadata schema standard reporting matrix, detailing the essential domain standards metadata elements required to accommodate multi-omics integration in areas of genomics, transcriptomics, proteomics, metabolomics (including imaging mass spectrometry). This guideline will be supported by multi-omics community example use

cases and curated Multi-omics Database Collection at FAIRsharing capturing use case records, if applicable/available. Documented use cases support diverse subject domain community examples of multi-omics standard integrations from existing cross-disciplinary group activities (societies, alliances, standard consortiums, and research projects focused on data standards harmonisation). Use case database collection will outline specific developments where multiple Omics standards have been successfully integrated for supporting multi-omics data platforms and/or metadata frameworks (Ex: National Microbiome Data Collaborative (NMDC), Multi-Omics Research Factory (MORF), etc.). Multi-omics data integration use cases with an included database record at FAIRsharing, identified during the landscape review, will be included in a new collection (as mentioned above) consisting of use-case multi-omics data repositories and/or knowledge bases implementing at least 2 different domain Omics data standards (reporting guidelines, controlled vocabularies and ontologies, etc.) reflecting key points of integration captured in the standard reporting matrix.
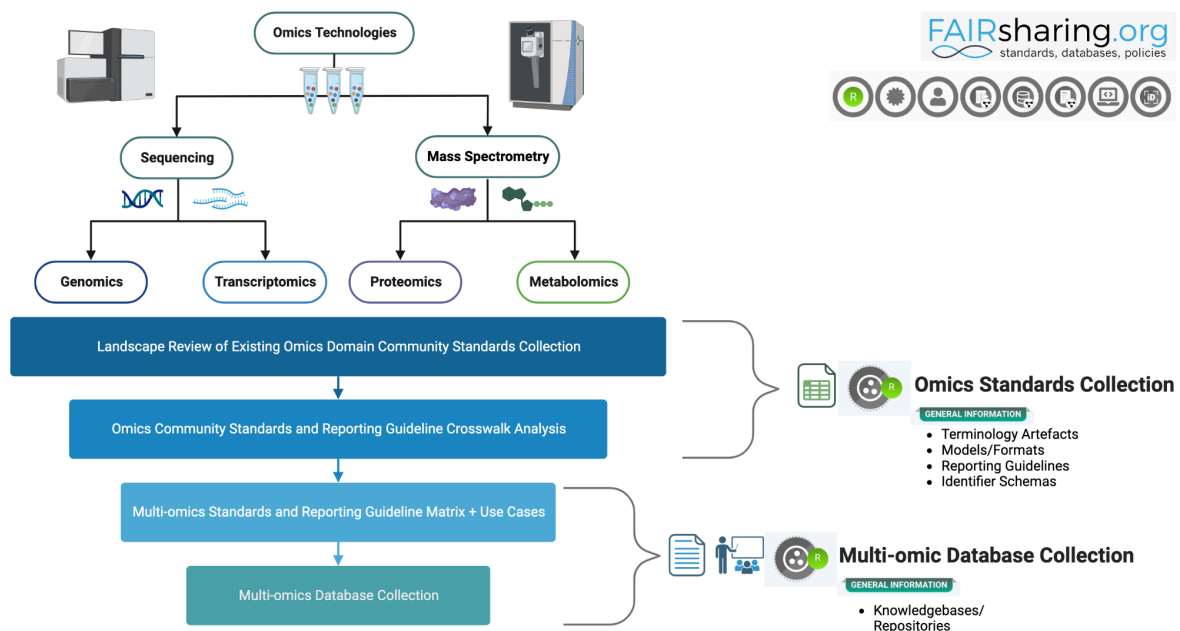
Created by 0000-0002-8741-7823 using BioRender.com

Figure 2. Omic domain technologies and corresponding core Omics working group deliverable outline.

2. Value Proposition

By leveraging preexisting Omics domain metadata standards, ontologies, and reporting guidelines, this working group will advance current siloed community harmonisation efforts with much needed cross-disciplinary diversity for creating more sustainable standard reporting guidelines fully representative of various data types and computational formats.

The deliverable produced by this WG aims to provide value and impact for the following adopters:

Adopter	Value/Impact
Researchers (data generators/users)	To advance awareness of existing and ongoing developments of multi-omics standard best practices for adoption within their research expertise.
Data Support Professionals (data/project managers)	To gain a better understanding of the multi-omics data landscape and improved knowledge of standards and best practices required to provide data management support for researchers integrating multi-omics data.
Developers & Data Curators (data consumers/generators)	To implement community-developed (meta)data guidelines, models, ontologies, schemas and formats for multi-omics for improved machine-actionable data discovery.
Research Organisations (performing institutions)	To implement and comply with policy stakeholder recommendations at an organisational and domain community level to promote standards and best practices in Omics/multi-omics research communities.
Publishers (service providers)	To make informed data sharing recommendations to authors, journal editors, and reviewers regarding appropriate metadata standards, data preservation, and policy best practices for multi-omics data.
Funders (policy stakeholders)	To make informed data policy recommendations to research performing institutions and projects developing multi-omics data management and sharing plans. Informed recommendations provide added guidance to publishing and reviewer stakeholder groups for evaluation of disseminated multi-omics data and computational analyses.

3. Engagement with existing work in the area

As mentioned in Section 1, the landscape review of Omics standards and best practices will entail thorough research and consultation of preceding and existing organisations and initiatives that develop standards for managing and integrating multi-omics data.

A primary resource of interest for this activity is FAIRSharing since it comprises several Omics standards and educational material for users on standards which will be harvested and used to create: (i) the crosswalk of common metadata standards linking Omics domains; and, (ii) multi-omics metadata schema standard reporting matrix. In turn, additional standards discovered during the landscape review will be collected and shared in the form of a new FAIRsharing Multi-omics database resource collection for ongoing community curation.

Other relevant Omics work in various areas

General Developments
- OBO Foundry Standard Ontology Identifier Policy; view FAIRsharing record
  - Example Ontology: EDAM; see term label example for "Omics"
- ISA Metadata Framework Model, instantiations in several formats, and tools; view FAIRsharing record
  - Example Community Standard Adoption(s): ISA-TAB Nano, NASA GenLab, Recommended ELIXIR Interoperability Platform
  - Example Training Material: Making Omics Data Matrix FAIR (FAIR Cookbook recipe)
- SciData examples of data modelling (updated by SJC)
- Linked (Open) Data Modeling Language (LinkML) (metadata schema)
- National Microbiome Data Collaborative (NMDC) (microbial microbiome use-case); view FAIRsharing record
  - Example Multi-Omics Metadata Portal: https://data.microbiomedata.org/
- The Environmental System Science Data Infrastructure for a Virtual Ecosystem
- (ESS-DIVE)
  - Example Identifier Schema: IGSN; view FAIRSharing record
- AgBioData (agricultural use-case); view FAIRsharing Community Collection
- ELIXIR multi-omics Genotype and Phenotype Core Resources
  - Example Service: X-Omics data analysis and integration stewardship

Relevant Literature
- Semantic concept schema of the linear mixed model of experimental observations (article)
- Ten quick tips for avoiding pitfalls in multi-omics data integration analyses (article)
- SODAR: managing multi-omics study data and metadata (article)
- Machine actionable metadata models (article)

Genomics
- GA4GH Variation Representation Specification; view FAIRsharing record
  - Example Standard Ontology: SO; view FAIRsharing record
  - Example Model Schema: VRS Information Model
- Genomic Standards Consortium (GSC); view FAIRsharing Collection
  - Example Metadata Standard: MIxS; view FAIRsharing record
- NHGRI-EBI GWAS Catalog (EMBL-EBI)
  - Example Standard Ontology: EFO; view FAIRsharing record
- Fair Header Reference (FHR) metadata structure framework for reference genomes

Proteomics
- The Proteomics Standards Initiative at Twenty Years: Current Activities and Future Work

- [HUPO Proteomics Standard Initiative (HUPO-PSI)](#); [view FAIRsharing Collection](#) (established domain standards)
  - Example Reporting Guidelines: [MIAPE](#); [view FAIRsharing record](#)
  - Example Reporting Guidelines: [MAGE-TAB](#); [view FAIRsharing record](#)
- [European Proteomics Association Standard Initiative (EuPA)](#)
  - Example Domain Collaborations: [Metaproteomics Initiative](#) (HUPO-PSI/EuPA)
- Example Project: [Critical Assessment of MetaProteome Investigation (CAMPI): a multi-laboratory comparison of established workflows](#)

Metabolomics
- [MEtabolomics standaRds Initiative in Toxicology (MERIT)](#)
- [Imaging Mass Spectrometry Society (IMSS)](#) (standards in development)
  - Example Standards: [Sample Reporting Survey](#)
- [The Metabolomics Society Initiative (MSI)](#) (established standards)
  - Related Activities: [ELIXIR Metabolomics Community](#), [Chemical Analysis Working Group (CAWG)](#)
  - Example Database: [Metabolights](#); [view FAIRsharing record](#)
    - Example Model Format: [ISA-TAB](#)
  - Example database: https://www.metabolomicsworkbench.org/
- [Metabolomics Association of North America (MANA)](#); Example [Standard Adoptions](#)

Ontologies and controlled vocabularies
- General Experimental
  - [Ontology of Biomedical Investigations (OBI)](#); [browse terms](#)
- Metabolomics and Proteomics
  - [MSIO](#); [browse terms](#)
  - [PSI-MI](#); [browse terms](#)
  - [CHEBI](#); [browse terms](#)
- Genomics
  - MIXS; [browse terms](#)
  - [SO](#); [browse terms](#)

RDA groups

- [FAIRsharing Registry: Connecting data policies, standards and databases RDA WG](#) (standards, databases, policies)
- [RDA COVID-19](#)
  - [RDA-COVID19-Omics](#) ([The final version of the RDA COVID-19 Recommendations and Guidelines for Data Sharing, published 30 June 2020](#)) (Omics sharing)
- [Research Metadata Schemas WG](#)
- [Life Science Data Infrastructures IG](#) (data lifecycles)
- [Persistent Identification of Instruments (PIDINST) WG](#) (instrument metadata schemas)
- [Data Repository Attributes WG (DRAWG)](#) (deliverable pending)
- [Chemistry Research Data Interest Group (CRDIG)](#)

4. UN Sustainable Development Goals (SDGs)

This WG contributes to several [United Nations Sustainable Development Goals (SDGs)](#).

Due to the revolutionary role of multi-omics in advancing our understanding of health and disease, this WG primarily contributes to [Goal 3: 'Good health and wellbeing - To ensure healthy lives and promote well-being for all at all ages'](#).

Owing to the broader role of multi-omics in the Life Sciences, the work of this group also contributes to:
- Goal 13: '[Climate action - Take urgent action to combat climate change and its impacts](#)' since [Omics research is used to predict climate change response](#).
- Goal 14: '[Life below water - Conserve and sustainably use the oceans, seas and marine resources for sustainable development](#)'
- Goal 15: '[Life on land: Protect, restore and promote sustainable use of terrestrial ecosystems, sustainably manage forests, combat desertification, and halt and reverse land degradation and halt biodiversity loss](#)'.

5. Work Plan

A work plan has been defined that facilitates an efficient and timely delivery of WG deliverables. Working Group members will meet virtually via Zoom (for max. 90 mins) monthly from November 2023. Tasks will be divided and allocated to task groups within the WG, and work undertaken by task groups in between meetings as required. Meetings may involve lightning updates from task groups and may include presentations from external speakers if relevant/applicable.

Month/Year	Preliminary WG Activities
September 2023	● First brainstorming workshop & publication of case statement ● Workshop slides, collaborative notes workshop 1 & workshop 2
October 2023	● Endorsement of case statement (Community, Council & TAB)
November 2023	● 1st WG meeting (WG kick-off meeting & member consultation, training material overview)
December 2023	● 2nd WG meeting (i. Presentation of WG aims, objectives, deliverables and timeline. ii. Allocation of task groups by domain expertise) ● Outreach (internal & external)
January 2024	● 3rd WG meeting (Landscape review) ● Outreach (internal & external)
February 2024	● 4th WG meeting (Landscape review) ● Outreach (internal & external)

March 2024	● 5th WG meeting (Landscape review) ● Definition of WG recommendations & outputs structure
April 2024	● 6th WG meeting (finalise landscape review - creation of Omics Domain collection in FAIRSharing mining efforts from task groups) ● Outreach (internal & external)
May 2024	● 7th WG meeting (crosswalk) ● Outreach (internal & external)
June 2024	● 8th WG meeting (crosswalk) ● Outreach (internal & external)
July 2024	● 9th WG meeting (finalise crosswalk and start collecting use cases). *This activity will help to identify the core metadata schema reporting matrix and invitation of use-case example representatives providing feedback on how well their resource captures elements from the matrix for more than one omics domain.* ● Outreach (internal & external)
August 2024	● 10th WG meeting (metadata schema standard reporting matrix and use cases) ● Outreach (internal & external)
September 2024	● 11th WG meeting (metadata schema standard reporting matrix and use cases) ● Outreach (internal & external)
October 2024	● 12th WG meeting (metadata schema standard reporting matrix) ● Final WG Recommendation Community review ● Outreach (internal & external)
November 2024	● Final WG Recommendation Endorsement (Council) & Press campaign

6. Adoption Plan

This work aligns with the RDA's mission to build the social and technical infrastructure to enable researchers and innovators to openly share and re-use data across technologies, disciplines, and countries.

For transparent and accessible collaboration, the WG will use a Google Folder for its documentation. Updates will be regularly posted to the WG wiki page summarising meetings and sharing important updates relating to WG progress and timelines. The WG will organise regular dissemination activities and solicit community feedback during specific phases of the project. Community consultation (e.g., calls to action, surveys) may be employed to identify different standards, ontologies, reporting guidelines, and best practices within different

Omics domains. The WG will also collaborate with organisations and initiatives that develop standards for managing and integrating multi-omics data, such as FAIRSharing.

It will be important to validate the: (i) Landscape review; (ii) Crosswalk of common metadata standards linking Omics domains; and, (iii) Metadata schema standard reporting matrix for multi-omics (Section 1) with the global multi-omics community (including researchers, data support professionals, research tool developers/providers, research performing organisations, publishers, funders, and policymakers) at various stages of the WG's lifecycle.

7. Initial Membership

The WG will represent international perspectives from a variety of stakeholders, including researchers, data support professionals, system/service providers, policymakers, publishers, and librarians. Following two brainstorming workshops held in September 2023, the WG comprises the following initial membership and leadership*:

No.	Name	Affiliation(s)	Country	Participation
1	Lindsey Anderson	Pacific Northwest National Laboratory / FAIRsharing Omics Champion	USA	Co-chair
2	David Molik	USDA ARS - ABADRU	USA	Co-chair
5	Christine Ballard	Oracle, Inc.	USA	Member
6	Francis P. Crawley	GCPA & SIDCER	Belgium	Member
7	Anupama Gururaj	NIH	USA	Member
8	Rob Hooft	Health-RI	Netherlands	Member
9	Flavio Licciulli	CNR - ITB	Italy	Member
10	Ugis Sarkans	EMBL-EBI	UK	Member
11	Elisha Wood-Charlson	Lawrence Berkeley National Laboratory / KBase	USA	Member
12	Stuart Chalk	University of North Florida / IUPAC	USA	Member
13	Maira Elahi	University of Windsor	Canada	Member

14	Chris Mungall	Lawrence Berkeley National Laboratory / ISB (Biocuration)	USA	Member
15	Natalie Meyers	University of Notre Dame	USA	Member
16	Adam Wright	OICR	Canada	Member

*Upon endorsement, the WG aims to recruit members from Asia-Pacific countries (East Asia, South Asia, Southeast Asia, and Oceania).