SF-RAD Data Standards and Informatics Infrastructure to Support Data Collection, Harmonization, and Sharing for Effective Surveillance of COVID-19 Outbreaks

[1]Dušica Vidović, [2]Daniel Cooper, [3]Kristina M. Babler, [3]Mark E. Sharkey, [4]Xue Yin, [4]Collette A. Thomas, [5]Benjamin B. Currall, [5]Sion L. Williams, [5]Natasha Schaefer Solle, [5]Naresh Kumar, [5]Melinda M. Boone, [6]Bhavarth S. Shukla, [7]Sreeharsha Venkatapuram, [7]Julio Perez, [7]Nakul Datar, [7]Christopher Mader, [7]Kenneth Goodman, [8]Krista Ryon, [5]George S. Grills, [6]Christopher E. Mason, [4]Helena M. Solo-Gabriele, [1,5,7]Stephan C. Schürer

[1]Department of Pharmacology, University of Miami, Miami, FL, [2]DataGrade Solutions, Miami, FL, [3]Center for AIDS Research, [4]Environmental Engineering Laboratory, [5]Sylvester Comprehensive Cancer Center, [6]UHealth, [7]Institute for Data Science and Computing, University of Miami, Miami, FL, [8]Weill Cornell Medicine, New York, NY

Abstract

Early detection of localized COVID-19 outbreaks has been facilitated through measurements of the RNA of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in wastewater from domestic sanitary sewer systems. South Florida Miami RADx-rad (SF-RAD) project was established with the goal to generate, optimize, standardize, and compare SARS-CoV-2 human and wastewater surveillance with various sampling, processing, detection, and analysis approaches, and to integrate wastewater data with community and hospital COVID-19 prevalence, with the aim of developing predictive models of local and community level spread of COVID-19.

In order to facilitate the SF-RAD data collection, harmonization and sharing, we have developed the data and metadata standards specifications for the various wastewater and human subject data. The internally defined specifications were harmonized and aligned to the RADx-rad and CDC data dictionaries, as well as the NIH Common Data Elements. The specifications were further formalized by utilizing the controlled vocabularies, numeric ranges, along with other information. They were stored in a local PostgreSQL database that enabled automated processes to generate the submission forms and data types' queries. These formalized representations were used in the OpenSpecimen to facilitate the submission and validation of the data. The collected data were further processed and visualized in the SF-RAD Dashboard.

FAIR Metadata Standards

Wastewater Data

The wastewater samples were collected at multiple locations at the University of Miami campuses corresponding to the student residential dormitories and a hospital that treats COVID-19 patients. These samples underwent further processing such as viral control spiking, concertation by different methods (such as magnetic beads, electronegative filtration), RNA extraction, and different water measurements.

Based on the processing steps, the wastewater data was organized into two major data categories: (i) wastewater sampling and quality information such as Field Data, Water Quality measurements, Bacterial Culture analysis, and sample processing data such as Pretreatment, Concentration, and RNA extraction, and (ii) the wastewater analysis results produced by PCR and RNA-seq assays to measure the viral load in the sampled wastewater. For each data category, the variables describing the data were developed in collaboration with the corresponding groups and individuals.

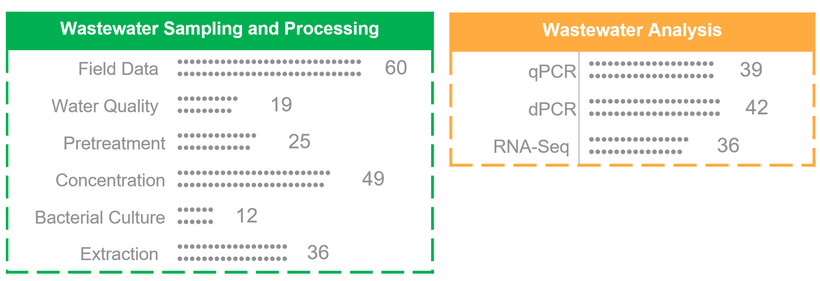

Figure 1. Wastewater data categories and counts for developed metadata standards specifications.

Human Subject Data

Human subject data collected under the SF-RAD project consists of three main categories of data: (i) COVID patients' information collected by the hospital, (ii) student testing and vaccination records on aggregated and individual person level, and (iii) UM counts of positive personnel and students.

The metadata standards specifications used to describe human subject data were adopted from the NIH Common Data Elements and expanded internally with parameters that were considered important for prediction models.

In order to make the human subject data compliant with the HIPAA privacy rule, we have implemented de-identification methods to ensure that any identifiable patient information is removed from the data before it can be shared with the community. The methods include statistical analysis and data processing (primarily data aggregation), and removal of personal identifiers as defined by the Safe Harbor.

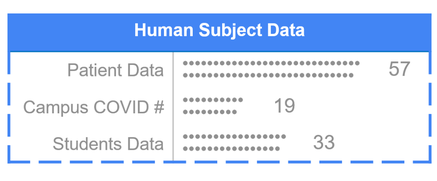

Figure 2. Human subject data categories and counts for developed metadata standards specifications.

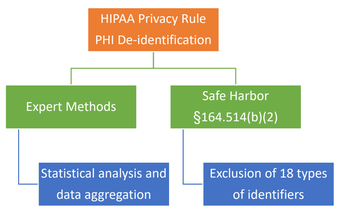

Figure 3. PHI de-identification methods.

Integration with Communities

The metadata standards specifications developed by other groups such as CDC, and groups of RADx-RAD program, were integrated into SF-RAD specifications.

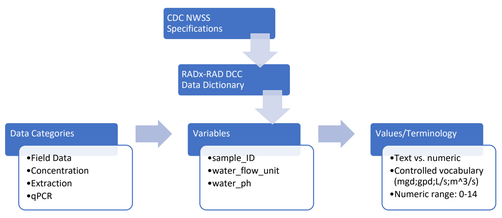

Figure 4. The metadata standards development integration with other communities.

SF-RAD Field	DCC Field	NWSS Field	SRA Field
sample_id	sample_id	sample_name	sample_name
sampling_date	sample_collect_date	collection_date	collection_date
sample_collection_method	sample_type	ww_sample_type	ww_sample_type
population_served	population_served	ww_population	ww_population
sample_collection_team	collected_by	collected_by	collected_by
institution_type	institution_type	ww_sample_site	ww_sample_site
sample_id	sample_id	ww_surv_system_sample_id	ww_surv_system_sample_id
water_flow_rate	flow_rate	ww_flow	ww_flow
water_temp	collection_water_temp	ww_temperature	ww_temperature
sample_tss	tss	ww_total_suspended_solids	ww_total_suspended_solids
water_ph	ph	ww_ph	ww_ph
water_salinity	ww_sample_salinity	ww_sample_salinity	ww_sample_salinity

Figure 5. Example for mapping variables between the SF-RAD and RADx-RAD, NWSS, and SRA values. The mapping enables efficient data processing and submission to these communities.

Informatics Infrastructure

Data Processing

In order to enable automated data processing, validation, and sharing, we have developed and implemented a robust operational infrastructure that integrates multiple databases, applications, API calls, and scripts. The data

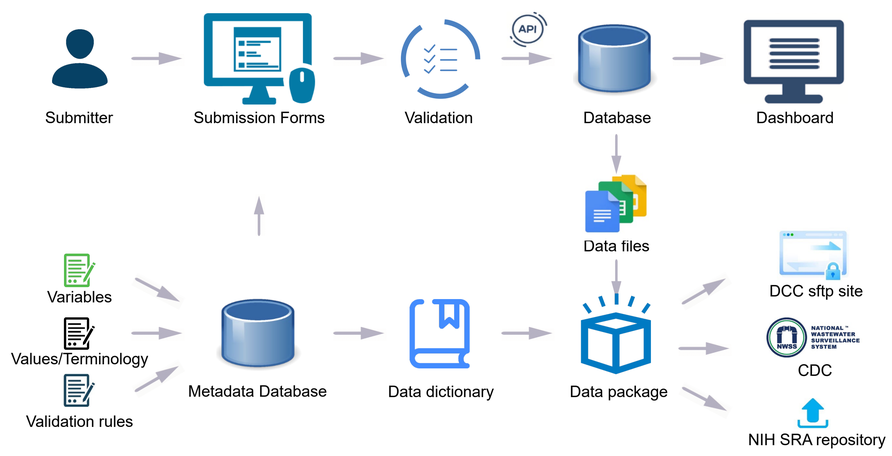

Figure 6. Implemented informatics infrastructure to support data collection, validation, and sharing.

Metadata Management

The metadata variables, values, terminologies, and rules are managed by centrally storing them into a PostgreSQL database. The data is formally described by JSON schemas and the corresponding submission forms and metadata templates can be generated from the information stored in the database.

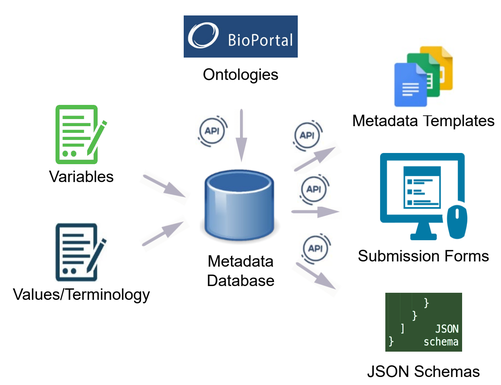

Figure 7. Database used for metadata management.

Data Submission and Validation

Data submission and validation is managed through the OpenSpecimen platform. The wastewater samples and their derivatives are registered, and the corresponding data associated with them.

For each data category, the submission forms are generated from the metadata database in a semi-automated way. The data validation rules are built into the submission forms, and only valid data can be uploaded into the OpenSpecimen.

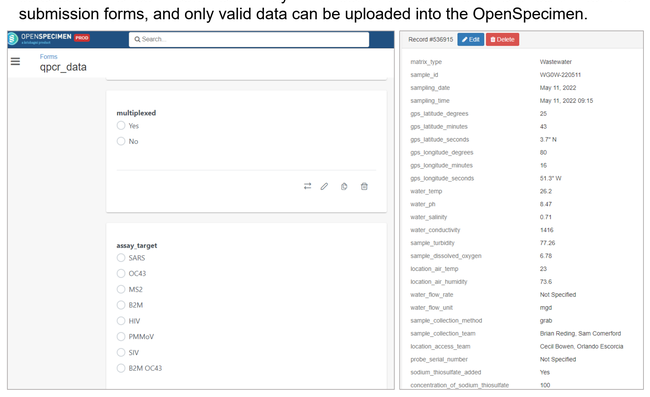

Figure 8. An OpenSpecimen form with implemented validation rules and a valid submitted data record.

Data Sharing

Collected and validated SF-RAD is exported from OpenSpecimen and packaged with the corresponding data dictionaries to comply with the submission requirements to the RADx-RAD DCC, CDC NWSS, and NIH SRA repository.

The variables' mapping enables a fast data conversion between different entities.

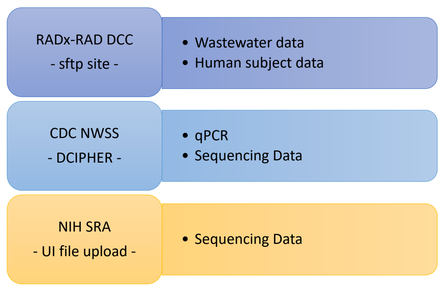

Figure 9. Data types and mechanisms to share the SF-RAD data with different communities.

SF-RAD Dashboard

SF-RAD data is visualized at the Dashboard. The site is available publicly at https://sfrad.idsc.miami.edu.

The visualization includes time series for the viral load in the wastewater, time series for the water quality measurements, and prediction models based on the viral load and positive-tested individuals contributing to the corresponding wastewater sampling location.

Additionally, the website integrates the geographic information system (GIS) via a set of sampling location GPS coordinates projected into the basemap (here Web Mercator projection into Stamen map) and associated with the corresponding data through the sampling location names. The GIS information is available for sampling stations, sewer lines, and building footprints.

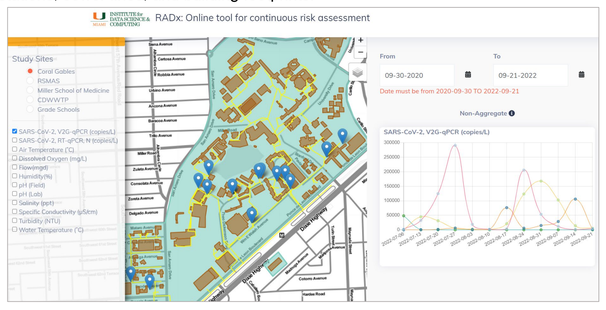

Figure 10. Time series and GIS visualization in the SF-RAD Dashboard.

Acknowledgements: This project is supported by the National Institutes of Health (NIH) Rapid Acceleration of Diagnostics-Radical (RADx-rad) grant 1U01DA053941-01; the NIH/NCI Cancer Center Support Grant 1P30CA240139-01 for the Sylvester Comprehensive Cancer Center; NIH grant P30AI073961 for the Miami Center for AIDS Research (CFAR); and NIH grant 1UL1TR000460 for the University of Miami Clinical and Translational Science Institute (CTSI). This project is also supported by University of Miami internal funding for wastewater surveillance at the university.

Correspondence: dvidovic@med.miami.edu; sschurer@med.miami.edu*