BUREAU OF RECLAMATION

Improving California-Great Basin Region Program's Water Quality Data Management to Enhance User Access, Analysis and Decision-Support

Science and Technology Program
Research and Development Office
Final Report ST-2021-7124-01

Delta Mendota Canal at milepost 64 near Santa Nella, CA (photo credit Laurel Dodgen)

U.S. Department of the Interior

September 2021

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE ***(DD-MM-YYYY)***
09-30-2021

2. REPORT TYPE
Research

3. DATES COVERED *(From - To)*
03-01-2017 – 09-30-2021

4. TITLE AND SUBTITLE
Improving California-Great Basin Region Program's water quality data management to enhance user access, analysis and decision-support

5a. CONTRACT NUMBER
19XR0680A1 RY.15417124.WD20811

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
1541 (S&T)

6. AUTHOR(S)
Laurel Dodgen, Richard Stack, Tiago S Oliveira

5d. PROJECT NUMBER
Final Report ST-2021-7124-01

5e. TASK NUMBER
7124

5f. WORK UNIT NUMBER
CGB-150

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
U.S. Department of the Interior
Bureau of Reclamation
California-Great Basin Region
Division of Environmental Affairs (CGB-150)
Information Management Services (CGB-3100)
2800 Cottage Way, Sacramento CA 95825-1898

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Science and Technology Program
Research and Development Office
Bureau of Reclamation
U.S. Department of the Interior
Denver Federal Center
PO Box 25007, Denver, CO 80225-0007

10. SPONSOR/MONITOR'S ACRONYM(S)
Reclamation

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
Final Report ST-2021-7124-01

12. DISTRIBUTION/AVAILABILITY STATEMENT
Final Report may be downloaded from https://www.usbr.gov/research/projects/index.html

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Following Executive Order 2013-13642 covering open data policy, the Department of the Interior – Bureau of Reclamation (DOI-BOR) has been developing efforts to build/modernize information systems to maximize interoperability and information accessibility, maintain internal and external data asset inventories, enhance information safeguards, and clarify information management responsibilities. In this context, the S&T Project 7124 intended to create and enhance DOI-BOR California-Great Basin Region capabilities to better integrate the BOR-wide data flow plans specifically related with data/documentation storage, database development/enhancement and data/documentation availability. Project 7124 efforts resulted in the development of a database schema, user requirements and reporting requirements. The new database infrastructure allows the incorporation of environmental data having in consideration data integrity and traceability. This report summarizes the activities developed.

15. SUBJECT TERMS
Data management, Open data, Database infrastructure, Water quality

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**THIS PAGE**	**U**		Laurel Dodgen
U	U	U			**19b. TELEPHONE NUMBER** *(Include area code)* **916-978-5038**

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

Mission Statements

The Department of the Interior (DOI) conserves and manages the Nation's natural resources and cultural heritage for the benefit and enjoyment of the American people, provides scientific and other information about natural resources and natural hazards to address societal challenges and create opportunities for the American people, and honors the Nation's trust responsibilities or special commitments to American Indians, Alaska Natives, and affiliated island communities to help them prosper.

The mission of the Bureau of Reclamation is to manage, develop, and protect water and related resources in an environmentally and economically sound manner in the interest of the American public.

Disclaimer

Information in this report may not be used for advertising or promotional purposes. The data and findings should not be construed as an endorsement of any product or firm by the Bureau of Reclamation, Department of Interior, or Federal Government. The products evaluated in the report were evaluated for purposes specific to the Bureau of Reclamation mission. Reclamation gives no warranties or guarantees, expressed or implied, for the products evaluated in this report, including merchantability or fitness for a particular purpose.

Acknowledgements

The Science and Technology Program, Bureau of Reclamation, sponsored this research.

Improving California-Great Basin Region Program's Water Quality Data Management to Enhance User Access, Analysis and Decision-Support

Final Report ST-2021-7124-01

prepared by

Laurel Dodgen, Physical Scientist, Division of Environmental Affairs

Richard Stack, IT Specialist, Enterprise Operations Division

Tiago Oliveira, Supervisory Physical Scientist, Division of Environmental Affairs (former)

Peer Review

Bureau of Reclamation
Research and Development Office
Science and Technology Program

Final Report ST-2021-7124-01

Report Title

LAUREL DODGEN

Digitally signed by LAUREL DODGEN
Date: 2021.09.15 13:14:28 -07'00'

Prepared by: Laurel Dodgen
Physical Scientist, Division of Environmental Affairs, Region 10

VANESSA KING

Digitally signed by VANESSA KING
Date: 2021.09.15 12:02:46 -07'00'

Peer Review by: Vanessa King
Hydrologist, Division of Planning, Region 10

"This information is distributed solely for the purpose of pre-dissemination peer review under applicable information quality guidelines. It has not been formally disseminated by the Bureau of Reclamation. It does not represent and should not be construed to represent Reclamation's determination or policy."

Acronyms and Abbreviations

CAS	Chemical Abstract Service
CGB-150	Division of Environmental Affairs
CGB-157	Environmental Monitoring and Assessment Branch
CGB-3100	Information Management Services
EDD	Electronic Data Deliverable
EFR	Electronic Field Record
GUI	Graphical user interface
QA	Quality Assurance
QC	Quality Control
Reclamation	Bureau of Reclamation
UTM	Universal Transverse Mercator

Contents

Page

Mission Statements ..... **iii**
Disclaimer ..... **iii**
Acknowledgements ..... **iii**
Peer Review ..... **v**
Acronyms and Abbreviations ..... **vi**
Contents ..... **vii**
Executive Summary ..... **ix**
1.Introduction **1**
2.Methods **2**
2.1 Current Database Review 2
2.2 Operational Review 4
2.3 Schema Development 5
2.4 User Requirements Development 5
2.5 Electronic Data Upload Development 5
2.5.1 Electronic Field Record (EFR) Development 5
2.5.2 Electronic Data Deliverable (EDD) Requirements Development 6
2.6 Data Reports Development 6
2.7 Database Validation 7
3.Results **7**
3.1 Schema 7
3.2 User Requirements 8
3.3 Electronic Data Upload 9
3.3.1 Electronic Field Record (EFR) 9
3.3.2 Electronic Data Deliverable (EDD) 11
3.3.3 Quality Control Sample Documentation 11
3.4 Data Reports 11
3.5 Data Validation Progress 12
4.Data 12
4.1 File Location and Size 12
4.2 Software 13
4.3 Point of Contact 13
4.4 Keywords 13
5.Discussion **13**
5.1 Project Planning 13
5.1.1 Schema Scope 13
5.1.2 Quality Control Sample Documentation 14
5.1.3 Start-To-Finish Data Control 14
5.2 Staff Utilization and Engagement 14
5.3 Project Outcome 15
References **16**

Executive Summary

The Environmental Monitoring and Assessment Branch (CGB-157) is a fee-for-service group in the Division of Environmental Affairs, Region 10, Bureau of Reclamation. Since formation, CGB-157 has engaged in environmental monitoring efforts across California, Nevada, and Oregon to provide clients with data for effective, informed management. This monitoring has generated over a half-million data records, including physical, inorganic, and organic analytes in water, sediment, soil, and biota. However, due to lack of upgrades over time, CGB-157's current database meets very few current standards for data management. It also does not support metadata and formatting that are required to upload data records to the Reclamation Information Sharing Environment (RISE). CGB-157 leadership determined that building an entirely new database would 1) strengthen data integrity and oversight, 2) resolve architecture issues that hindered incremental improvements in data management, and 3) align the CGB-157 data rules with RISE requirements to facilitate future uploads to RISE.

From March 2017 through July 2021, this project has spent approximately 6500 staff hours on the development and validation of a new database to meet operational needs and meet RISE data standards. This report includes a discussion of how to successfully undertake a new database design and build. The new database is expected to provide quantitative and qualitative benefits over a multi-decade useful life. When fully implemented, the new database is expected to save about 1200 staff hours each year through automatic data uploads. It will also provide benefits in improved data integrity and usability, helping CGB-157 more quickly provide valuable environmental data and assessments to aid environmental management decisions. Moreover, by facilitating data upload to RISE, the database will provide CGB-157's high-quality environmental data to other users and contribute to the overall mission of Reclamation.

1.Introduction

The Environmental Monitoring and Assessment Branch (CGB-157) is a fee-for-service group in the Division of Environmental Affairs, Region 10, Bureau of Reclamation. It originated in the 1980s as part of the remediation effort for Kesterson Reservoir, a former national wildlife refuge that experienced waterfowl and fish deaths in the late twentieth century due to unsuitable water being delivered via the partially constructed San Luis Drain. Remediation efforts removed selenium-contaminated wildlife habitat, and cost Reclamation around $21 million. The unfortunate and expensive event proved the value of environmental monitoring data in both planning and performing environmental management activities.

Since formation, CGB-157 has engaged in environmental monitoring efforts across California, Nevada, and Oregon to provide clients with data for effective, informed management. This monitoring has generated over a half-million data records, including physical, inorganic, and organic analytes in water, sediment, soil, and biota. Data initially was managed as paper records, and then transitioned to manual entry into a Microsoft Access database in the early 2000s. No database or data management upgrades have been made since then, despite significant improvements in the fields of information technology and data management and the ability to perform electronic upload of data.

Current best practices for storing electronic data proscribe a relational database whose architecture meets design rules. The rules preserve data integrity and data oversight by controlling data entry and modification, minimizing data variation and duplication, improving data identification and usability, and describing a robust database structure that can be easily updated. However, CGB-157's Access database meets very few current standards for best practice. Any user can enter and modify data, most database fields are not controlled either in data type or in content, and users cannot easily discover and retrieve data. Also, the lack of good architecture means that the database cannot be incrementally improved over time to meet new CGB-157 operational needs.

The growing need to overhaul CGB-157's database due to operational requirements was enhanced by increasing government support, and eventual mandate, of open data efforts [2-7]. Open data is typically electronic, machine-readable, in a standard and non-proprietary format, and published to the public to the extent security guidelines allow. OMB Memorandum M13-13 in 2013 emphasized the need to treat information as a valuable national resource and as a strategic asset to increase operational efficiencies, reduce costs, improve services, support mission needs, safeguard personal information, and increase public access to valuable government information [5]. In 2014, the Department of the Interior (DOI) released the Open Data Policy [8]. This policy prompted many groups in Reclamation, including CGB-157, to prioritize data management efforts in order to comply with DOI's plan. These efforts were further supported by the passage of the OPEN Government Data Act [7] in 2019 which built on the OMB Memorandum M13-13 [5] and made key aspects of the Open Data Policy [8] permanent. The OPEN Government Data Act created minimum standards for making Federal Government data available to the public and required the Federal Government to use open data to improve decision making.

One of the programs that originated in response to DOI's open data policy was the Reclamation Information Sharing Environment (RISE). RISE is a data-sharing platform built with the intention to act as a central source of Reclamation data, storing a variety of data from operating information to environmental monitoring. Individual data creators within Reclamation, such as CGB-157, are encouraged to upload their data to RISE to make it discoverable and accessible to other Reclamation groups, external stakeholders, and members of the public who might use the data to facilitate decision-making. Uploading data to RISE also allows data creators a path to comply with DOI's open data policy. However, to upload to RISE, data creators must provide a variety of supporting metadata and re-format their records to meet RISE's requirements. For CGB-157, the required metadata (such as sampling station elevation) and required formatting are not supported by the current Access database. Any records chosen to upload to RISE would need to be supplemented and formatted manually, costing considerable staff resources.

CGB-157 leadership determined that building an entirely new database would 1) strengthen data integrity and oversight, 2) resolve architecture issues that hindered incremental improvements in data management, and 3) align the CGB-157 data rules with RISE requirements to facilitate future uploads to RISE. Overall, the improved database would make CGB-157's valuable environmental monitoring data more discoverable and usable to clients, Reclamation colleagues, and the general public.

2.Methods

2.1 Current Database Review

The CGB-157 Access database was built in the early 2000s when the capabilities and expectations for electronic databases were much simpler and more limited. It was built to have data manually entered, and that practice continues currently. It contains the tables listed in Table 1. Few of the tables were connected by relationships, leading to a lack of constraints. For instance, users can enter any value in the table 'tblSample_Result' field 'Method' because that field is not actually constrained to the approved values in the table 'tblMethod'. This leads to duplicate methods being created and incorrect methods being entered. Inside of a table, many reasonable constraints are missing. For example, users are not required to enter a Chemical Abstract Service (CAS) unique chemical identifier number in the table 'tblAnalyte' field 'CAS'. This lapse directly led to many organic analytes being entered multiple times under slightly different names, since the CAS could not be checked for duplicates. A third type of problem is also apparent. While the database had been created with tables to store quality control data from CGB-157 and the contract laboratories, these tables were not used, likely due to the added burden of manual entry.

Table 1. All tables in current Access database.

Table Name	Intended Use	Table Content	Issues

tblProject_Info	Project information	Code, name, description, and manager for a project	No data constraints in any field. Table does not constrain sample results.
tblAnalytes	Analyte information	Type, class, name, CAS, alternate name, active flag, association with projects	Duplicate analytes. CAS, type, and class not required or constrained. Associations with projects abandoned after first project.
tblStation_ID	Station information	Code, description, water type, project code, UTM, area office location	Duplicate stations. UTM not required. Field named 'WATER_SOURCE' but non-water stations collected.
tblLaboratory_Info	Contract laboratory information	Name, description, representative contact details, active flag, last audit date and remarks	Table info largely duplicated in stand-alone spreadsheet so table rarely updated and is incomplete and out of date.
tblMethod	Method information	Code, description, active flag	Duplicate entries. Table does not constrain sample results.
tblX_Ref_Anl_Meth	Associates analytes with valid methods for their analysis	Analyte name, method name	Table does not constrain sample results.
tblFld_Smpl_Info	Sample event information	Sample ID, station, collection date and time, other sample details, analyte results measured in field	Table contains sample results measured in field instead of placing them in tblSample_Result. Field results don't have units. Most fields (such as sampler name) not constrained and have duplicates.
tblSample_Result	Sample result information	Sample ID, analyte, CAS, USBR Batch, result	'USBR_BTH_ID' not mandatory. Many fields (such as units) are not

		details, lab details, QA details, verification flag	constrained and have duplicates.
tblUSBR_QA_BTH	Lists sample IDs in each QA batch	Batch ID, batch preparer and date, sample ID details	Inconsistent entry to fields. Not linked to sample results after abandonment of tblSample_Result field USBR_BTH_ID.
tbl_USBR_QA_Results	QA results for a QA batch	Batch ID, analyte, CAS, method, QA details	Inconsistent entry to fields. Not linked to sample results after abandonment of tblSample_Result field USBR_BTH_ID.
tbl_Lab_QB_BTH	Lists lab QC batch number	Lab batch ID, lab name, lab prep date	Abandoned shortly after database launch. Few lab QC batches recorded.
tbl_Lab_QC_Results	Lab QC results for a lab QC batch	Lab batch ID, analyte, CAS, method, lab QC details	Abandoned shortly after database launch. Few lab QC results recorded.

CAS- Chemical Abstract Service unique identifier number. UTM – Universal Transverse Mercator coordinate system. QA – Quality Assurance. QC – Quality Control.

2.2 Operational Review

After assessing the content and functionality of the current database, CGB-157's operational processes and needs were reviewed by staff members representing each core function to determine other processes or data that were needed or desirable in a new database. Several improvements were identified:

1. New database must support electronic upload of field data and laboratory results. Moving from manual entry to electronic upload would preserve data integrity by minimizing transcription errors and circumventing paperwork damage. It would also save approximately 1200 work hours each year since staff would spend seconds performing an upload instead of 15-180 minutes manually entering data on each field and lab document. Types of data that would need to be uploaded include field logbooks, QA processing books, and lab reports.
2. Ensure new database has the capacity to store all data necessary to comply with RISE upload requirements. Review showed that many fields were missing, including station elevation, station type, and sample type.
3. Develop a suite of reports to improve data usability. Current reporting options were limited and caused clients and users to perform multiple queries and extensive data manipulation to meet their needs.

4. Track additional metadata. Some types of metadata were desired to support data validity, such as equipment used to collect samples and QC standards incorporated in samples. Other metadata was desired to manage CGB-157 operations, such as associating consumable materials with specific projects for cost allocation and tracking staff workloads. A last type of metadata was needed to meet RISE data standards.
5. All QC data must be captured and associated with relevant sample results to ensure the legal defensibility of the environmental data and improve the data's usability. QC data is generated by both CGB-157 and by contract laboratories.

2.3 Schema Development

After identifying the data elements that needed to be captured in the new database, CGB-157 worked with Information Management Services (CGB-3100) to develop the database structure. Database structure is commonly described by a schema that lists each table in the database, the fields in the table, and the relationships among the tables. Similar to an architectural blueprint, a schema describes the structure of the database to facilitate the database's initial build and continuing maintenance. Collaborative work was needed to ensure the new database would meet current data management best practices while also meeting CGB-157 operational needs. During the schema development, a small team from CGB-157 met regularly with one member of CGB-3100 to discuss needs and find solutions. This small team interviewed CGB-157 staff to ensure that the details of all CGB-157 operations were evaluated. During the database build, regular emails and monthly meetings were used to solve new problems and communicate about progress.

2.4 User Requirements Development

The next project stage was developing a User Requirements document. This document guides the database developer in building the graphical user interface (GUI) by describing what features the GUI needs and how those features look and behave. A team of three CGB-157 staff members was selected to represent the diversity of database users. Each team member prepared separate parts of the User Requirements document according to their subject area, while working concurrently in an online document. The initial draft of the User Requirements was then reviewed by a schema team member to fill in technical details and standardize the database behavior among the sections. Lastly, the database developer reviewed the document to verify that specified operations were technologically possible and advisable.

2.5 Electronic Data Upload Development

A key feature of the new environmental database is the electronic upload of data. Three data upload types were eventually identified: field sampler data, laboratory result data, and QC sample documentation.

2.5.1 Electronic Field Record (EFR) Development

A variety of data is collected by samplers in the field and several efforts were undertaken to develop an upload process. Initial efforts focused on sourcing a ready-made product, but several commercial

platforms were evaluated and were not acceptable. Some were not allowed under Reclamation policy due to cloud features; others did not have all the data record features needed. It was determined that a bespoke instrument should be developed.

An experienced field sampler in CGB-157 led the effort to develop this instrument and termed it an Electronic Field Record (EFR). This staff member evaluated current paper documents used to record field data and interviewed samplers at CGB-157 to identify all data elements for inclusion. Then they designed a workbook in Microsoft Excel that would capture all data elements while balancing ease-of-use in the field. Many Excel features, such as data validation controls, formulas, and look-up tables, were used to improve data integrity. Look-up tables were created based on the approved content of database tables. For instance, TBL_WEATHER lists 6 approved values for temperature, 5 choices for sky, 6 for precipitation, and 4 for wind. These values were put in a table in Excel and EFR weather selections were constrained in a drop-down selection box to these approved values.

After initial development, the EFR developer, schema team member, and database developer went through several rounds of review to ensure that the EFR satisfied technical requirements and would interface correctly with the database. Then the EFR developer and schema team member developed a guidance document. Once finalized, the EFR was customized by each investigation lead to the needs of each active investigation by removing irrelevant data entry options and updating names and analytes. Each field sampler has government-furnished equipment, in the form of both laptops and iPads, previously issued and could immediately use finalized EFRs in the field.

2.5.2 Electronic Data Deliverable (EDD) Requirements Development

Contract analytical laboratories have historically produced paper, and later PDF, lab reports. The industry has now moved to a standard of also including an electronic data deliverable (EDD) that supplies laboratory results in a machine-readable format. A team of three members, representing the samplers, QA reviewers, and data management groups developed requirements for an EDD. To preserve data quality and usefulness, EDD requirements include look-up tables, lists of allowed values for particular fields. The team determined which data elements were mandatory and the allowed values in the look-up tables from a review of the database schema. A schema team member reviewed the draft to supply technical details and a data team member supplied content for the look-up tables. Then the EDD requirements were provided to all contracted analytical laboratories, and the schema team member worked with the labs to create compliant EDDs.

2.6 Data Reports Development

A small team of CGB-157 staff was selected to design the data report requirements, composed of a staff member who analyzes environmental data to write investigation reports, a schema team member, and the project lead. Together they evaluated the various data user groups, the types of data they use, and the best format for retrieving and viewing that data. Ultimately a suite of data reports were described in an Excel workbook using tables and mock-ups. The database developer referenced this document while building the reporting functions of the database.

2.7 Database Validation

The completed database must be validated to ensure that it was populated with the correct background functional data, that it records and retrieves data correctly, that the GUI functions operate properly and allows the user to perform the intended task, and that the data reports retrieve the intended data in the specified format. A Microsoft Excel online document was set up to describe each validation task, the assigned person to perform the validation task, progress notes, and a centralized log to track any problems found during validation. The log was evaluated by a member of the schema team to determine if the identified problem needed to be fixed, and whether it should be fixed by the database developer or by schema team member. Problems for the database developer were logged as Issues in GitLab. Problems for the schema team member were fixed through the database GUI.

3.Results

3.1 Schema

The schema for the database was drafted during thirteen months of development by the schema team with input from CGB-157 staff. Key groups in CGB-157 reviewed the final draft over several more months before finalization. The final schema is composed of over 70 highly connected tables encompassing over 600 fields. Figure 1 graphically displays a small portion of the schema as an entity relationship diagram, highlighting the connections to field sampling activity and sample result.

The backbone of the schema is five tables (TBL_ACTIVITY, TBL_SMPL_SPIKE_INFO, TBL_SMPL_RESULT, TBL_LAB_QA_RESULT, TBL_USBR_QA_RESULT) which encompass the core of CGB-157's environmental monitoring efforts from field sampling activities through QC sample incorporation and lab analysis to QA evaluation. Many tables support this central workflow with metadata. For example, TBL_USBR_QC_BATCH records associations between sample results so that a user can understand which sample results are affected by a QA determination. A second example is TBL_STATION, which records location and description details for sampling sites. Other database tables are solely supporting, such as TBL_UNIT which allows the database to constrain entry of sample result units to approved values. Database tables are managed by authorized users, removing the previous problem of unauthorized creation of incorrect or duplicate values.

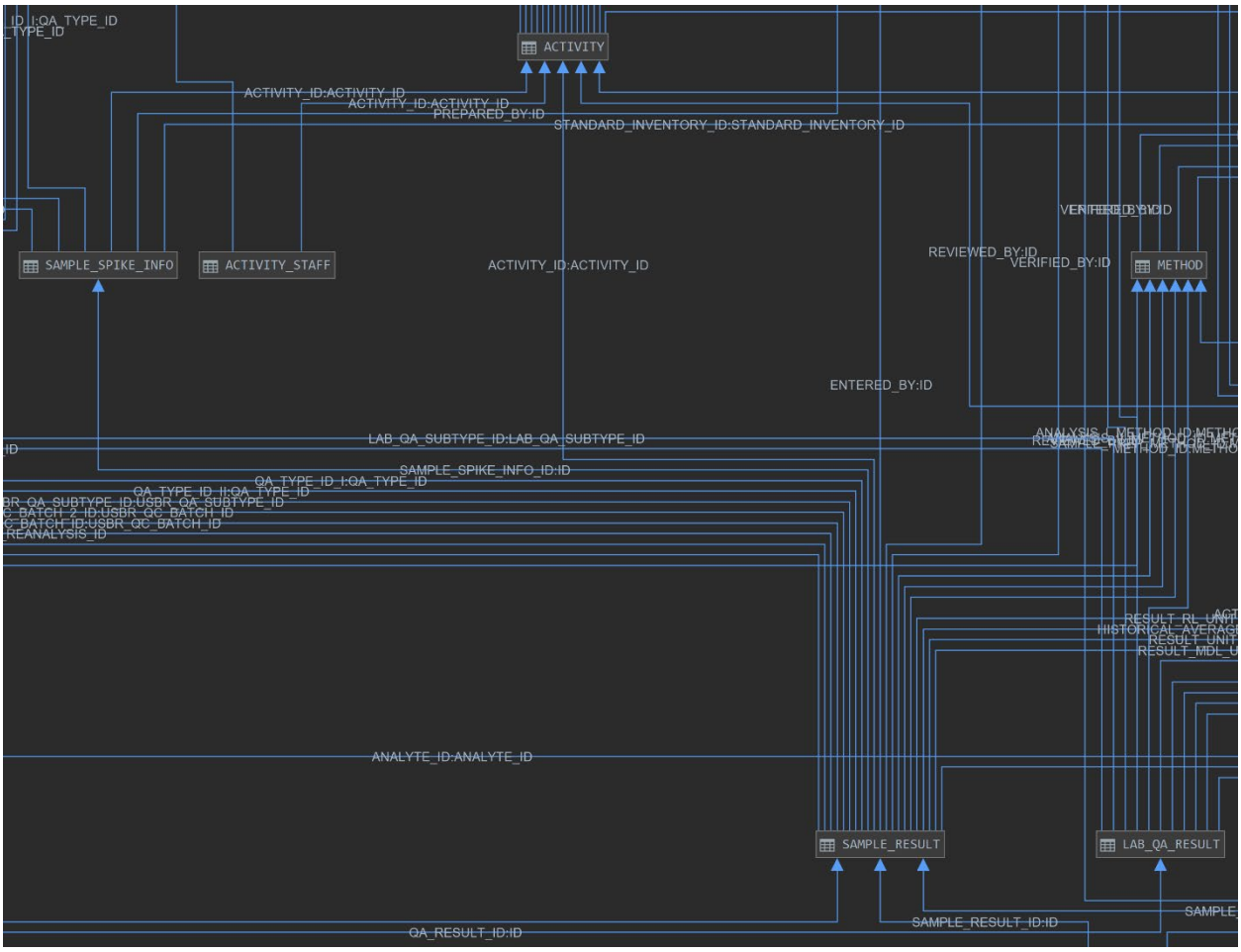

Figure 1. A small portion of the entity relationship diagram for the database schema, highlighting the connection to field activity and sample result. Gray boxes represent database tables. Blue lines depict relationships between fields in separate tables.

3.2 User Requirements

A major feature of the new database is its user-facing graphical user interface (GUI). The main headings of the GUI menu are listed in Figure 2. The GUI allows users to perform a variety of tasks including uploading data, manually entering data, approving or rejecting data, updating supporting information such as equipment and staff, and retrieving data reports. The User Requirements document that describes the appearance and functionality of the GUI was developed over four months by a three-member team of CGB-157 staff and then reviewed and finalized over three months by the CGB-157 supervisor and database developer. The final document is over a hundred pages long. The document describes processes in CGB-157's workflow. Each process is described in a step-wise fashion, and includes all necessary details such as database fields being queried or filled and the appearance of the GUI.

Figure 2. Menu options displayed in GUI of database. Arrows open sub-menus.

3.3 Electronic Data Upload

3.3.1 Electronic Field Record (EFR)

The EFR is used to record field sampling activities and format the data for database upload. It was developed over fourteen months by an experienced CGB-157 sampler and then reviewed and revised for another four months by the EFR author and a member of the schema team. During the final review, a companion guidance was written by both staff members to instruct field samplers how to fill in the EFR and customize it to specific sampling investigations. Each investigation lead used the guidance to adapt the generic EFR for their specific investigation details, such as by removing irrelevant columns and including the actual stations and analytes involved in the investigation. A 3-month test period was used by the samplers, where the original documentation process was used concurrently with the EFR. This test period was crucial to train samplers on correct documentation in the EFR. Findings from the test period also recommended a number of EFR changes, including which equipment details to record, new options for weather observations, and a change to how samples were noted in QA batches.

The EFR format is composed of one Excel workbook per sample event. Several worksheets in the book contain human-readable templates where samplers enter field data. For example, each sample station has a worksheet where the sampler can record information such as arrival time, equipment

used, field readings measured, and departure time (Figure 3). Some information, such as arrival date

	A	B
1		
2	Station Name	Station A
3	On Station Date (mm/dd/yy)	01/00/00
4	On Station Time	
5	Station Weather Temp (Select from list)	
6	Station Weather Sky Cover (Select from list)	
7	Station Weather Precipitation (Select from list)	
8	Station Weather Wind (Select from list)	
9	Collection Date (mm/dd/yy)	
10	Collection Time	
11	Equipment Calibration Date (mm/dd/yy)	01/00/00
12	Equipment ID 1 (e.g. Hand Held)	0
13	Equipment ID 2 (e.g. Sonde)	0
14	Equipment ID 3	0
15	Equipment ID 4	0
16	Equipment ID 5	0
17	Equipment ID 6	0
18	Equipment ID 7	0
19	Equipment ID 8	0
20	Temp (°C)	
21	DO (mg/L)	
22	DO (%)	
23	EC (us/cm)	
24	pH (units)	
25	Turbidity (NTU)	

Figure 3. Top view of a station worksheet from an Electronic Field Record workbook.

("On Station Date") autofills from other EFR worksheets to save sampler effort. The several human-readable worksheets are used to autofill one computer-readable worksheet. In that worksheet, records are organized to match the database ingestion process. Before a field activity, a sampler prepares the EFR and saves it to Microsoft Teams. In the field, a sampler fills in the EFR

by accessing Microsoft Teams if there is network service or a pre-loaded EFR as is common in remote stations. After the sampler completes sampling, the EFR goes through a review process to ensure data quality. Then the sampler navigates to the database GUI, select the menu item for field record upload, and select the reviewed EFR. The database ingests it, asks the sampler to approve the ingested records, and the records automatically populate the database. The finalized and uploaded EFR is archived in a Microsoft Teams folder and any further edits to it occur through CGB-157's document control process to ensure record integrity and accountability.

3.3.2 Electronic Data Deliverable (EDD)

The EDD is produced by an analytical laboratory so that sample results can easily upload to the database. The requirements were developed over five months, and then reviewed for a further month. At that point, look-up tables were added and reviewed over a further two months. The EDD requirements include 47 fields from 4 database tables and state the EDD must be provided as a tab-separated variable text file. While some fields are not relevant to unusual analyses, labs are required to provide data for all relevant fields. As contract laboratories came under new blanket purchase agreements that required EDDs, they were asked to produce an EDD for each lab report that met the requirements provided by CGB-157. This process required several rounds of review and conversation as laboratories developed a product that met specifications. In the process, the EDD requirements were edited for clarity and updated as labs found unusual elements in their own process that weren't covered by the initial EDD requirements (such as an analytical surrogate not in the EDD Analyte look-up table).

3.3.3 Quality Control Sample Documentation

Quality control samples are incorporated with environmental samples to demonstrate data quality. However, the upload of QC sample data was identified too late in the project to address during the project duration, and will be developed subsequently by relying on the robustness of the new database architecture to add that feature.

3.4 Data Reports

The suite of data reports was designed by two CGB-157 staff members over three months, with input from other data users. Each report specification contains a top table that lists general requirements such as type of data pulled and who can request the report, a middle table that shows database fields used to construct the report, and a bottom table that gives an example of the report layout. Figure 4 depicts this content for a report that summarizes lab delivery of analytical reports. Altogether, 28 data reports were developed to meet the needs of different types of users. Reports cover simple and detailed analytical results, simple and detailed QA summaries, field and QA comments, staff workflow rate, and more, allowing users to easily retrieve the desired information. As CGB-157 operations continue, new report options can be created by collaborating with CGB-3100.

Report should show all available data (including all QA_TYPES
Report available to all CGB-157 users
Date range selected by user
Data sorted by LAB_NAME then reverse chronological order

HEADER	COMMENT	
LAB_NAME	from TBL_SMPL_RESULT via LAB_REPORT_ID	lab samples sent to
METHOD_CODE	from TBL_SMPL_RESULT via METHOD_ID	method analyzed by lab
LAB_SUBMISSION_DATE	from TBL_SMPL_RESULT via LAB_REPORT_ID	date samples sent
LAB_REPORT_EXPECTED_DATE	from TBL_SMPL_RESULT via LAB_REPORT_ID	expected lab return
LAB_REPORT_NUM	from TBL_SMPL_RESULT via LAB_REPORT_ID	lab report number
LAB_REPORT_RECEIVED_DATE	from TBL_SMPL_RESULT via LAB_REPORT_ID	actual lab return
USBR_QC_BATCH_NUM	from TBL_SMPL_RESULT	QA batch IDs sent

EXAMPLE

Laboratory	Method	Submission Date	Expected Date	Received Date	Report Number	USBR QC Batch

Figure 4. Example of a data report specification.

3.5 Data Validation Progress

Database functions and data handling are being validated to ensure that it operates as intended and preserves data quality. Three areas of validation were identified: the content of supporting tables (such as TBL_UNIT and TBL_ANALYTE), the processes occurring in the GUI (such as uploading an EFR and validating a lab report), and the content of data reports. Overall, 105 processes were specified for validation and validation began in April 2021. As of July 2021, 30 processes have been fully validated, 177 problems have been recorded in the log, 4 problems have been corrected by CGB-157, 47 GitLab Issues have been filed for the database developer to address, and 30 Issues have been resolved. Database validation is expected to be complete in March 2022, and then CGB-157 will begin incorporating the new database into branch operations.

4.Data

4.1 File Location and Size

Data from this project is located on Reclamation servers at "\\ibr2mpfs002.bor.doi.net\MPRFS002\DEA\2017-S&T Project\Deliverables". The schema document is 2.85 MB. The User Requirements document is 161 KB. The EFR template is 351 KB. The EDD requirements is 161 KB. The EFR guidance is 220 KB. The data reports document is 895 KB. The data validation document is 225 KB. The total folder is 4.78 MB.

4.2 Software

The database is in Oracle RDBMS. Its platform is PHP 7 and its framework is in Symfony. The GUI is in KendoUI.

4.3 Point of Contact

Questions about this project can be directed to Laurel Dodgen (LDodgen@usbr.gov, 916-978-5038), Richard Stack (RStack@usbr.gov, 916-978-5494), or Dan Deeds (DDeeds@usbr.gov, 916-978-4467).

4.4 Keywords

Database; schema; data integrity; Open Data; RISE

5.Discussion

5.1 Project Planning

A project of this duration, with critical implications for the success of future CGB-157 operations, needed thorough planning. While a project scope and calendar were included in the project proposal, a more detailed plan was needed to successfully direct the granular elements. For example, the initial proposal stated in Task 3 that User Requirements would be developed to support the development of the database GUI. The actual plan to successfully execute this sub-task involved multiple steps, including identifying core users, forming a team from the users, identifying mandatory document content and style guide, assigning team members specific content, and determining a review system for the document. Most project tasks and sub-tasks were well-planned, with designated teams handling specific tasks using collaborative document tools. However, some elements were insufficiently planned or completely neglected, leading to overall increased project duration. A more comprehensive and detailed project plan should have been developed at project initiation.

Three critical elements were neglected in the initial planning: schema scope, electronic upload of QC sample documentation, and start-to-finish data control.

5.1.1 Schema Scope

The initial project plan identified a sub-task to update the database schema in order to make CGB-157's data meet operational needs and be compatible with RISE data standards. However, the schema updates were a major task, considering the starting condition of CGB-157's database and the numerous gaps in meeting RISE standards. Once the project team realized that this schema update was a substantial task, the team ought to have stopped development to plan schema scope. This

situation led to the schema update scope growing in size as more features were added and greatly added to development time. Had there been an initial project plan to guide the schema scope and update process, some features would have been wait-listed for future database updates and other features planned in a more holistic way. Overall, it would have allowed the project to complete sooner and more efficiently.

5.1.2 Quality Control Sample Documentation

Early in the project it became apparent that to serve CGB-157 operations and facilitate data import to RISE, data entering the database would need to be automatically uploaded from standardized files. The process to standardize field activity data and laboratory result data in machine-readable documents was reasonably planned and successful. However, close to project completion a gap was discovered: no process had been planned to allow upload of data related to QC samples. This oversight will be corrected in a future database upgrade, and until then this data will need to be manually entered through the database GUI. Had this gap been identified by performing an early, comprehensive planning process, this need could have been addressed in concert with the other standardized files.

5.1.3 Start-To-Finish Data Control

Central to CGB-157's mission is producing robust, legally-defensible environmental monitoring data for supporting management decisions and meeting contractual obligations. One element of robust data is having control of the data from generation through reporting, and being able to demonstrate that control. In CGB-157's previous database system, field samplers generated paper documentation that was reviewed and archived. The data was transcribed from the paper to the database and a second person reviewed the transcription. When data was used for reporting, a trained staff member reviewed the data before releasing it to clients. Data control could be demonstrated from generation to reporting, except for a major gap in that no control existed for users changing data in the database.

This gap was addressed in the new database, but other gaps developed and needed to be addressed, such as the preservation and archival of primary field records in the form of EFR documents. Ideally a comprehensive data control plan would have been developed as part of the project initiation plan. It should have reviewed the planned data process in the new database system and determined how to preserve document control through each step. Then, as the project plan was updated the document control plan would be revisited as well. During the project, data control measures were addressed in isolation by each team lead as each project part was developed. While this approach worked during the initial and middle stages of the project, it showed its deficiencies near project completion when it became apparent that an overall data control system was not ready. This situation led to late-stage efforts to sew together the individual document control measures.

5.2 Staff Utilization and Engagement

This project involved most aspects of CGB-157's operations, including sample collection, data management, QA processing and assessment, and data reporting. Project teams were composed of a diverse staff to represent the various stakeholders and provide subject-specific expertise. This approach was very successful in making a well-rounded product that would meet CGB-157 operational needs. It also became a useful staff development experience. Staff members were given

opportunities to grow into new roles and tasks, which sometimes required coaching to stretch their skill sets beyond their initial comfort. Careful management was needed to align project tasks with staff members' existing skills and interests. Since many tasks were far different than usual staff activities, staff were sometimes reluctant to start assigned tasks or complete them by due dates. Staff engagement became critical in encouraging staff members to complete assigned project tasks. During this long project, reminding staff of the end goal and acknowledging incremental achievements were helpful in maintaining engagement and momentum, especially when verbally supported by CGB-157's branch chief. Overall, the project benefited greatly by the appointment of one specific staff member who had scientific and data management expertise to champion the project and direct project efforts.

5.3 Project Outcome

From March 2017 through July 2021, CGB-157 and CGB-3100 have spent approximately 6500 staff hours on the development and validation of a new database to meet operational needs and meet RISE data standards. This project has been a considerable investment in future CGB-157 operations, requiring extra work effort on top of typical CGB-157 workloads. The new database is expected to provide quantitative and qualitative benefits over a multi-decade useful life. When fully implemented, the new database is expected to save about 1200 staff hours each year through uploading data (instead of manually entering). It will also provide benefits in improved data integrity and usability, helping CGB-157 more quickly provide valuable environmental data and assessments to aid environmental management decisions. Moreover, by facilitating data upload to RISE, the database will provide CGB-157's high-quality environmental data to other users and contribute to the overall mission of Reclamation.

References

[1] Krishnamurthy R., Awazu Y. (2016). Liberating data for public value: The case of Data.gov. *International Journal of Information Management*, 36, 668-672.
[2] Memorandum - Transparency and Open Government. (2009). *M-09-12*.
[3] Memorandum - The Open Government Directive. (2009). *OMB M-10-06*.
[4] Executive Order - Making Open and Machine Readable the New Default for Government Information. (2013). *Executive Order 2013-05-09*.
[5] Memorandum - Open Data Policy – Managing Information as an asset. (2013). *OMB M-13-13*.
[6] Digital Accountability and Transparency Act. (2014). *Public Law 113-101*.
[7] Open, Public, Electronic, and Necessary (OPEN) Government Data Act. (2019). *Public Law 115-435*.
[8] Open Government Plan 3.0. (2014). *U.S. Department of the Interior*.
[9] Blodgett D., Read E., Lucido J., Slawecki T., Young D. (2016). An Analysis of Water Data Systems to Inform the Open Water Data Initiative. *Journal of the American Water Resources Association*. 845 – 858.