BUREAU OF RECLAMATION

Leverage Existing Environmental Data for Improved Usability by Standardization and Migration to RISE-Compatible Database

Science and Technology Program
Research and Development Office
Final Report ST-2022-19210-01

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

U.S. Department of the Interior

September 2022

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
09-30-2022	Research	10-01-2018 – 09-30-2022

4. TITLE AND SUBTITLE
Leverage Existing Environmental Data for Improved Usability by Standardization and Migration to RISE-Compatible Database

5a. CONTRACT NUMBER
22XR0680A1 RY.15412019.WD19210

5b. GRANT NUMBER

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

6. AUTHOR(S)
Laurel Dodgen, Ph.D.

5d. PROJECT NUMBER
Final Report ST-2022-19210-01

5e. TASK NUMBER
19210

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)
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-2022-19210-01

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

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The Environmental Monitoring and Assessment Branch (CGB-157) recently built a new database to improve data integrity and facilitate data upload to Reclamation Information Sharing Environment (RISE). This project built on that effort; CGB-157's historical data was standardized, supplemented, and migrated to the new database for improved data accessibility and usefulness. A project process was developed that preserved data quality and efficiently used staff time. This process relied on automation through computer scripts and consistent documentation to form a reproducible workflow and support the traceability and reliability of final data products. Final data products were validated through a comprehensive review process to demonstrate that migrated data accurately represented the original documentation.

15. SUBJECT TERMS
Data management, Open data, Reproducible workflow, R software, RISE compatibility

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

Mission Statements

The U.S. Department of the Interior protects and manages the Nation's natural resources and cultural heritage; provides scientific and other information about those resources; and honors its trust responsibilities or special commitments to American Indians, Alaska Natives, and affiliated Island Communities.

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. This work represents contributions by the staff of the Environmental Monitoring and Assessment Branch (CGB-157).

Leverage Existing Environmental Data for Improved Usability by Standardization and Migration to RISE-Compatible Database

Final Report ST-2022-19210-01

prepared by

Laurel Dodgen, Ph.D.
Physical Scientist, Division of Environmental Affairs

Peer Review

Bureau of Reclamation
Research and Development Office
Science and Technology Program

Final Report ST-2022-19210-01

Leverage Existing Environmental Data for Improved Usability by Standardization and Migration to RISE-Compatible Database

LAUREL DODGEN Digitally signed by LAUREL DODGEN
Date: 2022.09.19 08:41:08 -07'00'

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

DEBORAH TOSLINE Digitally signed by DEBORAH TOSLINE
Date: 2022.09.16 15:54:18 -07'00'

Peer Review by: Deborah Tosline, RG, PMP
Hydrologist, Program Development Division, Region 8

Acronyms and Abbreviations

BHP	Regional Baseline Monitoring Investigation
CAS	Chemical Abstract Service
CGB-157	Environmental Monitoring and Assessment Branch
QA	Quality Assurance
QC	Quality Control
RISE	Reclamation Information Sharing Environment
U.S.	United States
UTM	Universal Transverse Mercator
WGS84	World Geodetic System

Definitions

Back-end
The computer server, application/software, and database that retrieves and stores data.
Completeness
The quantity of usable data collected compared to the quantity planned for collection.
Crosswalk
A table that shows equivalent fields in separate database tables.
Database Architecture
The complete design of the database, showing all the individual components.
Database Administrator
The information technology specialist who directs all activities related to the operation and maintenance of a database.
Data Element
A single point of data. A single Excel cell is often a data element.
Data Record
A single data observation with multiple data elements. A single row in Excel is often a data record.
Data Standards
A system of rules for how data is formatted and how the database functions.
Field
Commonly shown as a column in a table, a field is a set of data values in a table.
Front-End
The user interface that allows retrieval and access to data stored in a database.
Long Format Table
A table arranged to minimize width, with each observation arranged as a new row.
Lookup Table

A database table that states allowed values for a particular field. Often used to constrain entry to that field and disallow entered values that are not in the lookup table.

Metadata

Data that provides supporting information but is not a primary result.

Primary Key

A field or combination of fields that is unique for each record in a database table. Often sequential integers are assigned to an identification field that is used as the primary key.

Query

A request for data from a database.

Wide Format Table

A table arranged to minimize length, with all observations related to a particular sample arranged in the same row.

Contents

Page

Executive Summary

The Environmental Monitoring and Assessment Branch (CGB-157) recently built a new database to improve data integrity and facilitate data upload to Reclamation Information Sharing Environment (RISE). This project built on that effort; CGB-157's historical data was standardized, supplemented, and migrated to the new database for improved data accessibility and usefulness. This project evaluated the data products that were needed to migrate the data and identified four tables: station records, field activity records, sample result records, and quality assurance records. Gaps and incompatibilities in the historical data were assessed and many areas were identified that needed data standardization or supplementation to improve the data and make it conform to data standards for CGB-157's new database and RISE.

CGB-157 has over 600,000 environmental monitoring records and associated meta-data. The data sets were prioritized for migration so that highest-value data was available soonest. Data elements were prioritized for standardization and migration so that staff time was spent on high value tasks.

A project process was developed that preserved data quality and efficiently used staff time. This process relied on automation through computer scripts and consistent documentation to form a reproducible workflow and support the traceability and reliability of final data products. Final data products were validated through a comprehensive review process to demonstrate that migrated data accurately represented the original documentation.

From November 2017 through September 2022, staff spent about 3,500 hours to standardize, supplement, and migrate about 350,000 environmental records. Migrating records to the new database is expected to save CGB-157 about 35 hours a year in internal data searches and 110 hours a year facilitating external data searches. Data in the new database also has an enhanced pipeline to RISE. The new database already has a crosswalk that relates CGB-157 database standards to RISE standards. By migrating data into the new database and using that previously established crosswalk, staff do not need to develop individual crosswalks for each data set. This efficiency is expected to produce a one-time savings of 200 staff hours, with annual smaller savings of about 30 hours. Additionally, future data updates to RISE from CGB-157's database can happen automatically without need for staff time. Overall, this project made data more accessible to resource managers, more discoverable by the public, and more useful for environmental evaluations, supporting Bureau of Reclamation's mission and public trust.

1.Introduction

The Environmental Monitoring and Assessment Branch (CGB-157) performs environmental monitoring, generates monitoring data, and manages that data for users to make effective, informed resource management decisions. Data are frequently used to prepare Water Quality Monitoring Plans, Environmental Impact Statements, Environmental and Disposal Liabilities, and other official

documents; to demonstrate compliance with legal orders; and to prepare reports and recommendations. These uses require that the data demonstrate high integrity, completeness, and legal defensibility.

CGB-157 has generated over 600,000 data records with accompanying metadata at a cost of about $30,000,000. More than 2000 analytes covering physical, inorganic, organic, and toxicity characteristics have been measured in over 20 sample matrices including water, sediment, soil, and biota. Sample locations include more than 3000 sites across California, Nevada, and Oregon. When CGB-157 was created in the 1980's, data was managed as paper records. Data management transitioned in the early 2000s to manual entry into a Microsoft Access database front-end with Oracle back-end. A review of this database around 2016 showed that the database was not built according to current database standards, including lacking proper table relationships and data constraints. In 2017, CGB-157 started development of a modern, relational database to improve data quality, data usability, and workflow processes. That work was supported by the Science and Technology Program under project ID 7124. More information on the design and development of the new database is available in the report for that project [8]. CGB-157's new database also allows CGB-157 to comply with government open data efforts [1-5], particularly Department of Interior's Open Data Policy [6] and the OPEN Government Data Act [7].

The Reclamation Information Sharing Environment (RISE) was also developed in response to the Open Data Policy. RISE is a data-sharing platform that is intended to act as a central source of Reclamation data, storing a variety of data from facility operations 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 users, external stakeholders, and the public. During development of CGB-157's new database, RISE's data standards were evaluated. CGB-157's database was designed to interface smoothly with RISE and allow automated data uploads.

As new environmental monitoring data is generated by CGB-157, it will be entered to the new database. However, all of CGB-157's historical data remained in the historical database where it was not discoverable or accessible by the public. In this disposition, the data is not discoverable for resource managers and not compliant with open data requirements. A systematic process was needed to update the data so that it was compatible with the new database and with RISE and to produce data files suitable for migration to the new database. This process also needed to demonstrate data quality preservation and be efficient with staff time. For this project, a process was developed to reproducibly standardize, supplement, and format data sets. A reproducible workflow was implemented through documentation and automation to enhance data integrity and minimize staff time.

2.Methods and Results

This section will discuss the development of project goals, the project process, and the generation and verification of data products.

2.1 Determine Data Goals

Data is generated by CGB-157 to meet legal and contractual obligations and support resource management decisions. Data must have high integrity, completeness, and legal defensibility. During the database development process [8], it became clear that historical data was of variable quality. Over decades, data standards and operating procedures had improved and staff had turned over. For good data stewardship, records needed to be evaluated and improved for data usability.

Historical data encompassed direct result records and various types of metadata. Many kinds of metadata had already been improved and migrated to the new database as part of designing and building the new database. For example, the table of analytes was reviewed. Missing information, such as Chemical Abstract Service (CAS) numbers, were added as needed. Duplicate analytes were identified and consolidated (reducing the table from 2042 analytes to 1793). The final table of analytes was uploaded to the database and acts as a lookup table that constrains analyte result entry. Numerous other tables were also cleaned and migrated, including staff information, result units, investigation information, quality assurance (QA) materials, laboratory information, sample matrices, weather observations at sampling stations, and sampling equipment.

A review identified several groups of data that had not yet been migrated to the new database. These data were composed of records about field activities, sampling stations, analytical sample results, and quality control (QC). These data needed to be catalogued, matched to new database fields, and updated to meet data standards.

2.1.1 Station Records

Station data is composed of information about a sampling site, such as latitude and longitude coordinates, site type, and site name. These data are stored in the new database in the TBL_STN_INFO table. The historical station table had data in seven fields that needed to be migrated to TBL_STN_INFO.

2.1.2 Field Activity Records

Field activity data is composed of metadata about a sample collection, such as sampler name, sampling station, and observational comments. These data are stored in the new database in the TBL_ACTIVITY table. The historical activity table had 10 fields that needed to be migrated to TBL_ACTIVITY.

2.1.3 Sample Result Records

Sample result data is composed of measured results (e.g., analyte concentrations in a sample) and metadata (e.g., analytical method used for sample analysis). These data are stored in the new database in the TBL_SMPL_RESULT table. The historical result table had 14 fields that needed to be migrated to TBL_SMPL_RESULT.

2.1.4 Quality Assurance Records

QA data is composed of measured QA results (e.g., recovery and precision of QC samples) and metadata about sample results (e.g., relationship to sample results). These data are stored in the new database in the TBL_USBR_QA_RESULT table. The historical QA table had 16 fields that needed to be migrated to TBL_USBR_QA_RESULT.

2.2 Assess Data Gaps and Incompatibilities

After determining the scope of historical data remaining to be migrated, the data was evaluated against the data standards of CGB-157's new database and of RISE to determine which required data elements were missing from historical data sets and which data elements were not in compliance with the data standards.

Many data gaps were identified when comparing the required CGB-157 and RISE data elements to historical data. Data gaps fell into two categories: structural gaps and entry gaps.

CGB-157 data was already of high quality and completeness, in a general sense. Field activity documentation was generated and reviewed by trained staff. Lab report documentation was reviewed and QA validated by trained staff members. Documentation was manually entered to the historical database by trained staff and then all values verified by a second trained staff member. While developing the database, an initial data review suggested that only about 1% of records had inaccuracies, though many records had entry gaps (discussed in Section 2.2.2).

2.2.1 Structural Gaps

These missing data elements existed due to upgrades to the CGB-157 database capabilities, improvements in CGB-157 data standards and database architecture, and requirements to make data acceptable for RISE upload. When the database tables described in Sections 2.1.1 through 2.1.4 were assessed, it was determined that historical tables lacked fields to fill 25 fields in TBL_ACTIVITY, 2 fields in TBL_STN_INFO, 28 fields in TBL_SMPL_RESULT, and 15 fields in TBL_USBR_QA_RESULT.

Structural data gaps with high impact on data usability are described in Sections 2.2.1.1-2.2.1.4.

2.2.1.1 Field Result Units

Most historical data included analytical results that were measured at the sample location, namely, sample pH, electrical conductivity (EC), temperature, turbidity, reductive-oxygen potential (EH), and dissolved oxygen (DO). These data were housed in the same historical database table as field activity records (e.g., sampler name, date collected) in a misguided effort to simplify date entry. From a data hierarchy and data relation stance these data must be housed in the table of sample results. One of the problems with housing these data in the activity table is that all field result data were missing units for the measured values. The units needed to be supplied before the data could be migrated to the new database.

2.2.1.2 Station Elevation

RISE data standards require that information for sampling stations include elevation. CGB-157's historical database did not capture elevation information, so this required data element was missing from all records and needed to be supplemented before CGB-157 data could be uploaded to RISE.

2.2.1.3 Analyte-Specific Quality Control Subtypes

A QC subtype describes the category of a QC sample (e.g., "matrix spike", "matrix duplicate", "field blank", etc.). QC subtypes may vary among the analytes measured in a sample. For example, a triplicate sample may be collected and remain as a triplicate for low-level mercury but be spiked with

added chemical to be a matrix spike for boron. Therefore, analyte-specific QC subtypes are needed to accurately document the sample type and evaluate the sample results.

The historical database included two places to record QC subtype, in the field activity table and in the QA batch table. In the field activity table, each sample could only have one QC subtype (i.e., QC subtype is sample-specific). In the QA batch table, each sample in a particular batch could have one QC subtype. While assigning at the batch level provided more individualization, it did not allow different QC subtypes among different analytes in the same sample and in the same batch.

In the new database, analyte-specific QC subtypes were enabled by moving this field to the sample results table. This architecture improvement required that each result have an identified QC subtype while formerly only each sample/batch had a QC subtype. The historical sample QC subtypes needed to be extrapolated to fill the expanded analyte QC subtypes and the extrapolated values needed to be verified before the data could be migrated to the new database.

2.2.1.4 Database Identification Values

Current best practice for databases includes a requirement for each database table to have a field for record ID. This ID uniquely identifies each record in the table, simplifying constraints, queries, and table relationships by providing an alternate to a compound primary key. Since these record IDs were newly generated by building the new database, they did not exist in the historical data and needed to be supplemented to ensure data integrity during upload. For example, each sample station had a new, unique "STN_ID" that needed to be integrated with the historical data. Many data elements needed to have record IDs integrated with the data set, including investigation name, matrix type, sampler name, batch number, laboratory name, analyte name, unit, analytical method, data qualifier, and QC subtype.

2.2.2 Entry Gaps

These missing data elements existed when the historical database had fields to store the data, but the values were not entered. Two primary causes for these missing data are likely. Over decades, data entry procedures and staff behavior varied. Also, the historical database did not use many reasonable data constraints that would require entry before accepting a new record. These factors would easily lead to variations in data completeness.

2.2.2.1 Station Coordinates and Type

The historical sampling station table had fields to record the latitude and longitude coordinates and the station type (e.g., "river", "facility", "reservoir"). However, it lacked constraints to require users to enter these values when creating a new sampling station, and consequentially many values are missing. For example, the historical table only had coordinate values for 599 of the 3245 records. Further review also showed that some values had been entered with Universal Transverse Mercator (UTM) coordinates while others had been entered with World Geodetic System (WGS84). This missing and inconsistent data greatly diminished the utility of the data, because users could not search for sample results from their area of interest and could not perform spatial data analysis and visualization. Likewise, missing station types limited users from performing categorical data analysis and identifying data sets from similar station types. To make historical data more accessible and to make station records compatible with the new database and RISE, coordinates and station type needed to be supplemented from available documentation before the data could be migrated to the new database.

2.2.2.2 Quality Assurance Batches

The historical database had a minimally acceptable architecture for storing QC and QA data. One table contained records of sample results, which included a batch number that related all results in a QA batch. A second table housed the QC subtype of each sample in a batch. A third table showed QA results in the batch (e.g., the calculated spike recovery).

However, users were not required to enter a batch number for a result record, and not required to enter a batch number in the other two tables that matched the value in the result record or matched each other. From viewing historical data, it appears that users stopped entering batch numbers to sample result records nearly the moment that the database was first in use. This significant gap in data standards led to loss of relationship among sample results and QC subtypes and QA results.

About 5-9 years after the database came into use, the data entry procedure normalized on entering a batch number in the QC subtype table and QA result table (but not the sample result table). The entry was a combination of the lab report number and a sequential alpha character. For example, the first batch on lab report '123' became '123A' and the second batch became '123B'. This action was not ideal because the modified lab report number was not fully unique and could accidentally connect unrelated records. Also, because the entry was not constrained to match any field in the sample result table, there was no explicit relationship between the modified lab report number and a batch of samples. This modified lab report number could have been placed in the batch number field of the sample result table, thus making the relationship explicit. Unfortunately, it was not entered, and the batch number field of the sample result table is empty for most records.

However, because the modified lab report number was assigned in a systematic way, the batch number could be extrapolated from it when it was available. The extrapolated batch numbers needed to be verified and the missing values supplemented before the data could be migrated to the new database.

2.2.2.3 Quality Control Subtype

The historical table that housed QC subtype records was arranged in wide format, with each record representing a single batch (Table 1). Each record had a batch number (usually the modified lab report number). The user entered the sample name into the appropriate field for its QC subtype. In the example in Table 1, the sample name 'FOO101' is entered as the background sample for batch '123A'.

However, a field was omitted when creating the database: QC blank spike. If the table had been arranged long format, this issue could have been easily fixed by adding another approved value option to the 'QC subtype' field (Table 2). Since this table was arranged wide format, users had to select one of the similar existing fields (i.e., matrix spike or reference). The data shows high variability in blank spikes being entered to the matrix spike or reference fields, likely due to staff having different training and procedure shifting over time.

Table 1. Example of wide format table for QC subtype.

Batch Number	Background Sample	Duplicate Sample	Blank Sample	Matrix Spike Sample	Reference Sample

123A	FOO101	FOO102	FOO103	FOO104	
456A	BAM001		BAM003		BAM002

Table 2. Example of long format table for QC subtype.

Batch Number	QC Subtype	Sample Name
123A	Background sample	FOO101
123A	Duplicate sample	FOO102
123A	Blank sample	FOO103
123A	Matrix spike sample	FOO104
456A	Background sample	BAM001
456A	Reference sample	BAM002
456A	Blank sample	BAM003

The database lacked a constraint for a user to enter every sample name into a QC subtype field, so many result records have no associated QC subtype. The table of QC subtypes stored valuable data that could be parsed to provide QC subtype for result records (when the data was entered to the table). However, the parsed QC subtype values needed to be verified and the missing values supplemented before the data could be migrated to the new database.

2.2.2.4 Quality Assurance Results

QA results are integral to using result data in an informed manner. QA results can include accuracy recovery values, accuracy spike values, precision deviation values, and contamination detection values. The historical table that housed QA results was also arranged in wide format. For each record, users could enter batch number, analyte name, a spike value (lacking units), a certified reference value (lacking units), and a QA result for each of 13 QA indicators (e.g., matrix spike recovery, matrix duplicate precision). Just as users were not required to enter a batch number or assign a QC subtype for a record, they were also not required to enter QA results. This gap ranged from a few records for some investigations, which may have been user error, to about half of all records for other investigations, which suggests changes in procedure over the decades. The historical table of QA results stored valuable data that could be parsed. However, the missing QA results needed to be manually supplemented before the data could be migrated to the new database.

2.2.3 CGB-157 Database Compatibility

When the new database was being developed, the existing database structure and content was reviewed to understand the scope of data to be housed. Particularly, the 90,000-record data set from the 'Regional Baseline Monitoring' (BHP) investigation was deeply reviewed. This multi-decade effort included organic and inorganic analytes in environmental samples across California, Oregon, and Nevada. Due to its length of activity, breadth of sample locations, and large parameter set, it was considered representative of the overall database content. Reviewing the BHP data in the historical database showed many elements that were incompatible with data standards for the new CGB-157 database.

2.2.3.1 Collection Date and Time

In the historical database, sample collection dates and times were entered as a string data type in two separate fields. In the new database these data are stored in one field that is compliant with ISO

8601. In order to migrate historical data, collection date and time needed to be combined and translated to ISO 8601 format.

2.2.3.2 Data Qualifiers

The historical database had four fields in the sample result table to capture data qualifiers, with each type of qualification noted by its own field. For example, whether a result met its analytical holding time was noted in one field, while results that were outliers were noted in a second field. At some point in the past, qualification information stopped being entered to those four fields and instead was entered as string characters in the numeric (but unconstrained) sample result field. For example, a sample result record that measured boron in a sample at 460 µg/L but was found to be biased high due to QA results might have the value "460 H" in the result value field ("µg/L" is stored in the result unit field). This mixed data type was incompatible with the improved database architecture, and also problematic for data users since string characters needed to be cleaned from results before data analysis or visualization. To migrate the historical data, qualifier information needed to be scraped from all five fields and consolidated into a single, non-repetitive character string in its own field.

2.2.3.3 Field Results

As discussed in Section 2.2.1.1, field results were stored in the field activity table historically, but moved to the sample result table in the new database. This adjustment required field result data to have the units added to the result unit field for all records, and for field result records to be merged with other sample result records.

2.2.4 RISE Compatibility

For data creators to upload data to RISE, data sets must conform to RISE data standards for format, content, and metadata. All data that CGB-157 chose for upload to RISE needed to be supplemented and crosswalked to meet RISE's requirements. For example, samples of treated water in CGB-157's new database have the matrix type "WATER_TREATMENT" and matrix ID "17", while those same samples have a RISE matrix type "Treated Water" and matrix ID "19". All CGB-157 data and metadata needed to be matched to approved values in RISE lookup tables and crosswalked to produce RISE-conforming values. Any missing data, such as station elevation (see Section 2.2.1.2), needed to be supplemented.

To upload data sets from CGB-157's historical database, values in each data set would need to be individually crosswalked and formatted. This process would cost considerable staff resources, estimated at over 200 hours for the initial data sets. Subsequent data updates would require further staff time of about 30 hours each year. However, data sets in the new database have a stream-lined path for RISE upload. As part of the new database development, comprehensive crosswalks between CGB-157 tables and RISE tables were established that covered all current data values. Once data sets are in the new CGB-157 database, they can automatically upload to RISE after the user submits RISE's initiation paperwork. Future data updates can happen automatically, with no need of further staff time. While saving staff-time, these automatic uploads also prevent bottlenecks of human work that would delay data being uploaded to RISE and accessible to Reclamation employees and other data users.

2.3 Prioritize Data

2.3.1 Data Sets

CGB-157 has over 600,000 records of environmental monitoring data from over 70 investigations. While the project goal was to migrate all historical data into the new database, data sets needed to be prioritized so that high-value data was migrated soonest. The project lead contacted 101 stakeholders who worked with environmental data and environmental resources in the western United States (U.S.). Stakeholders were identified in Department of Interior, Bureau of Reclamation, state and regional water quality control boards, research universities, and non-governmental organizations. Stakeholders were requested to provide input on types of data or geographic areas of data that they used frequently. Stakeholders were also directed to CGB-157's data inventory if they wanted to recommend particular data sets for early migration.

Data sets were prioritized for migration using the following ranking. Number in parentheses denotes what portion of CGB-157 records fell into that rank.

1. Active investigations with large data sets (63%)
2. Active investigations with small data sets (10%)
3. High-profile inactive investigations (12%)
4. Recent inactive investigations (7%)
5. Remaining investigations (8%)

2.3.2 Data Elements

When reviewing historical data, it became clear that some data elements were crucial to the utility of the data set, such as the result value and unit, the sampling station coordinates and elevation, field comment, and QC subtypes. Other data elements were useful but not crucial, such as the date a lab report was generated or the weather at time of sample collection.

The project work required extensive staff time and focusing on crucial data elements gave the best return on investment. To illustrate, if 100 data elements could be reviewed and fixed by a staff person in an hour and the staffer reviewed both crucial and non-crucial data elements, that staffer might finish cleaning four records in an hour. But if that staffer reviewed only crucial data elements, they might clean 20 records in an hour. Non-crucial data elements were migrated without cleaning, leaning on CGB-157's high-quality data entry and review process. Focusing on crucial data elements allowed more data to be cleaned and migrated while preserving the utility of the data.

The review of historical data sets demonstrated that four database tables needed to be migrated to the new database (Section 2.1). When the four tables were assessed, it was determined that historical tables had structural gaps (Section 2.2.1) and lacked data to fill over 70 fields (Table 3). However, only eight fields were considered crucial. Reviewing entry gaps (Section 2.2.2) showed where the historical database had fields to fill the new database but many records were missing crucial data elements. Both crucial data elements identified from structural gaps and from entry gaps were addressed in the standardization and supplementation process.

Table 3. Assessment of crucial data elements that are missing from historical database tables.

Table Content	Missing Historical Data Elements	Crucial Missing Data Elements	Crucial Elements Description

Field Activity	25	3	QC types and batch
Sample Information	5	2	Station ID, elevation
Sample Result	28	1	Analyte-specific QC subtype
QA Result	15	2	Accuracy symbol, spike unit

Two additional data concerns arose during review of BHP data (Section 2.2.3), in addition to structural and entry gaps. Over time, the data entry procedure had placed insufficient emphasis on entering alkalinity and sediment data. As a result, many historical records did not identify or mis-identified the form of alkalinity for a result. For example, a bicarbonate alkalinity value may be analyzed and reported as "bicarbonate alkalinity as HCO_3" or "bicarbonate alkalinity as $CaCO_3$". Misidentifying one form of alkalinity as the other form can result in 20-40% error depending on the analyte. Similarly, sediment samples may have their results reported as "dry-weight basis" or "wet-weight basis". Many historical records did not identify or mis-identified the basis for the result. This omission caused error proportional to the percent moisture of the sample, which can commonly be 15-60% for environmental samples. Because of these serious issues, all alkalinity and sediment data were marked for review during the standardization process.

2.4 Develop Process Structure

A large data project needs a comprehensive plan to ensure data quality and a successful project outcome. Planning begins with identifying the data goals for the project (Section 2.1), and then identifying the scope of data to be managed and updated to meet the data goals (Sections 2.2 and 2.3). Finally, a reproducible workflow is devised that produces reliable data products through documentation and automation and that is resilient to project changes and staff turnover.

2.4.1 Documentation and Structure

A project folder was set up on a shared network drive with a systematically organized subfolder structure. All files were systematically labeled and stored in appropriate subfolders. A Google Sheets workbook was set up to facilitate highly collaborative tasks. It was later migrated to a Microsoft SharePoint Excel workbook when Department of Interior switched to Microsoft products. Both Google Sheets and Microsoft SharePoint host document versioning, which is valuable for viewing and reverting user changes to the document. Easily accessible and clearly labeled documents facilitated multiple needs, including:

- Keeping documentation updated
- Collaborating with team members
- Working on tasks in parallel
- Supervising work progress

2.4.2 Original Data

The first step of standardizing and supplementing each data set was making a backup of the original data. For this project, the backup was a download of the data that was stored in a subfolder of the shared network project folder. Creating this backup meant that at any time environmental

monitoring data could be reverted to the original content. It also served as useful reference to understand how a data piece had been updated due to the project process. This action was a critical step in ensuring data integrity.

2.4.3 Areas for Automation

Automation is a valuable tool to reduce staff hours and manage data in a more reproducible manner. Furthermore, very large data sets may cause software like Microsoft Excel to operate slowly or not run at all, creating barriers to performing tasks in a manual process. For this project, automation by bespoke computer scripts was implemented in many process areas after a review of what tasks could be successfully automated. All scripts were written in R and used in the integrated development environment R Studio. While some tasks were performed manually for best data quality or efficiency, many tasks were partially or fully automated. See Sections 2.5-2.6 for details.

2.4.4 Work Teams

This project involved work on many elements of each data set and on many data sets. For best data quality and efficiency, work was organized into discrete domains of station records, field records, and result records. Tasks for one domain could be performed independently from other domains and were assigned to specific teams. Teams were formed from staff members with relevant knowledge and experience for the task, but consideration was also given to the workload and enthusiasm of each staff member. Generally, QA and data entry experts worked on field and result records, while sampling experts worked on station records.

2.5 Generate Data Products

Four data products were generated by this project for each data set that was migrated. An Excel Workbook (.xlsx) was produced for each data product, corresponding to data from station table, field activity table, sample result table, and QA result table. The process to generate each data product was unique to the structure of that database table and the crucial data elements in the table.

2.5.1 Station Records

The entire content of the station table was placed in a Google Sheet worksheet to facilitate parallel workflow. Additional fields were added to the original data to hold updated station values and staff assignments. Each member of the station team was assigned rows of records to review, based on familiarity with the station area, such that all station records were reviewed.

2.5.1.1 Documentation Review

Team members reviewed their assigned stations, referencing investigation documentation such as Water Quality Monitoring Plans, Sampling and Analysis Plans, Quality Assurance Project Plans, client reports, and site photos, as well as personal knowledge. When possible, team members supplied missing information and updated incorrect information for latitude, longitude, station type, U.S. county, and U.S. state in the new fields of the worksheet. At the task initiation meeting, team members were given specific instructions on how to fill updated values:

- Latitude: WGS84 decimal degrees without a degree symbol
- Longitude: WGS84 decimal degrees without a degree symbol
- Station type: conforming to the allowed values in the geographic origin lookup table
- County: capitalized and unabbreviated

- State: 2-character, capital letter U.S. postal code

After an initial review, team members were able to supply values for 1754 stations of 3243 total stations, leaving 1489 unfilled. Of unfilled stations, 91% (1356 stations) had been sampled three or fewer times which suggested that they represented a small amount of data. Team members performed a second round of review where they self-assigned unfilled stations based on familiarity with the station or prioritizing stations that were frequently sampled. After the second review, a further 94 stations were filled. The data gains from the second review cycle highlight that teams are essential to project success because they increase work quality, in addition to work throughput.

2.5.1.2 Duplicate Review

The reviewed and updated station records were then automatically processed by the computer code script "stn.clean-up.reviewed.list.R" in less than five minutes. This script performed several functions:

1) Reformatted data to conform to new database structure.
2) Checked and corrected station types to conform to geographic origin lookup table.
3) Checked that only numbers and decimal marks were entered in latitude fields.
4) Checked that only numbers, decimal marks, and minus signs were entered in longitude fields.
5) Checked and corrected longitude values to all begin with minus sign.
6) Checked county values are valid U.S. counties and capitalized.
7) Checked and corrected U.S. state values to valid 2-character, capital letter postal code.
8) Grouped stations that were closely located by performing a recursive box-fitting to the coordinate data. Briefly, a box was drawn around an initial station that stretched 0.007 decimal degrees in each cardinal direction (about 2000 feet). Any station within that box was grouped with the initial station. The recursion ran, redrawing the box larger to include each newly encompassed station until no new stations were within the box. All stations within the box were assigned the same group ID. This systematic and automated process allowed closely located stations to be found quickly and accurately without extensive staff time.
9) Exported table of grouped stations. Any stations that were alone in their group were not exported.

The table of grouped stations was copied into a second worksheet of the Google Sheet workbook. The station team members reviewed grouped stations and provided assessments based on personal knowledge and documentation on each station's uniqueness from other stations in the group. For stations that were unique, the team members entered that station's record ID in the column headed by their initials. For stations that the team member judged to be a duplicate of an earlier station, they entered the earlier station's record ID in the column headed by their initials. Of 3243 stations, 1330 stations were flagged for duplicate review.

The reviewed table of grouped stations was automatically processed by the script "stn.duplicate.dissent.R". This script summarized the station assessments provided by team members. For stations where all team members agreed on the station's identity, no further action was taken. For stations with dissenting assessments, station records and assessments were summarized in a table and exported to a file. Of 1330 initially flagged stations, only 177 stations had dissenting assessments and these records were reviewed by the station team. Some records were easily resolved with discussion in the meeting and the consensus decision was marked in the file. A

few records were selected for documentation review by a station team member and further discussion in a follow-up meeting with all team members.

After all records with dissenting assessments reached a consensus decision, the decision was marked in the file, and the file was automatically processed by the second-half of the script "stn.clean-up.reviewed.list.R". This script summarized all decisions about station identities to create a file where each station was given a final station ID. For unique stations, their station ID was their record ID. For duplicate stations, their station ID was the record ID of the first station record at that location.

From the initial station table with 3243 stations, review was able to identify 267 duplicates and consolidate the table to 2976 stations. This process served the simple utility of reducing the length of the station table for easier browsing. More importantly, consolidating stations allowed the data collected under separate names to be pooled together into large data sets with more statistical power. In one case, five separate stations were identified that were the same sample location. By combining these five data sets, a single data set was made with 1517 data results. Users interested in data for that sample location can now pull all available data with one station name and have a larger data set for data analysis.

2.5.1.3 Elevation Supplementation

Elevation values are required by RISE. Stations with coordinates were automatically supplied elevation values by the script "LWQD.stn.GET.elevation.R". This script runs a call to the application programming interface (API) at "https://nationalmap.gov/epqs/pqs.php". This service is hosted by the U.S. Geological Service and provides data of elevation at specific coordinates across the U.S. The script automatically queried the API for the elevation of each station's coordinates and filled the elevation into the file of station information. If a user had performed this task by manually entering each coordinate set into the user interface website and then entering the returned elevation value into the station file, this task would have taken about 15 hours. The script performed this task in under 10 minutes and was not vulnerable to transcription errors.

2.5.1.4 County and State Supplementation

U.S. state values are required by RISE. Stations with coordinates were automatically supplied U.S. county and state values by the script "lat.long.parser.R" which was based on the function latlong2 in the package jvamisc. This script references a library of geospatial polygons to determine the county and state polygon that contains each coordinate set. If a user performed this task manually, it would have taken at least 50 hours and been highly susceptible to transcription errors. This script performed this task in under five minutes, ensuring data quality and saving valuable staff time.

The final file of de-duplicated stations with Stations IDs, station names, coordinates, elevations, counties, states, and other station data in the database-conforming format was supplied to the database administrator for upload to CGB-157's new database.

2.5.2 Field Activity Records

Data was standardized by the process documented in the project file "CGB-157 LWQD Project Workbook" tab "Transfer Process ver 5" and narrated in Sections 2.5.2.1-2.5.2.3.

2.5.2.1 Field Result Units

All field result data were missing units in the historical database. The script "LWQD.find.weird.field.result.R" was used to process field result data and flag ambiguous data for human review. Previous review of historical data had shown that six analytes were measured in the field and results were entered into the database without result units: pH, EC, temperature, turbidity, EH, and DO. After data review and discussion with the sampling team, it was determined that most were always measured with the same units: pH as units, EC as μS/cm, turbidity as NTU, and EH as mV. However, temperature was entered as degrees Celsius (°C) or degrees Fahrenheit (°F) and DO was entered as % or mg/L.

The script reviewed all field result data. For temperature results, if the result was ≤ 32 the result was presumed to be in units of °C. If the result was ≥ 50 and < 99, the result was presumed to be in units of °F. For DO results, if the value was ≤ 13 the result was presumed to be in units of mg/L and if the value was ≥ 40 and < 105 the result was presumed to be in units of %. These cutoff criteria were determined after data review and discussion with the sampling team. All temperature data between 32-50 and over 99 and all DO data between 13-40 and over 105 was reformatted and exported for human review.

This preliminary automated screening was useful for reducing staff time by focusing review on only ambiguous records. For example, for one data set containing 8268 records of field analyte data, no temperature records and only 10 DO records needed human review. Once a staff member reviewed original documentation, a unit was determined and marked in the exported file for later data incorporation.

2.5.2.2 Field Comments

Staff performing sample collection were able to document observational comments, which were later entered into the historical database exactly as written. This important tool led to variable and sometimes useless data as the standard for the content and the format of the field comment varied among samplers and over time. The script "LWQD.field.comment.R" was used to review and standardize field comments. The script first removed field comments that contained no useful information by using the coding technique of regular expressions. For example, at one time the convention had been to write the QC subtype in the field comment. However, this information is already captured more accurately in the historical database table of QA results. After removing these types of comments, the remaining comments were exported for human review and later data incorporation.

2.5.2.3 File Generation

After the steps in Sections 2.5.2.1-2.5.2.2 were complete, a portion of the script "LWQD.produce.FINAL.file.R" was used to synthesize, standardize, and supplement the field data and then arrange it in a format compatible with the new database. The script took fewer than five minutes to perform the following functions for a data set:

1) Loaded lookup tables and crosswalk tables.
2) Loaded historical field activity table and new field comment table.
3) Formatted collection date and time into one column of data conforming to ISO 8061.
4) Retrieved and filled historical record info for:
    a. sample name
    b. sample depth

5) Matched historical record info and filled database-conforming values for:
    a. updated (i.e. de-duplicated) station ID and name
    b. investigation ID and name
    c. matrix ID and name
    d. sample type ID and name
    e. sampler(s) ID and name
6) Supplied updated field comment to related record.
7) Exported table for human review and database upload.

2.5.3 Sample Result Records

Data was standardized by the process documented in the project file "CGB-157 LWQD Project Workbook" tab "Transfer Process ver 5" and narrated in Sections 2.5.3.1-2.5.3.7.

2.5.3.1 Automated Data Flagging

After the original data sets were downloaded and placed in the folder for original historical data, the sample result records were processed by the script "LWQD.find.weird.chem.result.R". This script performed the following functions:

1) Loaded sample result records without altering them.
2) Flagged records where:
    a. analyte was measured < 20 times
    b. unit was used for an analyte < 4 times or < 25%
    c. unit is missing or invalid (e.g., "null", "uu", "-")
    d. method was used for an analyte < 4 times or < 25%
    e. method is missing or invalid (e.g., "NA")
    f. all alkalinity records (as discussed in Section 2.3.2)
3) Created a field to note which data element was flagged for review.
4) Sorted flagged records by lab report number and collection date so that related records are co-located for easier human review.
5) Exported table of flagged records.

The criteria for flagging sample result records were developed by a team after reviewing historical records, considering the critical data elements (Section 2.3.2), and considering the data structure for CGB-157 records. Nearly from the beginning of CGB-157 environmental monitoring, samples were batched together. A batch usually represented a single collection campaign and had QC samples included to evaluate and demonstrate data integrity. Batches usually were composed of 1 to 10 environmental samples and related QC samples, giving a sample total of 3 to 13 samples in a batch. The criteria described above were chosen to identify and flag a single batch that was mis-entered for a critical data element.

2.5.3.2 Manual Data Review and Correction

The exported file of flagged records was copied into the project Google Sheets workbook. A team of four experienced QA staff members reviewed the flagged records and compared the data to the original documentation. Staff noted any needed corrections in a comment field. A team of two experienced data entry staff members made the required corrections in the historical database, and then entered the date completed and their initials to the workbook. The original data was preserved in unchanged, downloaded files.

This method was chosen over updating the data in a downloaded file for three reasons. First, the data entry staff were experts in the historical database and were more likely to make the changes quickly and accurately in that location. Second, some changes required changes to multiple database tables. For example, if a method needed correction, this affected data in the sample result table and the QA result table. These related changes were quickly done with a targeted database query. Third, when this project was begun, the new database was not fully operational and available to clients. By correcting the data in the historical database, clients had access to the best quality data immediately. Other changes to the data (e.g., adding field result units, updating comments) were performed in files outside the historical database (not in the original downloaded files). These corrections were done in files rather than the database because 1) changes were not supported by the structure of the historical database or 2) the change did not affect a crucial data element and was not worth staff time to enter to the historical database.

After all review and corrections were complete, the sample result data set was downloaded again and placed in a folder for verified historical data.

2.5.3.3 Sample-Specific Record Review

As discussed in Sections 2.2.2.2 and 2.2.2.3, some information was available in the historical database for the QC subtype and batch number of a record. The available data was impaired because the batch number in the QC subtype table and QA result table was not entered to the batch number field of the sample result table, breaking the relationship between a sample result and its QC/QA data. Also, some records were not entered to the QC subtype and QA result tables. The upper portion of script "LWQD.batch.R" was used to scrape available data and then generate a file for human review that had QC subtype and batch number for each sample record. The script performed the following functions:

1) Formatted QC subtype table into long format with record ID, QC subtype, and sample name fields.
2) Matched and filled new QC subtype values to each sample record when data were available.
3) Created batch number(s) for each sample, using the record ID(s) in the reformatted QC subtype table, and assigned batch number(s) to sample record when data were available.
4) Sorted all records by sample name and collection date so that related records are co-located for easier human review.
5) Export table with sample name, QC subtype, batch number(s), and collection date.

Each exported table for a data set was reviewed by a member of the QA team. The separate tables were distributed among the team. Instead of splitting an exported table for parallel review, each exported table was reviewed by a separate person for logistical and efficiency reasons. When a single staffer reviewed a table, the original (paper) documentation did not need to be passed among multiple people. Sharing documentation took time and became a limiting factor during coronavirus stay-at-home orders. Also, as the staffer reviewed the table, they became very familiar with the structure of the environmental monitoring investigation and lab report data, which improved their efficiency at finding needed documentation and data.

Records that were successfully filled by the script were not reviewed, under the premise that historical data was entered and reviewed by trained personnel. Avoiding re-review saved considerable staff time. For example, one investigation had 901 total sample records but only 209

records were missing data. Staff focused on records that were missing QC subtype or batch data, reviewed original documentation, and entered correct values to the exported table.

When the file was finalized, it was processed by the lower portion of script "LWQD.batch.R", which ran in under five minutes. This script parsed the batch numbers and supplied new database-conforming batch ID(s), batch year(s), batch collection number(s), and batch group(s) for each sample record. It then produced two files. The first file summarized the final batches in a human-readable format for a manual spot-check. The second file conformed to the database format for the batch table, and was copied into the main batch data table for later data incorporation.

2.5.3.4 Preliminary File Generation and Flagging

The process to supplement QC subtype and batch number (Section 2.5.3.3) was sample-specific, but as discussed in Section 2.2.1.3 the new database requires analyte-specific QC subtype and batch data. An initial data review showed that the QC subtype and batch ID were mostly identical among a sample's analytes and so filling the sample-specific data to the analyte-specific records would generate valid data while saving staff time. Computer script was used to review the extrapolated data and flag records that had invalid QC subtype or batch data for further human review.

To illustrate, imagine a sampler collected 4 samples to be analyzed for 5 analytes. If a QA member entered sample-specific data for one QC subtype field and three batch fields, that would require entering 16 data elements. However, they would need to enter 80 data elements if the records were analyte-specific. By first filling sample-specific data, then reviewing it and correcting it, the process overall saved staff time while preserving data quality.

A portion of the script "LWQD.produce.FINAL.file.R" was used to integrate the corrected sample result table, new sample-specific QC subtype table, and new sample-specific batch table and then generate a table for analyte-specific data review. The script performed the following functions:

1) Loaded lookup tables and crosswalk tables.
2) Loaded updated sample result table.
3) Loaded new sample-specific QC subtype table and new batch table.
4) Checked and corrected human-entered QC subtype values to database-conforming values.
5) Generated a lookup table for sample names and associated batch ID(s).
6) Retrieved and filled historical record info for:
    a. sample name
    b. laboratory ID
    c. laboratory report number
    d. laboratory sample number
7) Matched historical record info and filled database-conforming values for:
    a. batch ID(s), year(s), collection number(s), and group(s)
    b. result symbol, value and unit
    c. reporting limit value and unit
    d. analyte ID, name, fraction, and CAS number
    e. preparation method ID and name
    f. analytical method ID(s) and name(s)
    g. QA comment
    h. de-duplicated data qualifier ID(s) and code(s)
    i. QC type ID(s) and name(s)

    j. QC subtype ID and name
8) Filled new database fields that lacked historical data with `NA` data type.
9) Subsetted table to only include records with unacceptable QC subtype and batch data, plus related records. Records were flagged when a sample result record had more than two batch IDs or when an analyte-specific batch had more than one sample result record with a QC subtype of "Regular" (i.e., the background sample for matrix duplicates and matrix spikes).
10) Sorted all records by sample name, batch ID, and analyte so that related records are co-located for easier human review.
11) Exported table.

2.5.3.5 Analyte-Specific Record Review

Records flagged for unacceptable analyte-specific QC subtype or batch data were reviewed by the QA team. Like the initial process to supplement the data, each exported table was reviewed by a separate person and the tables were distributed among the team. Staff reviewed the original documentation for all flagged records and updated the QC subtype or batch IDs in the table, following the process documented in the project file "CGB-157 LWQD Project Workbook" tab "Excess Regulars Process". The updated table was integrated into the preliminary sample result file using a portion of the script "LWQD.produce.FINAL.file.R". The script performed the following functions:

1) Loaded reformatted sample result table.
2) Loaded updated analyte-specific QC subtype and batch table.
3) Checked and corrected human-entered QC subtype values to database-conforming values.
4) Checked that all flagged records had been resolved, else threw an error message with the failing record(s).

The project process had the automated script partially fill sample-specific QC subtype and batch data, staff fill in missing data, automated script assess the validity of the filled data, and then staff correct the flagged records. This approach saved considerable staff time over having staff fill analyte-specific data from the beginning. For example, one project had 10,698 sample records and 69,602 analyte result records, with about 1/3 missing information. If staff had filled in missing analyte-specific data from the start, that process would have required filling about 22,300 records of missing data. By using the sample-specific approach first, staff had to fill in 3436 sample records, and then correct 10,887 analyte-specific records, reducing staff time for this step by about 35%.

2.5.3.6 Accuracy Record Supplementation

Some accuracy results were available in the historical QA result table (Section 2.2.2.4). This data had limitations because 1) the relationship between these data and the sample result data had been impaired when batch numbers were not entered in the sample result table and 2) only a portion of accuracy data had been entered.

The script "LWQD.clean.Access.QA.results.R" was used to clean up available accuracy data from the historical database. This script performed the following functions:

1) Loaded table of QA results downloaded from historical database.
2) Formatted QA result table into long format.
3) Dropped all records except accuracy type (e.g., matrix spike, blank spike, reference).
4) Standardized entered values.
    a. removed non-numeric values

    b. removed empty values
    c. removed values less than 20 and greater than 180
5) Merged separate fields for concentration of matrix spike and reference.
6) Merged separate fields for sample result of matrix spike and reference.
7) Exported table of accuracy records.

A portion of the script "LWQD.produce.FINAL.file.R" was used to integrate the exported accuracy records with the updated sample result table and then generate a table for accuracy record supplementation. The script performed the following functions:

1) Loaded exported accuracy table and updated sample result table.
2) Merged accuracy records with sample result records using compound matching with the sample name, analyte name, and any portion of lab report number.
3) Extrapolated missing spike/reference units using an evaluation of magnitude variation among the spike/reference sample value and sample result value.
4) Subsetted table to only include accuracy records and to only include columns helpful for human review, such as collection date, lab report number, and fields requiring entry.
5) Sorted table by sample name, lab report number, and analyte so that related records are co-located for easier human review.
6) Exported table.

Data entry team members reviewed the exported accuracy table using the process documented in the project file "CGB-157 LWQD Project Workbook" tab "Accuracy Process". Records that were missing accuracy result values were filled as possible from original documentation by human entry to the exported accuracy table. As team members attempted to enter missing accuracy results, they also entered spike/reference values, spike/reference units, and QC subtypes.

The filled accuracy table was processed by a portion of the script "LWQD.produce.FINAL.file.R". The script performed the following functions:

1) Loaded the filled accuracy table and the updated sample result table.
2) Checked analyte names, QC subtype names, and units were database-conforming values.
3) Checked if any QA results with QC subtype of matrix triplicate (a precision type) were erroneously included in the accuracy table and updated those records.
4) Updated QC subtype values in sample result table to match accuracy table.

2.5.3.7 File Generation

After the table of sample results had been improved by the sample-specific entry, analyte-specific updates, and accuracy supplementation, it was processed by a portion of the script "LWQD.produce.FINAL.file.R". The script performed the following functions:

1) Reformatted field result data to match sample result data and database requirements.
2) Matched field result records and filled database-conforming values for:
    a. batch ID(s), year(s), collection number(s), and group(s)
    b. lab ID and name
    c. lab report number
    d. lab sample number
    e. result symbol, value and unit
    f. reporting limit value and unit
    g. analyte ID, name, fraction, and CAS number

h. preparation method ID and name
i. analytical method ID(s) and name(s)
j. QA comment
k. data qualifier ID(s) and code(s)
l. QC type ID(s) and name(s)
m. QC subtype ID and name

3) Merged field result records and sample result records.
4) Exported table for human review and database upload.

2.5.4 Quality Assurance Records

Data was standardized by the process documented in the project file "CGB-157 LWQD Project Workbook" tab "Transfer Process ver 5" and narrated in Sections 2.5.4.1-2.5.4.2.

2.5.4.1 Quality Assurance Result Mining and Supplementation

The historical database had some QA results available, but the data was limited and impaired as discussed in Sections 2.2.2.2 - 2.2.2.4 and 2.5.3.6. An initial data review showed that many accuracy results in the historical table were usefully formatted and could be successfully scraped, related to their result records, and manually supplemented. The script and staff effort to perform that work is described in Section 2.5.3.6.

The precision and contamination records were in worse condition. Many of these records had text values that could not be easily updated to database-conforming values. For example, users had entered various versions of "+-RL", "<RL", and ">RL" as precision results, but the new database requires numeric results of the calculated difference between the background and duplicate results. For best data integrity and utility, as well as efficient staff time, a script was written to automatically calculate the numeric precision and contamination values using the sample results. Automatically calculating precision and contamination results, instead of going through a similar process as Section 2.5.3.6, saved approximately 1,200 staff hours from reviewing about 24,000 records.

2.5.4.2 File Generation

After the table of accuracy records had been updated through human review and the finalized table of sample results was exported, the table of QA results was generated with a portion of the script "LWQD.produce.FINAL.file.R". The script performed the following functions:

1) Loaded finalized tables of sample results and filled accuracy results.
2) Set up function to find the background sample for a QC sample.
3) Matched historical record data and filled database-conforming values for:
    a. sample name
    b. analyte ID and name
    c. laboratory report number
    d. method name
    e. QC subtype
4) Calculated the result value and result limit for precision and contamination samples.
5) Determined the result value, upper limit, lower limit, spike/reference value, and spike/reference unit for accuracy samples. If accuracy result was available in the accuracy table, it was used. Else, if spike/reference value, accuracy value, and background value were available, then accuracy result was calculated.
6) Exported table for human review and database upload.

2.6 Validate Data Products

CGB-157 produces environmental monitoring data to fulfill legal and contractual responsibilities. The data must be high quality, complete, and traceable. To demonstrate the data integrity of records after they completed the standardization and supplementation processes, the data products were validated. Station records were validated through discussion at team meetings. Records for field activity, sample result, and QA result were validated as described in Sections 2.6.1 and 2.6.2.

2.6.1 Process

Results for field activity, sample result, and QA result were evaluated holistically by comparing original documentation to finished data product. This approach discovered data errors regardless of the source (e.g., original data entry, computer script, lookup table typos, supplementation entry, etc.). By comprehensively evaluating the data products, they were demonstrated to accurately represent the true values of the original data.

Validation occurred through a combination of automation and staff review. The script "LWQD.verify.FINAL.R" performed the following functions:

1) Loaded lookup tables and crosswalk tables.
2) Loaded finalized tables for field activity, sample result, and QA result.
3) Randomly selected batch IDs in the data set for validation. Review target was 15% of records, with a minimum of one batch and a maximum of seven batches.
4) Subsetted tables to data in the selected batches.
5) Generated supplemental files to aid human review. Supplemental tables provided descriptions and related records to contextualize the records selected for review.
6) Exported subsetted tables and supplemental files for human review.

Team members with QA expertise reviewed the exported files using original documentation and supplemental files and following the process documented in the project file "CGB-157 LWQD Project Workbook" tab "FINAL Verif. Process". Briefly, staff reviewed every data element in all three data products. Any discrepancies between the original documentation and the data products were noted in a Bug Log.

2.6.2 Findings and Corrections

Bug Log entries recorded the file with the discrepancy between original documentation and data product, a description of the discrepancy, and initials of the reviewer. Each entry was reviewed by a staff member with QA and database expertise who compared the original documentation to the flagged data element and determined a correct resolution for the entry. A description of the research, correction (if any), staffer initials, corrector initials, and discrepancy type ID were entered into neighboring columns.

Over 400 entries were made to the Bug Log. About 63% of entries were not actual data errors; they were differences in how primary documentation named samples and handled results compared to the new database. For example, the original documentation recorded one contamination result as "<2RL", but the new database recorded the same result as "300" (the reporting limit of the sample which was below quantification) and the contamination limit as "600" (twice the reporting limit). Another common example stems from the updates to analytical method names. During the new database development, a crosswalk was developed to relate the analytical methods entered with

results to the correct name of the published method. In the original documentation, a method might be named "SM3500Se fluor" or some other name used by the laboratory. In the data products, these colloquial names were updated to the correct, published name ("SM 3500 Se-C"), generating a discrepancy although not a data error. No corrective action was taken for these Bug Log entries.

Another 21% of Bug Log entries was due to errors in the original entry to the database. These errors varied widely. Some examples are failing to enter a QA comment or entering an incorrect analytical method, sample type, or lab report number. For these Bug Log entries, the correct value was determined from original documentation, the data element was corrected in the historical database, the updated historical data was downloaded, and the final data products were easily re-generated by re-running the computer script. While these errors could have been fixed in the finalized data product, this process was used for two reasons. First, some of these errors trickled to multiple data elements in the final data products so fixing them once in the historical data assured that all finalized data was updated. Second, this approach was resilient to other processes. Once fixed in the historical database, the data element was permanently fixed. If the error was fixed in the data product, it would need to be fixed again when the data products were re-generated in response to script updates or other Bug Log entries.

About 4% of errors were due to problems with the computer scripts. For example, the wrong data selection criteria were initially used to fill in the units of dissolved oxygen results. A more complex example involved results for radioactive analytes. For these analytes, the historical database housed their minimum detection areas in the reporting limit field because it lacked a detection limit field. The script corrected this issue by identifying affected analytes and moving their reporting limit values to the new detection limit field. However, the script incorrectly identified the chemical "uranium" as the analyte "uranium radioactivity" and incorrectly moved the reporting limit. All script errors were corrected in the scripts. The final data products were re-generated by the scripts and then checked to confirm the error had been resolved.

The remaining 12% of errors were due to unsystematic sources. For example, an error was found in the analyte name "Nitrite (as NO2)" that was caused by a typo in the analyte lookup table for the new database. This issue was fixed by updating the lookup table in the new database, downloading the updated table, and re-generating the final data products using the scripts. One error was found from a data entry staff member filling the wrong value to the accuracy supplementation file. This error was addressed by correcting the data element in that accuracy file.

After all Bug Log entries were evaluated and resolved, the corrected historical tables were downloaded again and the scripts were run to re-generate the final data products with the best possible data. By using automated scripts, the final data products were quickly re-generated with minimal staff time and 3-25 minutes of computer running time for a data set (depending on data set size).

3.Data Resources

3.1 File Location and Size

Data from this project is located on Reclamation servers at "\\ibr2mpfs002.bor.doi.net\MPRFS002\DEA\2019-S&T Project - Leveraging Historic Data\Deliverables". The total folder is 250 MB. The PowerPoint presentation of this project is 237 MB. The PowerPoint slide deck is 10 MB. The project workbook "CGB-157 LWQD Project Workbook" is 2.6 MB. Computer scripts are listed in Table 4.

Table 4. Table of computer scripts discussed in the report.

File Name	File Size (KB)
stn.clean-up.reviewed.list.R	22
stn.duplicate.dissent.R	3
LWQD.stn.GET.elevation.R	3
lat.long.parser.R	5
LWQD.find.weird.field.result.R	10
LWQD.field.comment.R	7
LWQD.produce.FINAL.file.R	111
LWQD.find.weird.chem.result.R	18
LWQD.batch.R	13
LWQD.clean.Access.QA.results.R	5
LWQD.verify.FINAL.R	9

3.2 Software

This project used R (4.1.0), RStudio (1.2.1335), and Git (2.35.1.windows.2).

3.3 Point of Contact

Questions about this project can be directed to Laurel Dodgen (LDodgen@usbr.gov, 916-978-5038) or CGB-157's branch chief, Dan Deeds (DDeeds@usbr.gov, 916-978-4467).

3.4 Keywords

Data management, Open data, Reproducible workflow, Automation, R software, RISE compatibility

4.Discussion

4.1 Gap Records

Data sets were prioritized for migration by considering stakeholder input, geographic area, and data set size and utility. Consequently, currently collected investigations with large data sets were prioritized because the data was wanted by stakeholders, timely for current resource management decisions, and useful for data analysis. However, an unexpected problem arose from this decision. Since the new database was not in full implementation when this project was started, gap records were created representing data generated after this project started but before new data was entered directly to the new database. These gap records will be addressed with future standardization and supplementation that will follow the developed process. The gap records will be compared to all records so that data are flagged by comparing to a whole population of data, instead of a subset that may be skewed.

This project decision created slightly more work but did allow the highest value data sets to be more quickly available to stakeholders. The advisability of this practice for future data sets will depend on the data content and stakeholder need.

4.2 Cost of Data Migration

Due to staffing constraints, this project addressed about 2/3 of CGB-157's historic data. The remaining data is planned to be migrated soon using this developed process. About $20 million of government data was standardized and supplemented at a cost of about $500,000, demonstrating the high cost of preserving and improving data. Migrating the remaining historical data is likely to cost a proportional amount (about $250,000) or slightly more. Some of the remaining historical data is stored in an alternate historical database, and some processes and scripts may need to be re-developed to address it.

While the cost of data preservation remains, the data improvement costs will largely be avoided in the future due to the good data standards enforced by the new database. Several features contribute to the good data practices. The new database implements rules and constraints that minimize or prevent incorrect or duplicated data entry. The new database supports automatic upload of machine-readable files after testing for data compliance. This step greatly reduces staff time entering data, while also enhancing data quality by eliminating transcription errors. Additionally, data in the new database undergoes QA review after entry, assuring that QA experts see and approve the final disposition of the data. Lastly, the database structure follows best data practices that make the data easily updated, managed, and migrated in the future.

4.3 Value of Automation

Automation was a critical component of successfully performing this project. Completing some tasks by computer scripts saved considerable staff time, allowed staff to focus on high-value

supplementation tasks, and enhanced data quality by supporting process reproducibility. While the exact processes and scripts used in the project would need updating to be relevant to other Reclamation projects, the principle of thoughtful automation is generally applicable.

The large data set BHP was standardized and supplemented as part of the new database development. That process was completely performed by humans using Excel functions (e.g., summary statistics, pivot tables) and visual inspections. While these are valuable tools, the insufficient documentation for that effort meant that it was not reproducible or reversible. Overall, the process improved that data and ended up identifying and correcting errors in about 1% of the BHP records. But the process could not be reliably repeated for other data sets and could not be reversed if a process problem was identified.

When this project implemented a reproducible workflow through automation and documentation, it enhanced the quality of the data sets and saved staff time. Review of the migrated data sets show that the scripts were able to flag 6-10% of data records for human review. Of flagged records, about half needed corrections. This approach saved substantial staff time by allowing staff to skip correct records and focus on problematic records. For example, one investigation family had 62,325 records. The script flagged 9.9% of records for human review, of which 51% needed corrections. The number of corrected records amounted to 3.3% of the data set. A combined process of automated and human review found and corrected more errors than the human-only review process.

Automating workflow also allows the project to flexibly adapt to problems or changing project needs. For example, errors were found in the lookup tables for both analyte and matrix during this project. While this would have been a major task to manually update all instances of the affected values in all the finalized files, this was a very minor task with automation. The lookup tables were corrected, the scripts were run, and the final files were automatically re-generated with all affected values corrected. Human correction would have taken considerable time and a typo by a staffer performing this task might never be found due to the random nature of human mistakes. A second example is the implementation of batch group. While the data fields for batch ID, batch year, and batch collection number were implemented very early in the development of the new database, batch group was implemented very late after testers discovered a problem with batching. Adding this new data element to final files was very simple; a few lines of code were added and the script rerun to produce final files with the correctly added data element.

Automation should be incorporated into a project in a thoughtful, systematic way. After project goals and tasks were identified, all tasks were evaluated to determine their suitability for automation. Suitability was assessed by the effect of automation on data quality, but also staff time. For this project, a staff member was already proficient at coding in R so automating tasks represented a significant time savings. If the project team lacked coding experience, the project lead would have evaluated options including outsourcing coding tasks, training staff in coding, and automating fewer tasks.

Automation comes with unique documentation needs. Just as a standard operating procedure may be updated and has version numbers, a script may be updated. The versions of a script need to be tracked in a systematic manner. This project used the version control system Git to track script versions. Git tracks changes to documents, timestamps version, and allows previous versions to be viewed or restored. This tool allows code to be developed cleanly, without keeping old versions of a script or keeping old code in a script "just in case" it's needed again. It also allows traceability in the

project process. If there is a question about how a particular document was generated at a particular time, the version of the script at that time can be referenced with Git.

4.4 Data Integrity

Preserving, enhancing, tracing, and documenting data integrity is crucial to a data management project. This project stewarded data in several ways:

- Preserved original data from changes
- Documented process steps
- Documented changes to the data
- Automated tasks for reproducibility and error-prevention
- Versioned documentation with built-in Google and Microsoft features
- Versioned scripts with Git
- Validated final data products against original data

Any data management project should form a data integrity plan before work is performed.

4.5 Project Outcome

From November 2018 through September 2022, CGB-157 spent about 3,500 staff hours on the standardization, supplementation, and migration of about 350,000 environmental records. This project was an act of stewardship and investment in that environmental data. Data was formerly split between the historical database and the new database, requiring users to navigate to both locations to access data. In that condition, CGB-157 would spend about 35 hours each year navigating between the databases to find the needed data. CGB-157 would spend an additional 110 hours each year facilitating data searches from external stakeholders. These staff resources are saved with data migrated to the new database. Data in the new database also has an enhanced pipeline to RISE. The new database already has a crosswalk that relates CGB-157 database terms to RISE terms. By migrating data into the new database and using that single overarching crosswalk, about 200 staff hours are saved from doing individual crosswalks for each data set. Additionally, future data updates can happen automatically without need for staff time. Overall, this project made data more accessible to resource managers, more discoverable by the public, and more useful for environmental evaluations, which supports Bureau of Reclamation's mission and public trust.

References

[1] Memorandum - Transparency and Open Government. (2009). *M-09-12*.
[2] Memorandum - The Open Government Directive. (2009). *OMB M-10-06*.
[3] Executive Order - Making Open and Machine Readable the New Default for Government Information. (2013). *Executive Order 2013-05-09*.
[4] Memorandum - Open Data Policy – Managing Information as an asset. (2013). *OMB M-13-13*.
[5] Digital Accountability and Transparency Act. (2014). *Public Law 113-101*.
[6] Open Government Plan 3.0. (2014). *U.S. Department of the Interior*.
[7] Open, Public, Electronic, and Necessary (OPEN) Government Data Act. (2019). *Public Law 115-435*.
[8] Dodgen L., Stack R., Oliveira T. (2021). 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*.