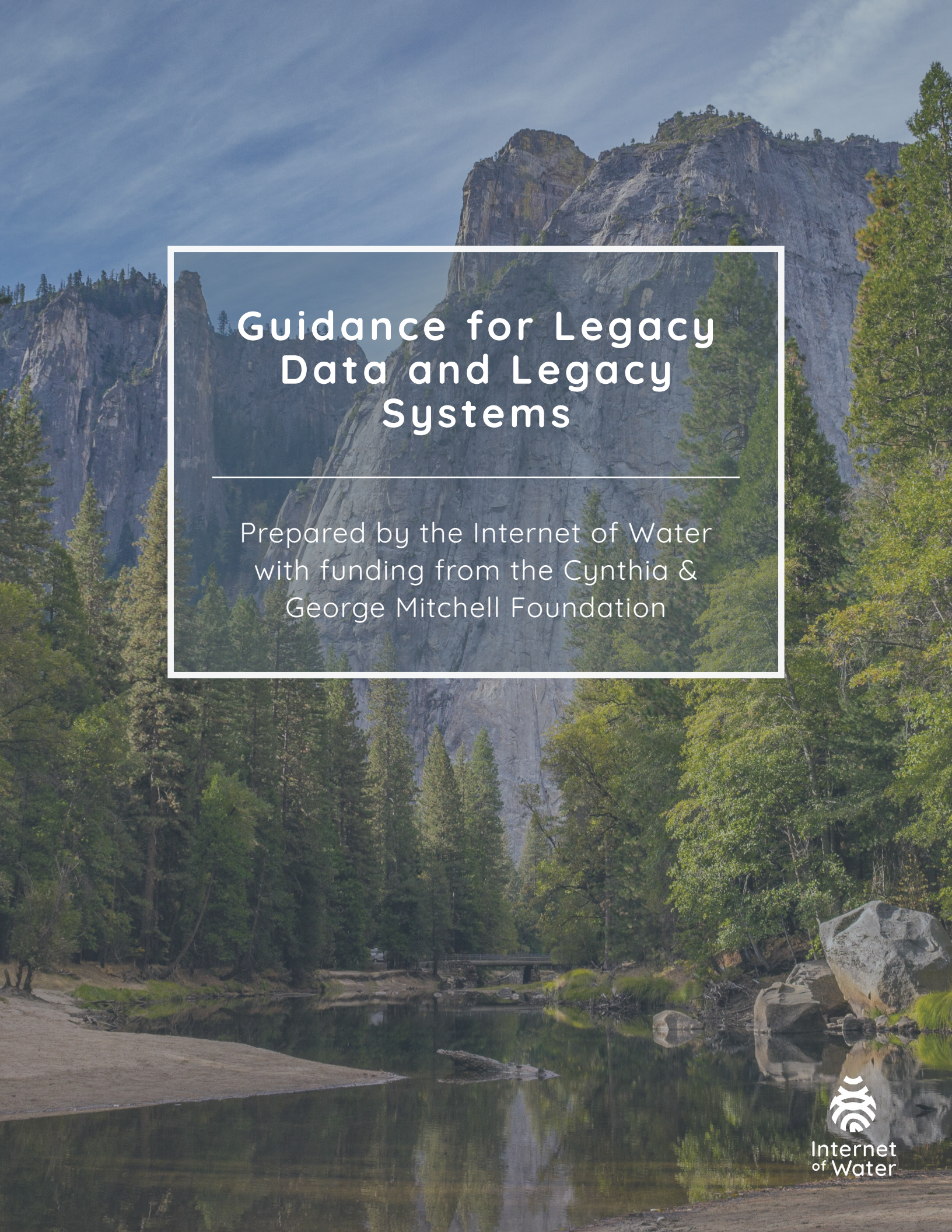

Guidance for Legacy Data and Legacy Systems

Prepared by the Internet of Water with funding from the Cynthia & George Mitchell Foundation

Internet of Water

Contents

INTRODUCTION

In Texas, surface and groundwater resources are governed by separate regulatory frameworks and managed by different entities. Texas does not have the capacity to readily understand the interaction of surface and groundwater flows because they do not have the ability to integrate surface and groundwater (SWGW) information, a key factor for effective water resources management. Instead, surface and ground waters are treated as separate resources and the water data on SWGW interactions is of limited use in decision-making due to fragmentation and accessibility issues. And yet SWGW interactions directly impact water resource availability and quality. Over-pumping of aquifers affects groundwater baseflow and springs, which in turn affect surface water resources and environmental flows. Direct discharges from wastewater treatment plants to downstream losing streams impact groundwater quality, especially in the state's karst terrain. Future development and increasing demand for water resources will put even more pressure on the state of Texas to understand SWGW interactions.

In the face of increased growth, pumping, and wastewater discharge, the public, regulators, and policymakers, especially in the Hill Country and areas around growing urban centers, will be severely limited in their ability to sustainably manage water resources. The guidance provided in this document is designed to address a critical barrier to understanding and effectively managing SWGW interactions: legacy data and legacy systems. Data held in legacy systems are difficult to find, access, and integrate with other data, and because of this, these data are rarely used in decision-making processes. Improving the findability and accessibility of legacy data is a foundational step in improving water management decisions about SWGW interactions in Texas. However, this same guidance can be applied to address challenges with legacy data and systems for a variety of other applications.

MODERNIZING LEGACY DATA SYSTEMS

The solutions to modernizing legacy data and systems vary, but both legacy data and systems present barriers to greater organizational efficiency, improved and timely decision-making, and the ability to maximize opportunities that data analytics and big data provide. In each case, best practices should be considered:

- Ensure that data converted from legacy systems or formats are compatible with new or modern systems
- Take specific measures to ensure the accurate transfer of data
- Define and implement data quality standards
- Aim for consistency across all systems

As important as legacy data can be to water management, efforts to modernize legacy data and systems can be stymied by common challenges around data migration and conversion.

Legacy data: data stored in old or obsolete systems that are difficult to access, process, or integrate with other solutions.

Legacy systems: outdated computing software and/or hardware that are still in use. These systems still meet the needs for which they were originally designed; however, they cannot interact with newer systems.

Types of data that may be included in legacy data and legacy systems are paper files and records, pdf reports, videos, photos, audio files, physical samples, hard copy maps, etc. These data are often valuable and frequently referenced for a variety of purposes. However, due to their lack of findability and accessibility, can be costly and time-consuming to process and use effectively. In addition, because they pre-date current, open-data policies, standards, and formats, they are at risk for permanent damage or loss.[1]

1 See USGS guidance: https://www.usgs.gov/centers/cdi/science/developing-a-usgs-legacy-data-inventory-preserve-and-release-historical-usgs?qt-science_center_objects=0#qt-science_center_objects

Common challenges in the data migration and conversion process include:

- Difficulty merging unstructured and structured data: Unstructured data, such as video, audio, or image files do not fit easily into a data structure based on rows and columns, such as spreadsheets. Structured data is specific with pre-defined formats.
- Conerns about time and monetary investment: The migration and conversion of legacy data is often large in scale and scope, possibly spanning many decades or across many agencies or divisions, creating concerns about the time and monetary investment needed to successfully execute such initiatives.

Data migration: moving data between devices, locations, or systems.

Data conversion: a transformation to updated or new applications through the ETL process (extract, transform, load). Extraction refers to moving data from its original source, whether that is another database or an application. Transform refers to data clean up, duplication, and integration. And Load refers to its transfer to the target database or application.

CONDUCT A LEGACY INVENTORY

A successful data migration and conversion project is dependent upon a robust and detailed plan. However, in order to develop a plan that can be successfully executed, agencies must first conduct a legacy inventory. To conduct a legacy inventory, identify and catalog all existing legacy data or legacy system components.

DEVELOP A PLAN

After conducting a legacy inventory, follow the steps below to develop a migration and conversion plan:

- Prioritize data and/or systems for conversion
- Develop specification and standards for data selection for integration into the new or modern system

- Identify data conversion formats (SGML, XML, HTML, etc.)
- Identify the extent of digitization, if needed.
- Assess the compatibility of legacy data and systems with the new system
- Develop a timeline for the data conversion process
- Specify costs associated with data conversion
- Develop frequency of data conversion

In Texas, other organizations have identified three activities to find sources for data to be converted: 1) a pre-2002 review including Slade and others (2002)[2]; 2) a post-2002 review of the literature, including U.S. Geological Survey, Texas Water Development Board, Texas Commission on Environmental Quality, and Texas groundwater conservation districts; 3) other post-2002 sources as needed.

ARTICULATE THE BENEFITS

In addition to the development of a migration and conversion plan, it is helpful to articulate the value of data conversion initiatives to leadership. This likely includes benefits from these areas: a) employee efficiency; b) sustainability; c) security.

Improved employee efficiency often accompanies data migration and conversion due to enhanced data processing and management. Benefits related to efficiency may be direct, such as reducing the amount of time employees spend processing data, but can also be indirect such as reducing an organization's over-reliance on a select number of employees who know how to manage or maintain legacy systems. Legacy systems may also mean longer and more extensive onboarding processes as new employees are less and less familiar or trained on existing legacy systems.

The sustainability of systems is often improved with data migration and conversion. Many legacy systems cannot be integrated with modern cloud or SaaS solutions, making it difficult to incorporate new tools and programs and often requiring specialized coding. This impacts the ability to easily maintain legacy systems. This lack of integration with modern systems frequently results in data silos within organizations, where different departments across the organization cannot fully access the data they need.

And finally, concerns over cyber security continue to grow, and for good reason. Outdated systems are often prime targets for cybercriminals because they can exploit the weak points to gain access. As news headlines showcase significant and costly

2 See Texas Water Development Board Groundwater Database (GWDB) Reports: https://www.twdb.texas.gov/groundwater/data/gwdbrpt.asp

cyber-attacks, leaders have become more aware of the need to reduce the vulnerability to cyber-attacks within their own organizations.

ESTABLISH A WORKFLOW

Establishing a data cleaning and quality workflow is essential to ensure the efficient migration of data and the quality of that data once migrated. This includes data mapping activities that extract data fields from legacy data and systems and match them with target fields in the destination system. The template developed during the data mapping process will be used throughout the migration and conversion process to ensure the proper transfer of data to the new system.

There are three main data mapping techniques: a) manual data mapping that requires an IT professional to hand-code or manually map the data source to the target schema; b) schema mapping, which is a semi-automated strategy that employs mapping software to establish a relationship between a data source and the target schema; and c) fully-automated data mapping that uses a drag-and-drop data mapping UI that allows non-technical users to carry out mapping tasks in an easy-to-use format. Once the data mapping is complete, a small test batch should be executed to test the transfer.

A final task in the workflow development should be a mechanism to identify and remove duplicate data.

SCAN AND DIGITIZE

Legacy data can exist in many formats: paper files and records, pdf reports, photos, audio files, physical samples, hard copy maps, etc. Converting these data can transform them into usable digital assets. This process frequently involves both scanning and digitizing the data.

Scanning data is the process of document imaging in which a hard-copy document is scanned as an image and converted to a digital file. This is a physical process that can be facilitated by batch-scanners which exist as both stand-alone machines as well as parts of photocopier machines. This is often fast and cost-effective. However, the location of such scanned images should be planned in a manner similar to the ETL process mentioned above. Scanned image files should be located in a coherent file system with sufficient metadata and indexing for quick discovery and reference of any particular image. Moreover, stopping at scanning hard-copy documents fails to provide the most benefit from data conversion initiatives.

Digitizing data is the process of transforming information (often from scanned images) from data that cannot be readily processed by computer programs into data that can. For example, a collection of scanned water well logs can be digitized so that its content, including specific well metadata elements, is consistently searchable across documents. In the case of imagery such as aerial photography, scanned image files can be digitized into georeferenced raster data suitable for analysis with other geospatial data products. It is this process that provides the greatest return on investment. Therefore, when possible, the migration and conversion of legacy data should include digitization.

CONCLUSION

When valuable data is trapped in legacy systems organizations are prevented from gaining the important insights that it holds. The modernization of legacy data and systems can transform an agency's ability to understand and manage the resources it is responsible for. In Texas, the modernization of SWGW legacy data and systems would give the state the tools it needs to develop a more complete understanding of SWGW interactions. Only with this understanding will Texas be able to meet the challenges presented by increased development and more frequent and prolonged droughts. These same challenges with legacy data and legacy systems exist across the country with different water resources and different water management challenges. We hope this guidance can serve as a tool for a variety of water management agencies to begin the legacy data modernization process.