Spencer Naar

FINAL REPORT

Data and Metadata Profile

Selected Repository:	USGS Water Mission Area NSDI Node
Dataset File:	https://water.usgs.gov/GIS/dsdl/sir2013-5079_Groundwater_Depletion_Study_Files.zip
Dataset Link:	https://water.usgs.gov/lookup/getspatial?sir2013-5079_Groundwater_Depletion
Dataset Name:	*Groundwater depletion in the United States (1900−2008)*
Dataset Contributors:	E.A. Achey, S.M. Feeney, D.P. McGinnis, and J.J. Donovan assisted with analyses and calculations for some of the aquifer systems. U.S. Geological Survey colleagues G.N. Delin, D.L. Galloway, E.L. Kuniansky, and R.A. Sheets provided helpful review comments. D.J. Ackerman, E.R. Banta, J.R. Bartolino, L.M. Bexfield, J.B. Blainey, B.G. Campbell, A.H. Chowdhury, B.R. Clark, J.S. Clarke, J.B. Czarnecki, R.B. Dinicola, J.R. Eggleston, C.C. Faunt, R.T. Hanson, R.E. Heimlich, C.E. Heywood, G.F. Huff, S.K. Izuka, L.E. Jones, S.C. Kahle, M.C. Kasmarek, Eloise Kendy, A.D. Konieczki, A.L. Kontis, S.A. Leake, Angel Martin, Jr., J.L. Mason, D.P. McAda, E.R. McFarland, V.L. McGuire, Jack Monti, Jr., D.S. Oki, S.S. Paschke, G.A. Pavelis, D.F. Payne, M.D. Petkewich, J.P. Pope, C.L. Stamos-Pfeiffer, G.P. Stanton, S.A. Thiros, F.D. Tillman, B.E. Thomas, and J.J. Vaccaro kindly provided information about computer simulations, model results, depletion analyses, and (or) review comments for specific areas. S.A. Hoffman provided valuable assistance with Geographic Information System (GIS) tools and map preparation. This work was supported in part by funding from the U.S. Geological Survey's Office of Groundwater and Groundwater Resources Program.

Data | This profile looks at long-term cumulative groundwater depletion data in the United States between the years of 1900 and 2008. The data is published by the U.S. Geological Survey under the Department of the Interior and the dataset is publicly accessible on the USGS Water Mission Area NSDI (National Spatial Data Infrastructure) Node page. Data was collected from 40 separate aquifer systems and 1 land use area to estimate overall rates and the magnitude of change in the volume of groundwater stored in the earth's subsurface. The most reliable method for calculating these estimates was direct measurements of water-level changes in the aquifer systems.

The dataset is provided as a .zip file, a compressed archive which can be opened with Windows XP. The file is organized into *database* and *spreadsheet* folders. The *database* folder is organized by *aquifer (365 files)* and *basecamp (27 files)* folders and includes a number of different file formats. Database files (.dbf) can be read using dBase, Excel or Access; Plain text files (.prj) are used for coordinates and projection data and can be read using ESRI ArcGis. .sbn, .shp, .shx, and .sbx are part of ESRI's shapefile format and can be read using ESRI software, a geographic information system. Content from these files often provide position and speed data and are collected by outdoor, waterproof GPS navigation devices. Extensible Markup Language (.xml) data can be viewed using a web browser however Oxygen XML developer or another xml editor would be helpful. The *spreadsheet* folder has six .xls formatted spreadsheets with groundwater depletion average rates from 1961-1970 and 2001-2008, depletion volume data from 1900-2000 and 1900-2008, Map Data and High Plains data and can be accessed using Microsoft Excel. The dataset does not come with any usage restrictions. It is intended for public access and no license information is provided.

The key stakeholders for the data are i) its *creators/managers*: employees of the U.S. Geological Survey and the larger Federal Department of the Interior ii) its *users*: environmental scientists and researchers, data analysts and estimators of groundwater depletion iii) *individuals & organizations*: academics and scholars who monitor and study water resources (UNM Water Resource program for example) and nonprofit environmental organizations (NMWCA for example) iv) *policy makers* who may enact legislation around water rights or access based on the data and projections and v) the *larger public* interested in water and environmental issues.

Metadata | The data comes with relatively comprehensive metadata focused on *identification, data quality, spatial data organization, spatial reference, entity and attribute, distribution and*

metadata reference information. The metadata is provided on the USGS Water Mission Area NSDI Node page. The original metadata URL was not accessible from the data.gov page. The *identification information* provides details on citation, description of the dataset article including abstract, purpose and supplemental information, time period of content, status, spatial domain, keywords by theme and location, access constraints, use constraints, contact information, graphic and data set credit, security information, and details on the native data set environment. The *data quality information* centers around attribute accuracy, logical consistency, completeness, positionality and lineage. The *spatial data organization information* focuses on spatial reference methods and descriptions. The *spatial reference information* looks at horizontal and vertical coordinate system definitions. *Entity and Attribute Information* provides descriptions; *Distribution Information* focuses on contact information, distribution liability and ordering; and the *Metadata Reference Information* provides information on standards. The metadata is structured according to the FGDC Content Standards for Digital Geospatial Metadata. The standard is considered out of date but is still used across USGS data centers for defining digital geospatial data. In 2010 the FGDC encouraged federal agencies to transition to ISO 19115, an internationally-adopted schema for describing geographic metadata, but little movement has been made on this front.

Publications | The primary publication based on the dataset is *Groundwater depletion in the United States (1900-2008)*, a Scientific Investigations Report by Leonard F. Konikow and published in 2013 by the U.S. Geological Survey. At least 82 articles cite this article and dataset in their scholarship. These resources were identified by locating the article in UW Libraries Search, selecting *references* and clicking on *sources* which have cited this article.

Enrichment | To improve users' ability to discover the data set in a repository environment, I would assign new descriptive file names which better reflect the contents of each file as well as add tags or descriptors to the file metadata to provide additional context. I would also include information on the overall organization of the dataset. While there are designated folders, there are few details on how the files are arranged and what they specifically capture. This would be useful metadata while searching and when trying to quickly review data contents. To assist someone unfamiliar with the data and to make use of the dataset for new purposes I would add additional descriptions on the process for creating the data. While some lineage and provenance metadata are included I would provide assumptions about the primary input data and limitations around calculations estimating groundwater depletions. I would also contribute more context around data compatibility given that data was collected from over 40 different aquifer systems using different methods over a period of a century. Differences in samples as a result of equipment and procedure should be addressed, especially if it is anticipated that the dataset will be used for new research and scholarship purposes.

References

DataOne Best Practices Primer, accessed 26 January 2023, https://dataoneorg.github.io/Education/bestpractices/

Groundwater depletion in the United States (1900-2008): Metadata: USGS Water Mission Area NSDI Node, accessed 26 January 2023, https://water.usgs.gov/GIS/metadata/usgswrd/XML/sir2013-5079_Groundwater_Depletion.xml#stdorder

Konikow, L.F., 2013, Groundwater depletion in the United States (1900−2008): U.S. Geological Survey Scientific Investigations Report 2013−5079, 63 p., http://pubs.usgs.gov/sir/2013/5079.

Research Data Alliance Metadata Directory, accessed 26 January 2023, http://rd-alliance.github.io/metadata-directory/

Water Resources Groundwater Software, U.S. Geological Survey. Access on 3/7/23 https://water.usgs.gov/software/lists/groundwater

Repository Profile

Selected Data Repository	USGS National Water Information System
Repository URL	https://waterdata.usgs.gov/nwis
Re2data Repository URL	https://www.re3data.org/repository/r3d100011035

The repository selected for the dataset is the USGS National Water Information System. This repository stores data on the quantity, quality, use, distribution, and movement of groundwater and surface water for all fifty US states. In addition to storage and preservation the repository serves to disseminate data to the public, State and local governments, public and private utilities, and other Federal agencies involved with managing water resources which is in alignment with the nature of the dataset. The data repository is closed to general data submissions but access to the research data repository is open to the public.

The repository accepts many different types of data including comprehensive information on data collection "site characteristics, well-construction details, time-series data for gage height, streamflow, groundwater level, precipitation, physical and chemical properties of water and water use data, peak flows, chemical analyses for discrete samples of water, sediment, and biological media" (Additional system background, 2011). The collected data fits into broad categories of surface water and groundwater but also water usage and water quality which includes figures on temperature, pH, nutrients and pesticides (About the USGS Water Data, 2021*)*. All data is contributed by researchers and scientists at the United States Geological Survey (USGS) and all information and guidance on what the repository provides to the data submitters specific to the "submission information package" is not publicly viewable. All consulting details and metadata structure and standards information for the submitter appear to

be internal. All data submitted is received at the National Oceanic and Atmospheric (NOAA) Wallops Command and Data Acquisition Station (WCDAS) in Wallops Island, Virginia. Data is then sent to the USGS Water Science Centers for data processing. Data is backed up on receivers located at the Earth Resources Observation and Science (EROS) Data Center in South Dakota (About the USGS Water Data, 2021*)*.

Data is publicly viewable in the repository by category of data and by geographic area. There does not appear to be a login required to download data and there are multiple mechanisms in place for accessing data within the repository. The repository has six "production services" to download and retrieve current condition data, daily values, hydrologic sites, groundwater levels, water quality and statistics. Users can also view graphs of current conditions, water levels, and water quality; tabular output in HTML and ASCII tab-delimited (rdb) data files and summary lists for selected sites. While .xml and .json are common files for sharing data, .rdb data files are the primary output format for this repository. That being said, the new Water Services site does allow for instantaneous, daily values and groundwater levels data to be downloaded in the .xml format and data-friendly formats, such as Microsoft Excel spreadsheets and .kml, are being worked on to enable integration with Google Maps, Google Earth and GIS formats (About Output Formats, 2021). It is not possible for the entire repository to be downloaded in a single action due to the volume of data, however data can be acquired slowly over time across geographic areas. The only access restriction is that a single data request can not exceed 100,000 site records, "a limitation intended to prevent any one data consumer from unduly affecting other users of the system" (Automated Retrieval, 2021) The repository has an entire page dedicated to automated retrieval of its data which is significant because the designated community relies on active, current and up-to-date water data for research and policy decisions.

The repository does display metadata using the same content standard as the selected dataset which is the FGDC Standards for Digital Geospatial Metadata. USGS provides a number of resources on metadata creation and following specific standards as part of the data description and input process. In terms of the "dissemination information package" the repository provides a number of custom options for output information and formats depending on the needs of the designated community.

One thing that was interesting to me was the repository has a notification service to inform users of the automated data retrieval community about planned outages, unexpected system problems and changes to the system that might affect third party sites and research. Since water data is collected at millions of location sites around the country and maintained by different USGS Water Science centers, data can suddenly become unavailable or data output formats can quickly change. This communication feature which proactively engages its stakeholders highlights the credibility and integrity of this repository and its recognition of the importance of its designated community which relies on its repository for timely, accurate and quality data.

References

About Output Formats (November, 2011). USGS National Water Information System: Help System. https://help.waterdata.usgs.gov/faq/output-formats

About the USGS Water Data for the Nation site. (November, 2021). USGS National Water Information System: Help System. https://help.waterdata.usgs.gov/faq/about-the-usgs-water-data-for-the-nation-site

Additional System Background. (November, 2011). USGS National Water Information System: Help System. https://help.waterdata.usgs.gov/faq/additional-background

Automated Retrievals. November, 2021). USGS National Water Information System: Help System. https://help.waterdata.usgs.gov/faq/automated-retrievals

Content Standard for Digital Geospatial Metadata. (CSDGM). (1998). Federal Geographic Data Committee. https://www.fgdc.gov/metadata/csdgm-standard

FAQ. (November 2011). USGS National Water Information System: Help System. https://help.waterdata.usgs.gov/faq

USGS National Water Information System. (March, 2022) Registry of Research Data Repositories. http://doi.org/10.17616/R3S333 last accessed: 2023-02-11

USGS Water Data for the Nation. (February, 2023). USGS National Water Information System https://waterdata.usgs.gov/nwis

Additional Information

Recommended Data Citation (Based on static USGS data citation guidelines**)**	Konikow, Leonard F. 2014, *Groundwater depletion in the United States (1900-2008)*: U.S. Geological Survey digital data release, accessed February 21, 2023 at: https://water.usgs.gov/lookup/getspatial?sir2013-5079_Groundwater_Depletion

Digital Preservation | Ensuring the enduring value and authenticity of data as well as long-term access to data files is a constant challenge. File format obsolescence and technology failure are a few of the many risks at play. A general guidelines document for the preservation of digital scientific data at USGS is available to access but nothing specific to the groundwater depletion dataset is provided in terms of preservation metadata. The data files of this project are in a variety of formats and despite most being in acceptable formats for long-term access, a number of them could be migrated to high digital preservation level formats. In terms of the structured text files, xml is considered the best preservation format for future use. The Microsoft Excel files (.xls) could be converted to xlsx or csv files for better tabular data preservation. Database files (.dbf) are moderately preserved for future use however if there is a possibility of migrating materials to SQL DDL (.sql) that would be ideal. In terms of geospatial data, ESRI Shapefiles (.shp, .shx, etc) are also moderately preserved but it would be ideal to migrate these files to Geographic Markup language (.gml) or GeoTIFF (.tiff) files for greater confidence in their

long-term preservation and access. In terms of software, the native dataset environment used for this dataset was Microsoft Windows 7 Version 6.1 (Build 7600) ; Esri ArcGIS 10.2.0.3348. Windows 11 is the latest major release of Microsoft Windows NT operating system (as of October 2021) and ArcGIS 10.8.1 is the current release of ArcMap (supported through March 01, 2026). As software development continues to advance there is a possibility that newer versions will not support older file formats unless they are migrated to recommended preservation level formats.

Copyright License Statements | According to the USGS website, "USGS-authored or produced data and information are considered to be in the U.S. Public Domain" which is considered a Creative Commons Zero license. Using CC0 one can waive all copyrights and related or neighboring rights….including database rights and rights protecting the extraction, dissemination and reuse of data. There are no use restraints for this dataset. Acknowledgement of the U.S. Geological Survey is appreciated in products derived from their data (Creative Commons, 2023).

Ethical Issues | Data ethics should be kept at the core of research data management services including collection, storage and access. This dataset does not require any efforts to ensure data anonymization for privacy reasons. The dataset is strictly concerned with environmental measurements and calculations. There are however ethical issues at stake as data collection for this project could potentially impact sensitive environmental habitats, flora and fauna populations and endangered species. The USGS does provide an Environmental Management Policy which states that USGS is "committed to protecting the environment through complete compliance with environmental laws, regulations, and outstanding efficiency in the conduct of [their] operations." This includes complying with environmental laws and regulations, minimizing impact of operations through regular evaluation and restoration, and conducting

audits to measure environmental performance and establish accountability to correct deficiencies, to name a few. This dataset and selected repository are also transparent in advocating for data sharing and have clearly communicated any license and access restrictions so users of their data are aware of how data should be reused and redistributed.

References

Creative Commons (2023). Accessed on March 7, 2023: https://creativecommons.org/choose/zero/

Environmental Management Policy Statement. *USGS*. Accessed on March 7, 2023: https://www.usgs.gov/environmental-management-policy-statement

USGS Guidelines for the Preservation of Digital Scientific Data. Accessed on March 7, 2023: https://d9-wret.s3.us-west-2.amazonaws.com/assets/palladium/production/s3fs-public/atoms/files/USGS%20Guidelines%20for%20the%20Preservation%20of%20Digital%20Scientific%20Data%20Final_508compliant.pdf