# Data processing handbook of WDC-RRE



**Institute of Geographic Sciences and Natural Resources Research,**

**Chinese Academy of Sciences**

**March 2016**

# 1 Data Selection and Evaluation

## 1.1 Criteria

**(1) WDC-RRE's data are selected according to the following 4 principles:**

1. WDC-RRE seeks data that have demonstrated importance to the resource and environmental science community as determined by substantive value for research and/or instruction, enduring archival value for research and/or instruction, and uniqueness.

2. WDC-RRE seeks data that support its mission.

3. WDC-RRE seeks to acquire data in core resource and environmental science areas.

4. WDC-RRE seeks data to which current and emerging research questions and techniques can be applied.

**(2) Datasets that meet these principles are further reviewed by WDC-RRE staff. Datasets are accorded a high priority for inclusion in the archive when:**

1. The data are not available anywhere else or are not likely to be available elsewhere in the future.

2. The data are in the public domain.

3. There are no copyright disputes or violations.

4. Copyright owners agree to WDC-RRE's data usage policy.

5. The dataset adheres to the standards for privacy and confidentiality.

6. The data statement file and metadata are complete.

7. The data format complies with that in the data preference list in the data storage policy.

## 1.2 Evaluation Criteria Details

After identifying a dataset given the considerations listed above, WDC-RRE staff applies the following criteria to assess the dataset's priority for acquisition. The following appraisal criteria are applied simultaneously. Data are immediately approved for possible acquisition when there are no concerns lowering the priority of the acquisition. If there are one or more concerns lowering the priority level of a dataset, WDC-RRE considers the potential benefits and costs associated with acquiring the data and acquires, in the short term, only those datasets, which it has the capacity to acquire. Data with lower priority for collection that are not acquired in the short term are either deferred for possible acquisition by WDC-RRE at a later date or referred to another archive whenever possible.

**(1) Data Availability**

1. If a dataset is available at an alternative site at a reasonable price and if there is confidence

that the availability of the dataset will continue over time, WDC-RRE may lower the priority for acquiring the dataset.

2. WDC-RRE may provide links to data available on the Internet as an alternative to physical possession of files, when long-term archival conservation is not compromised.

**(2) Security, Privacy, and Confidentiality Considerations**

1. WDC-RRE requires that studies deposited in the archive meet recognised standards for privacy and confidentiality of studied subjects.

2. WDC-RRE prefers to acquire data that can be available in the public domain.

**(3) Copyright and Other Legal Issues**

1. WDC-RRE prefers to acquire data with discernibly identified owners with explicit or implicit intellectual property rights over distribution of copies of the data to the public through WDC-RRE.

2. WDC-RRE requires that the person, or institution, that has explicit or implicit intellectual property rights over the data being submitted to WDC-RRE agree to WDC-RRE's Data Storage Specification.

3. WDC-RRE requires the "owner" of the dataset to grant permission to WDC-RRE to use the dataset for the following purposes:

①To disseminate copies of the dataset in a variety of media formats

②To promote and advertise the dataset in any public domain (in any form) by WDC-RRE

③To describe, catalogue, validate and document the dataset

④To store, translate, copy or re-format the dataset in any way to ensure its future preservation and accessibility

⑤To incorporate metadata or documentation of the dataset into public access catalogues

⑥To enhance, transform and/or rearrange the dataset, including the data and metadata, for any of the following purposes: protect respondents' confidentiality and/or improve usability of the data

**(4) Data Quality**

1. WDC-RRE strongly prefers datasets that have comprehensive data description files providing ample information on dataset's content features, data sources, and methods of the data acquisition and processing to allow users to assess the quality and analytical reliability of the data.

2. WDC-RRE considers the acquisition of lower quality data, if the data have unique historical value.

3. WDC-RRE prefers data in the most complete and original form, with the exception of data extracts specifically intended for instructional purposes.

**(5) Data Format**

WDC-RRE has clear regulations for data format. For details, please see the WDC-RRE data storage policy. The following four points describe the principles in detail:

1. WDC-RRE prefers data in a readily useable format, accessible to people with a variety of computing and technological capabilities.

2. WDC-RRE prefers data formats that promote easy access and use without compromising their research value.

3. WDC-RRE requires that data files deposited in a raw format be transformable or convertible into formats useable by a variety of software.

4. WDC-RRE prefers data files unaccompanied by value-added software.

## 1.3 Data Sources

Some of the major sources of WDC-RRE data are listed below:

1. Depositors. WDC-RRE acquires data from researchers conducting their research studies. Researchers are seeking a data archive that can make their data available to others and preserve it for future scholars.

2. Funding agency mandates. Many grants have requirements that results of studies funded under these grants be deposited in a public archive. Most of the data in the Thematic Collections are deposited in WDC-RRE under the terms of contracts and grants that funded the data collection.

3. Expert recommendations. WDC-RRE's staff, expert committee, user committee, will continuously monitor data from science and technology projects and academic papers that cause interest and can be recommended for inclusion to WDC-RRE.

4. Series collections. The most recent updates of a large number of serial data collections are automatically added to the archive.

5. Digitisation and data entry. Many historical collections at WDC-RRE were digitised by WDC-RRE staff to create quantitative data files.

# 2 Ingestion

The ingestion phase of the digital content life cycle involves the deposition, acceptance, and enhancement of the content and culminates in (1) the repository specifying a set of materials to be preserved for the long term, and (2) the repository making the resource available with all the information necessary to understand and use it.

At WDC-RRE, ingestion involves the following important data curatorship steps: ①Receipt of Submission ②Data Enhancement ③Data policy file ④Metadata ⑤ Data visualisation

## 2.1 Receipt of Submission

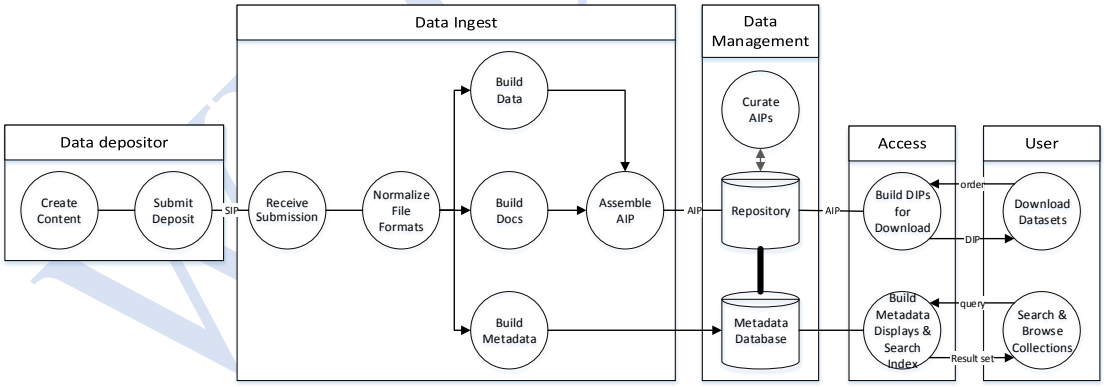
When a data provider agrees to deposit data, there must be checks in place to validate and process the submission. Secure handling of the data is also critical. Many repositories establish an access register to identify and track all submissions.

At WDC-RRE, all data are deposited via an electronic Deposit Form. Files uploaded via this secure system are given unique deposit IDs and moved to the appropriate area for further processing. After data providers complete depositing the data, WDC-RRE sends confirmation of the receipt of materials. Physical submissions that arrive at WDC-RRE on external storage devices such as CD-ROMs or DVDs are immediately copied to a secure location, and the copying process is verified to ensure that all files are transferred properly.

## 2.2 Data Enhancement

The goal of data processing and enhancement is to render data usable to researchers interested in accessing it, after the data have been deposited in a repository. In the resource and environmental sciences, archives add value to data by making them easier to use for secondary analysis. There is a wide variation in archival practices, often depending on the data archival conditions and the goals of a particular repository or discipline.

At WDC-RRE, once data are submitted in a Submission Information Package (SIP), the data pass through a 'pipeline' for processing and enhancement.




The specific steps depend on the unique characteristics of each dataset, but in general, WDC-RRE data processors always perform the following procedures: ①Review data for confidentiality issues; ②Examine the consistency and integrity of a dataset and data description; ③ Convert hardcopy documentation to electronic format; ④ Convert software-specific documentation to PDF/A; ⑤Generate multiple data formats for dissemination and preservation; ⑥Assign unique identifiers; ⑦ Create thumbnails

(1) WDC-RRE uses unique identifiers for identification of data, data description files, and

metadata and to differentiate between different versions of data, data description files, and metadata. All updated information about the data and metadata is recorded in the data description file.

(2) At WDC-RRE, dedicated staff and user committee review the data, data description files, and metadata to ensure that data format and quality comply with the regulations in the data storage policy. One such regulation is that the data description file and metadata should contain all required information. If not, the employee responsible for data storage will be requested to modify the data, data description files, and metadata before resubmission for review.

## 2.3 Metadata

Metadata are critical for an effective use of the data, as they convey information that is necessary to fully exploit the analytic potential of the data. Because it is often impossible for secondary researchers to ask questions from the original data collectors, metadata become the de facto form of communication between secondary researchers and original data collectors. Complete and thorough metadata make it possible to more completely understand a dataset, search for data values including variables, and employ a variety of options to display the data on the Web. Preparation of high-quality metadata can be a time-consuming task, but the cost can be significantly reduced by planning. In view of this, WDC-RRE has formulated metadata standards. These standards were referenced for local and overseas national standards, industrial standards on metadata, and are formulated based on a combination of criteria for creation, storage, and service characteristics of renewable resource data. WDC-RRE data management staff will carry out checks and updates based on metadata information provided by data providers to ensure good metadata quality.

## 2.4 Data Description File

Different from metadata, the data description file is an independent and more comprehensible source of data description. This file is included in the Dissemination Information Package (DIP) to help users fully understand the data. WDC-RRE data management staff will revise and improve the raw data description files provided by data providers according to the technical specifications for data files. The data description file mainly includes eight items, namely dataset/atlas content features, subject/industry scope of dataset/atlas, accuracy of dataset/atlas, dataset/atlas storage management, quality control of the dataset/atlas, sharing and usage method of the dataset/atlas, intellectual property rights of the dataset/atlas and others (optional). For details, please see the technical specifications for data files.

# 3 Data Deposit

WDC-RRE will compress entity data, metadata, data description files, and thumbnails into an Archival Information Package (AIP) for storage and will provide a unique identifier. WDC-RRE has made further additions to its archival storage processes by identifying multiple and varied methods and locations to back up its holdings. WDC-RRE currently maintains three copies of its data (and requires that any off-site backup be encrypted): ①One copy put on tape once a month on-site, with each copy held for 12 months；②One copy held locally；③One copy stored in Alibaba cloud. For details, please see the Data Storage Specification.

WDC-RRE expert committee and user committee will carry out integrated judgement according to the agreement signed by the data archiver and the scientific value of the data to determine the storage category for the data and whether the long-term storage is required.

# 4 Access and Dissemination

WDC-RRE disseminates data among researchers, students, policymakers, and journalists around the world based on its data usage policy. Users may download all data directly from WDC-RRE. Access to data is sometimes restricted and users are expected to adhere to norms for responsible use.

## 4.1 Data Policies of Dissemination

1. Responsible Use of Data

Those downloading data are expected to comply with standards of responsible use. Before gaining access to data, users are asked to read a Responsible Use Statement that says the following:

①The datasets are to be used solely for analysis and reporting of aggregated information.

②The confidentiality of research participants is to be guarded in all ways.

③Anything that can potentially breach participants' confidentiality is to be reported promptly to WDC-RRE.

④The data are not to be redistributed or sold to others without the written agreement of WDC-RRE.

⑤The user will inform WDC-RRE of the use of the data in books, articles, and other forms of publication.

2. Delayed Dissemination

All data archived at WDC-RRE must eventually become available to interested parties. Usually this happens as soon as the process of depositing and archiving is completed. However,

the dissemination of data can be delayed in some circumstances, as described in the following paragraphs.

①Protection of Human Subjects. Investigators and WDC-RRE may delay dissemination when there is a significant risk that research subjects can be identified. This risk may persist even when direct and indirect identifiers have been removed. Often a subset of such datasets can be released with restricted-use and Data Enclave protections.

② Self-Dissemination. Researchers may choose to disseminate datasets themselves, especially while funding is available to support this activity.

③Embargoes. Researchers sometimes ask to embargo data until a book or article is published or a report is written and submitted to a funder. Embargoes will typically last no longer than one year from the time WDC-RRE receives the data. During the embargo, WDC-RRE will process the data in the usual manner, but not release the data to the public. When the embargo is lifted, data would be ready for immediate dissemination.

④Handling of Delayed Data. WDC-RRE will agree to delay dissemination of data for a good reason but will insist that the data provider agree to a release date. Then, WDC-RRE will archive a "preservation-only" copy of the data for general safekeeping and for learning how to work with it, while providing access to available knowledgeable staff members.

3. Data Dissemination

When the data user satisfies the data usage policy requirements, WDC-RRE will provide a DIP for the user to download from the Internet. This package will contain the data itself and data description files.

## 4.2 Citing Data

Professional associations in the resource and environmental sciences are increasingly cognizant of the importance of proper data citation in their publications to encourage the replication of scientific results, to improve research standards, and to give proper credit to data providers. As other peer-reviewed journals and data stakeholders follow suit, consistently applied data citation standards will ensure that research data can be discovered, reused, replicated for verification, credited for recognition, and tracked to measure usage and impact. Accurate citation of data promotes more and better scientific conduct, and we believe that all data stakeholders can do more to improve the data citation. Citing data is straightforward. Each citation must include the basic elements that allow a unique dataset to be identified over time: Title, Author, Date, Version, Persistent identifier. For details, please see the data reference criteria in the data usage policy.

If there are any suggestions or comments about this document, please contact us:

Tel:+86-10-64889048-8006

E-mail: wdc-rre@lreis.ac.cn

Address：11A, Datun Road, Chaoyang District, Beijing, 100101, China, Institute of Geographic Sciences and Natural Resources Research, CAS.