**Data Management Plan**

**Western Pennsylvania Regional Data Center**

[**www.wprdc.org**](http://www.wprdc.org)

**Last updated 4/23/2018**

**Introduction**

The Regional Data Center is managed by the University of Pittsburgh’s Center for Social and Urban Research, and is a partnership of the University, Allegheny County and the City of Pittsburgh. The Regional Data Center provides an infrastructure for civic data sharing. This infrastructure has several dimensions, including: a technological infrastructure in the form of a community open data portal; a legal infrastructure for data exchange; a governance infrastructure for both data stored on the portal and the project itself, and an infrastructure for community engagement.

This plan describes basic data management practices and policies employed by the Regional Data Center. These practices are designed to enhance data quality, ensure proper documentation, provide data in acceptable formats, and reduce the chances that sensitive information will inadvertently be shared as open data.

Key partners in data management include Digital Scholarship Services at the University Library System and the School of Computing and Information - both at the University of Pittsburgh

The Regional Data Center’s open data portal operates on a federated model. It serves as both a data repository, and can also serve as a data catalog, providing a hyperlink to data that may live elsewhere.

**Data**

1. **Data:** The Regional Data Center hosts data from project partners and a number of additional government and nonprofit organizations in western Pennsylvania. Publishers are able to share as many datasets as they would like, provided that privacy and other quality and publishing guidelines outlined in this plan are met, and a data sharing agreement has been executed.
2. **Formats:** The Regional Data Center encourages the use of open, non-proprietary file formats by its publishers. Frequently-used formats preferred by the Regional Data Center comma-separated variable (CSV), plain text (txt), JavaScript Object Notation (JSON) and related geographic format (GeoJSON). Many of the datasets being shared by the City of Pittsburgh and Allegheny County through the WPRDC are already being made available in these formats. Publishers are also able to share data in other proprietary formats, with shapefile formats common for geographic data (bundled in a a .zip archive), pdf formats used for several documents, and Microsoft excel formats (xls) also in use. The Regional Data Center also plans to incorporate emerging file formats such as “data packages” developed through the Open Knowledge Foundation’s Frictionless Data initiative.
3. **Metadata:** The metadata standard is compliant with the DCAT and Project Open Data (POD) standards. The metadata record is tied to the dataset record. The metadata standard used by the Regional Data Center contains the following fields: Title, Description, Tags, Theme/Category, Organization, Department, Temporal Coverage, Geographic Unit, Data Notes, Related Documents, Data Dictionary, License, Format, Last Updated, Frequency Data Change, Publishing Frequency, Unique Identifier, Unique ID/URI (created automatically by CKAN for each dataset and resource), URL, Public Access Level, Public Access Level Comment, Data Steward Name, Data Steward Email, and Endpoint. To enforce the use of structured vocabularies and standardized metadata, the WPRDC open data portal was customized to enable the use of drop-down menus for metadata creation. Metadata is also harvested by the Federal government’s open data portal, at data.gov. The metadata standard was developed in collaboration with digital scholarship services at Pitt’s Hillman Library (thanks Mike Bolam) and draws heavily on the city of San Francisco’s metadata standard.
4. **Software:** Following a software selection process ([described in more detail](http://www.wprdc.org/news/inside-our-data-portal-selection-process/) in a blog post), CKAN was selected as the open data software platform for the WPRDC. CKAN is open source software initially developed by the Open Knowledge foundation. A development community has since emerged around the software, which also powers the Federal Government’s open data catalog available at [www.data.gov](http://www.data.gov). The Regional Data Center has worked with consultants at OpenGov and Ontodia to help manage and add features to the open data portal. [CKAN features](http://docs.ckan.org/en/ckan-2.2.3/features.html) include customizable metadata, an internal data store, ability to upload data manually or through automated processes, data viewers and other integrated data visualizations, integrated APIs, and download functionality.

1. **Model:** The Regional Data Center operates on a federated model, meaning that publishers have the option of adding data to the repository and/or link to data that is stored in a separate location.

1. **Security:** Publishers are asked to assume that all data shared with the Regional Data Center project should be shared publicly. All publishers are instructed to not share any private or privileged data with the Regional Data Center, and staff working under the project should not accept private or privileged datasets that cannot be published openly online. Transformations to make private data public may be supported or developed by the Regional Data Center, but should always be performed on the client-side.
2. **Privacy:** Data sharing by the City of Pittsburgh is governed by both the Pennsylvania Right to Know Law (RTKL) and City Open Data Legislation - The City is the only Regional Data Center publisher with an open data mandate in the form of legislation. Public records in Pennsylvania are subject to the presumption of openness standard in the RTKL. This standard dictates that all records are presumed to be open to the public unless: **①**Disclosure is prohibited by law, regulation, or judicial order; **②**Disclosure is prohibited by privilege, such as attorney-client or doctor-patient privileges; **③**The requested information meets one of 30 criteria outlined in Section 708 in Pennsylvania’s Right-to-Know Law; **④**Agencies deem the risks of public disclosure outweigh the benefits.

The presumption of openness standard places the burden of proof on agencies to demonstrate any exemption criteria are met. Most of the exemption criteria in Section 708 are related to the release of PII, drafts of internal documentation, or records that would pose public risks if released. It should be emphasized that public agencies can release information that meet the Section 708 exemption criteria if they deem such release in the public’s best interest.

Terms of use are provided to data users through a click-through license on the WPRDC data portal. The terms outline the general terms of use to the end user, require users to report a privacy breach, and also emphasize the importance of the data license. The WPRDC privacy policy is also included on the site, and mentions how staff may use information provided by users. Attorneys consulting on the project have questioned the legal enforceability of terms of service, and thus have recommended the use of a click-through license.

**Data Management**

1. **Repository:** The WPRDC uses the University of Pittsburgh Network Operations Center (NOC) to host the software running the open data portal and the data housed in its repository. The NOC is a state-of-the-art technical facility that houses servers and network equipment to ensure stable and reliable service for University enterprise systems. It is a centralized management center that is capable of identifying, notifying, and repairing problems when they occur and projecting when and where they might occur. The NOC is staffed 24x7 and systems are constantly monitored for problems. The Enterprise Monitoring (Netcool) and Enterprise Backup (Netbackup) services are provided via this facility, and used to provide data and system backups for the WPRDC. NOC engineers provide support for the research network and server infrastructure. The NOC provides a FISMA-compliant environment to researchers working on projects that require special handling of sensitive data. The NOC has redundant power sources along with uninterruptable power supplies and backup generators to prevent widespread system failure due to electrical outages. The functions of the Technology Help Desk are also closely integrated with the NOC to provide the best possible level of service to the University community.Data is stored at University of Pittsburgh Network Operations Center (NOC).
2. **Feedback:** Feedback from users can directly improve data quality. Users may provide feedback as a comment on the open data portal, email and phone calls, and by attending a data user group activity organized by the WPRDC. Users also have the ability to make an informal (non-RTKL) request through the open data portal. WPRDC shares this feedback with publishers as it is received.
3. **Backup:** Data is backed-up through the University NOC
4. **Breach Response Plan:** If a data breach occurs, the plan for managing it includes:
	1. Turning the dataset “private” and immediately removing all resources from the open data portal.
		1. These files may be temporarily saved locally. All copies should be deleted outright once it has been formally confirmed by the publisher that a breach has occurred.
	2. Contacting the publishing organization to inform them of an issue and develop a plan to respond. Information about what happened to cause the breach should also be collected and shared for staff discussion. Outcome of the discussion should be used to improve data management processes.
	3. The dataset record itself should be removed from the open data portal if there are no data resources remaining in the dataset record. If a new resource can replace the old data (e.g suppressed or summary data), the dataset may remain on the open data portal. If not, the dataset itself should be removed from the open data portal.
	4. Public notifications of a breach identifying a particular dataset should not be made. We do not want to call attention to any sensitive data that may be archived and still available elsewhere. We should capture a record of the breach and basic details about it in the Data Center performance statistics spreadsheet.
5. **Archiving:** The [Future-Proofing Civic Data report](https://drive.google.com/file/d/0B3MMB2pFQdI4MXlkTUhLOGNvdXlIeGViQURmZGQ4S2gxY1lN/view) from Temple University has been a informative and helpful resource for us. If data is published, but a new or improved dataset is created to take the place of an older dataset, the “decomissioned” data should clearly be marked in the title or description of the dataset or resource as an older version, and users should be directed to use the newer data. Alternatively, publishers may choose to remove the older resource or dataset from the open data portal itself if its presence will serve to confuse data users, or provide poor-quality data. The Regional Data Center should save a copy of the decommissioned data locally in a “dark archive,” or local copy. Older data may also be removed from the open data portal (stored locally) or archived into a zip file if storage becomes an issue. Metadata records should include notes on whether a particular dataset has been archived, and how to access archival data. It is important the older data be retained in the event that a researcher or someone else has made a reference to the decommissioned data.

**Data Sharing**

1. **Users:** Potential users of data include and are not limited to public, private, and nonprofit employees, journalists, librarians, students, researchers, data intermediaries, residents, entrepreneurs, and civic data enthusiasts. Anyone can use data from the open data portal provided they agree to the terms of use and adhere to the dataset licensing terms.
2. **Legal Infrastructure:** The legal infrastructure of the Data Center provides a framework enabling the University of Pittsburgh to operate a shared open data repository in collaboration with Allegheny County and the City of Pittsburgh. Additional data users in Western Pennsylvania are also able to use the Data Center’s infrastructure to publish open data. The three components of the Regional Data Center’s legal infrastructure include:
	1. **Data Deposit Agreement:** The first part of this legal infrastructure consists of a [data deposit agreement](http://www.wprdc.org/wp-content/uploads/2015/08/WPRDC-Project-Agreement_Exh-F_RDC-Data-Deposit-Agreement.Final_.07.16.15-WEBSITE-COPY1.pdf) between publishers and the University of Pittsburgh. This agreement enables these organizations to share data on the Data Center’s open data repository. By completing a two-page data deposit agreement and a publisher training session, any nonprofit, governmental, and academic organization in Western Pennsylvania  can publish open data through the WPRDC at no cost.
	2. **Terms of Use:** The second piece of the infrastructure governs the [terms of use](http://www.wprdc.org/terms-of-use/) for the Data Center’s open data repository. Data users are prompted to agree to these terms (via a click-through license) prior to using the open data repository.
	3. **Data License:** The third piece of the infrastructure involves the data license. Data licenses are important in that they allow data users to clearly understand conditions under which data can be re-used. The Data Center asks each data provider to either assign a license to each database they share, or explicitly assign the database to the public domain. In the Data Deposit Agreement, data providers are also asked to provide a default license to be assigned by the Data Center if no license is specified at the time of publication.
3. **Publishing Responsibility:** Publishers are responsible for publishing their own data unless the WPRDC is involved directly in the publishing process. A direct role is common for automated or harvested datasets. This responsibility includes determining what data to share, including any relevant privacy protections. Data providers are largely responsible for ensuring data quality standards are met, creating metadata records, and providing data dictionaries which define each data field.

The Regional Data Center staff can work to improve data quality through automated and non-automated processes. Where possible, the WPRDC works with publishers to automate publishing of critical datasets. These processes are commonly known as an “ETL” process - Extract-Transform-Load. Processes with direct involvement by the Regional Data Center are given an “ETL” tag in a dataset’s metadata record, and the number of automated datasets published is a key performance measure for the Regional Data Center. Additional GIS datasets are harvested from City of Pittsburgh and Allegheny County GIS (ESRI) open data portals on a weekly basis. The WPRDC developed a toolkit to extract data from other systems, apply necessary transformations, and load it to the open data portal. This process also validates the structure of datasets and send validate and send automated error notification to data publishers.

1. **Privacy Check:** The WPRDC provides a second layer of privacy assessment on all datasets shared on the open data portal for the first time. One of the WPRDC project team members reviews each dataset intended for publication before it becomes publicly accessible on the open data portal. The WPRDC can refuse to publish any dataset that staff believe would create undue harm to an individual if shared as open data. The WPRDC has also organized informal privacy roundtables to talk through the privacy implications of sharing sensitive data, and develop a involving staff from the Digital Scholarship Services team at the University of Pittsburgh’s Hillman Library, along with staff and students at the School of Information Sciences. Data will be released for access as soon as the final review of quality and privacy is assured. The Regional Data Center makes every attempt to conduct all reviews within 24 hours of each request (excluding weekends).
2. **Data Updates:** When updates are behind schedule, or automatic processes are not working properly, the Regional Data Center staff will notify data stewards when it detects an issue..

**Other Plans**

**Data user guides:** The WPRDC also leads the development of data user guides. User guides are designed to be a companion to metadata.  The Regional Data Center is now producing documentation for some of the most-used and also the most-complex datasets on the data portal. These guides are designed to give data users a sense of the purpose for which the data was collected, the business processes, software, and applicable standards involved in creating data,, suggested applications for the data (and how to use it), and other details that will help others make responsible use of the data.

**Training and tutorials:** Training classes and training materials are offered to staff at publishing organizations looking to load data, create metadata, and add a data dictionary on the open data portal. Training also encompasses information about how to prioritize datasets for release as open data, information related to selecting a license (public domain licenses are encouraged), and techniques and strategies for protecting personally-identifiable information.