LESSONS ON DATA MANAGEMENT PRACTICES FOR LOCAL DATA INTERMEDIARIES

AUGUST 2017

Rob Pitingolo

NNIP
NATIONAL
NEIGHBORHOOD
INDICATORS
PARTNERSHIP

ACKNOWLEDGMENTS

I appreciate the thoughtful suggestions on the draft guide from Minh Wendt and her colleagues in the Office of Minority Health at the US Department of Health and Human Services, Bob Gradeck of the University of Pittsburgh Center for Social and Urban Research, Noah Urban and Stephanie Quesnelle of Data Driven Detroit, and April Urban of Case Western Reserve University. Leah Hendey and Kathryn Pettit (technical reviewer) at the Urban Institute also provided valuable comments. We are grateful to the staff at Data Driven Detroit, NeighborhoodInfo DC, Metropolitan Area Planning Council, and the Community Research Institute at Grand Valley State University for generously sharing information for the case studies.

This guide was supported by the Office of Minority Health at the US Department of Health and Human Services. The views expressed are those of the author and do not necessarily represent those of the Office of Minority Health, Data Driven Detroit, the Urban Institute, its trustees, or its funders.

Table of Contents

INTRODUCTION

Have you ever had to deconstruct a data set or analysis a former staff member created? Have you thought to yourself there's got to be a better, more efficient way to get the data I need? Does the time you spend with data processing cut into the time you could be spending on analysis and helping local stakeholders use the results? This brief is designed for data intermediary organizations to help you develop and improve an important part of data management known as the ***ETL (Extract, Transform and Load)*** process.

The National Neighborhood Indicators Partnership (NNIP) believes that data intermediaries should work to improve the efficiency of their data management to contribute to organizational sustainability, spend more time on understanding their communities, and helping people in the community use the data. Coordinated by the Urban Institute, NNIP consists of organizations in more than 30 cities that have a shared mission to help community stakeholders use neighborhood data for better decision making, with a focus on assisting organizations and residents in low-income communities. Local data intermediaries, assemble, transform, and maintain data in order to fulfill this mission. Good data management practices also contribute to the credibility of Partners by helping them ensure consistency and quality in data processing.

This brief is not intended to be a step-by-step guide to developing an ETL process. Rather, it presents a collection of advice and lessons learned from data intermediaries in NNIP who developed successful ETL processes. Every organization has different staff skills and technology environments, so there is not a single correct ETL process. Many existing ETL processes reflect the reality of legacy systems and the accumulation of many decisions made over time. Partners' experiences demonstrate the value of periodically assessing the ETL process to determine if current practices are meeting community needs or could be improved. Building or changing an ETL process takes time and it is not unusual for data intermediaries to start small. If tackling everything at once feels overwhelming, consider developing an ETL strategy that begins with a few datasets and can be expanded over time.

Section 1 of this guide provides advice for getting organized and discusses each stage of the extract, transform and load process. Section 2 provides case studies from five NNIP Partners to illustrate how these steps are put into practice.

SECTION 1

THE ETL PROCESS

SECTION 1: THE ETL PROCESS

GETTING ORGANIZED

Organization is critical to the success of an ETL process and provides the backbone for the ETL stages. Getting organized includes determining the hardware or service to store the data and related documents, creating a file directory structure, and establishing a system for backing up your files.[1]

First, you will need to designate a central server where you can store your data as well as documentation related to the data such as licenses, memoranda of understanding, program instructions, etc. An additional benefit to using a central server is that you can assign permission to multiple staff members who may need access to the data. Avoid storing data on a single computer or an external hard drive that can be damaged or lost. The central server can be hosted inside your organization, or in a cloud service like Google Drive or Amazon Web Services. To decide what solution is appropriate, you should carefully evaluate factors like whether you will be handling any confidential or sensitive data, what data security capabilities and process you already have in place, what you need to develop, how the system performance and monitoring will work, and if there are any privacy, legal, or compliance concerns.[2]

Second, design a logical directory structure to organize your source data received from a data provider, processing code, and processed data. You can organize by topic (e.g. crime, property, health) or by source (e.g. police department, county assessor, health department). Mixing and matching topic and source at the same directory level may make it more difficult to for anyone looking for the data to find what they need, and cause problems when updating the data.

Lastly, think about a process for backing up both your source and processed data. At a minimum, it can be as simple as saving data to an external hard-drive on a regular basis, but it is recommended to use a script or software program to automatically backup data to a server or

[1] *As one resource, the Digital Curation Centre offers a Checklist for a Data Management Plan that includes prompts about data collection, documentation, ethics, storage, and sharing, as well as other general resources.*

[2] *Though about education data specifically, the Privacy Technical Assistance Center's Frequently Asked Questions on Cloud Computing provides good resources and questions to consider on establishing a secure cloud environment for sensitive data.*

cloud storage service. Off-site backups are safest since the data are physically protected in case something happens to the main server (fire or water damage, for example). Choose a regular backup interval that makes the most sense for your organization. If you have data coming in at high frequencies, a weekly backup might work best, otherwise a monthly or quarterly backup could be appropriate.

READING IN DATA (EXTRACT)

The first step in any ETL process is bringing data that you receive from the source organization into your system. Source data refers to the data received from a data provider. This term is used because the data are typically direct extracts from the provider's databases and have not been processed or manipulated. In this brief, we refer to publicly available administrative data from local governments, but in other contexts it could mean private or confidential data as well.[3]

Obtain Source Data

Data can come from data providers in different ways. Some data are freely available to the public on open data portals, other data are public but must be "scraped" from websites, and some must be obtained through negotiations with the provider. In all cases, developing strong relationships with data providers will help you to understand the content and limitations of the data and to ensure you can receive data updates in the future.

For data that are received through negotiations, the staff member at the data provider handling the paperwork may not be a technical expert in the data. In this case, try to identify a point of contact who is familiar with the data and can clearly define the elements within the data set. This will give you someone to ask if you have questions or run into problems. Even for data that are publicly available, a point of contact can assist when the data have insufficient documentation or other questions arise during data processing.

If possible, request data in plain text (fixed width) or CSV (comma delimited) format. This will ensure integrity of the data. Avoid requesting data in Excel or other proprietary formats, which may disrupt the original format of the data. For example, in Excel, a string variable with numbers is often automatically converted to a numeric variable and leading zeros are dropped. Proprietary formats sometimes have limits to how many rows and columns are supported, which

[3] *Additional information about protecting confidential information is on page 50 of NNIP's Guide to Starting a Local Data Intermediary. A forthcoming brief on this topic will explore confidential data in greater depth.*

can be problematic for large datasets. Once the data are received and saved to your system, consider saving the file as "read only" so that someone cannot accidentally overwrite the file.

Ask your data provider to include a metadata record or data dictionary for the source data file. This should define, in detail, all rows and columns in the dataset. For example, in a real property dataset, a column might be named "sales price" but the metadata record would include information like whether that includes closing costs or taxes. The metadata record should also define any values for categorical fields. For example, in the same real property example, for a variable called "property type," a code of 1 might indicate that the property is a single-family home, and a 2 might indicate that it is a condominium unit. With administrative data, the reality is that documentation is often incomplete. In this case, one strategy is to create a blank metadata record that you can use to work with the data provider to complete.

Decide on Software

Once you receive and save source data to your central server, you need to decide what software to use to clean, process, analyze, and/or visualize the data. NNIP Partners all pledge to regularly update data over time, doing so efficiently requires strategically selecting software that can facilitate this. A software package that will help you accomplish this goal will allow you to program or script your manipulation (cleaning, processing, analysis, or visualization) of the data set.

You should establish a standard process and criteria to evaluate any software under consideration.[4] This will help staff compare different options systematically and base the decision on organizational interests instead of personal preferences. For example, what are the upfront versus ongoing costs? Does the software have the needed features? Is it reliable and regularly updated? Is it easy for other staff members to use and for knowledge to be transferred between staff members? An ETL process that is reliant on the knowledge of a single person or a small group of people may have limited value in the long term.

Avoid using spreadsheet programs like Microsoft Excel or Google Sheets to manipulate data. These are widely-used tools with many useful features; however, spreadsheets are not ideal for ETL because they do not produce a log tracking the changes made to data. This means the data processing cannot easily be reviewed or replicated in the future.

[4] *The University of Pittsburgh Center for Social and Urban Research has shared their general steps for deciding on software and the criteria they use on the NNIP website.*

Other factors you should consider when deciding on software include how well the software meets your organizational needs, the cost, learning curve, and tradeoffs among these factors. For example, you may have to choose between expensive proprietary statistical software and free open-source software. Or you may have to choose between a programming language that can perform powerful data manipulations but is time and resource intensive to learn and a database management system that is easy to learn but lacks some of the more advanced features.

The best software option for your organization is the one that you can afford, can use effectively, and eases knowledge transfer between staff members in the event of turnover on the project team or the arrival of new staff in the organization. As a network policy, NNIP does not endorse any products or services from for-profit or non-profit firms.

We classified potential ETL options for managing data into five main categories, which are not mutually exclusive. For example, programming languages can sometimes be used inside of other programs or they can be used on their own.

- **Statistical software** like R, SAS and Stata are code-based software designed specifically for data management and analysis. Costs vary from free (R) to several thousand dollars for an entry-level license (SAS). The learning curves for these tools can be high, but some of them are taught in college courses on business, economics, and other fields that incorporate the study of statistics in their programs, so data intermediaries often do not have to invest in additional training. Intermediaries at universities may also have access to enterprise-licenses for statistical software through their institution.

- **Database management systems (DBMS)** like MySQL, PostgreSQL and Oracle allow for the storage of source data, and the transformation and management of processed data. Like statistical software, DBMS are code-based and scripts can be written to perform actions on the data stored within the systems. Costs also vary, from free and open-source like PostgreSQL to enterprise-level pricing for Oracle. Learning curves vary depending on the specific system, but most use the SQL language that staff may already be trained on from college courses or previous database experience.

- **ETL tools** like Talend and Pentaho were developed for corporate applications but can be applied to the data intermediary context as well. Businesses use these tools to manage and organize their data flowing from a variety of internal sources. Data intermediaries can use them for managing multiple data streams from their providers. These tools have an easier learning curve because they use a workflow-based graphical user interface, but still produce code behind the scenes that can be saved, replicated, or re-run over time.

- **Programming languages** like Python and Ruby are general purpose computer languages that can be used for many different applications, including data management and analysis. These are free or low-cost tools but have the steepest learning curve, which means data intermediaries will have to hire programmers proficient in the language they choose, or invest substantial resources to train their existing staff. However, they are powerful and flexible in that they can perform more than multiple types of functions. For example, a python script could be used to connect to an API, process data received from it, and upload that processed data to a FTP server.

- **Geospatial software** like Esri's ArcGIS, QGIS and Carto can be used to manage geographic information in a database. These products also range in cost from free (QGIS) to enterprise-level pricing (Esri). Data intermediaries can use these tools for data that have geographic references, such as the location of crime incidents in a neighborhood. The learning curve for ArcGIS and QGIS is high, but newer tools like Carto are working to make geospatial software that is easier to learn and use.

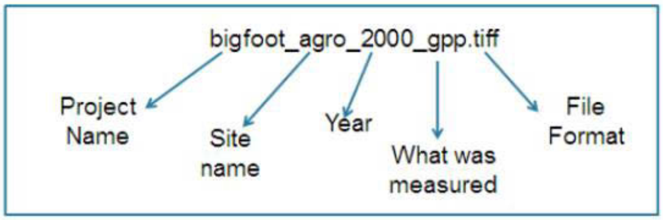
Across the NNIP network, data intermediaries successfully use tools in all five of the categories. Some data intermediaries use multiple tools from different categories to accomplish different tasks. It is not required to choose just one. The case studies included in section 2 go into more detail about how data intermediaries use these tools in their own ETL processes.

CLEANING AND PROCESSING DATA (TRANSFORM)

Transforming data refers to applying a set of rules or functions to the source data file to prepare it for its final application. This includes cleaning of the data for greater consistency and quality, transforming variables to create indicators, and manipulating the data into the format that their future use requires, whether that is summarizing for a neighborhood profile, analyzing for a report, or releasing to the public through an open data portal.

Develop Data Standards and Procedures

In addition to a consistent process, you should develop a common set of standards for your data. This includes developing guidelines for assigning descriptive names for data files. The names should be unique and avoid spaces. (See Cook 2012). The components of the file name should reflect the contents of the file, such as the source agency, the project, the geographic level or the year of the data.



Source: Cook, Robert. 2012. Managing Your Data: Assign Descriptive File Names. *http://commons.esipfed.org/node/702*

You will want to consider standardizing variable names, such as renaming fields in time series data (for example var_2015, var_2016) or a common field name for "census tract" across data files and sources. This will make eventual analysis and publishing much easier for staff members.

Finally, you should determine standard entries for values within the variables. For example, different data providers may format dates in many ways, but your cleaned data would always have those dates re-formatted to "mm-dd-yyyy". Or you might want to reformat census tracts to an 11-character variable based on combining state and county FIPS codes and the 6-character tract id, even if the data provider omitted part of that in the source data.

You should develop a consistent process to transform the data that you follow for every new dataset. Below, are common transformation steps and tasks. You may have additional steps based on your organization's needs.

- Evaluate and decide on adjustments for data quality[5]
    - Evaluate for missing data and decide how to handle missing values or observations (i.e. ignore them, impute them, assume missing data are zero, etc.)
    - Evaluate and decide how to handle outliers (extreme high or low values)
    - Standardize variable names and values based on organization standards
- Create or drop fields as needed for the final uses

[5] *For a primer on data quality issues, see the June 2017 webinar on Improving Administrative Data Quality for Research and Analysis by China Layne.*

    - Subset the data to keep only fields of interest
    - Recode values, such as collapsing categorical data into fewer categories
    - Construct derivative variables, such as crimes committed with a certain type of weapon
    - Geocode addresses and add coordinates for latitude and longitude[6]
    - Add relevant geographic fields (such as census tracts or neighborhoods) through geocoding process or geographic crosswalks

- Create derivative sets

    - Summarize data to required geographic levels for ease in fulfilling future requests

- Document the file and process.

    - Label all variables
    - Save your final data file(s) following organization standards in the appropriate directory
    - Save the final code or workflow for organization wide use
    - Develop metadata and documentation

Save Code or Workflows

Create code, script, programming, or workflows in the software that you use for data transformation. This code can be saved and reused in the future by other staff in the organization to repeat the manipulation performed on the data set. Saving your code or workflows has several benefits, including the ability to easily edit the code if a small change is needed without having to re-do the cleaning process from scratch. You can clean future datasets updates with minimal effort, assuming the underlying format of the source data has not changed.

Sometimes the source data does change and you need to update your code or workflows. Practicing version control will help with troubleshooting in case anything goes wrong and you need to roll back to a previous version of the code or workflow. Tools like GitHub work particularly well for code-based software and programming languages like R and Python. Some ETL tools have version control built-in, but for those that do not have this feature, other processes can manually accomplish the same result. For example, staff can save programs (with

[6] *NNIP's Tip Sheet: Geocoding Software and Services, provides an overview of the commonly used systems to geocode data with general information about benefits and drawbacks of each.*

standardized file names, including dates) as a new version of the code or workflow every time you update it.

Develop Metadata and Documentation

Documenting the data is a critically important, but easily overlooked, step in the ETL process. As discussed in the Extract section, you should always request metadata for source data, but you may want to enhance the original metadata with information that you ascertained through conversations with the data owner or during processing.

You should also produce new metadata to match each derivative data file, as new fields are created and data are otherwise manipulated during the Transform process. A good place to start in developing a standard for your organization is with some of the already established metadata standards, such as the Dublin Core Metadata Initiative, the Federal Geographic Data Committee, or Project Open Data Metadata. Most importantly though, use a consistent internal standard (even if it is not an official one).

Creating a process that makes it easy for staff to enter metadata is critical to having good metadata– both in getting staff to enter metadata and to ensure quality control. For example, rather than creating metadata from scratch each time, a standard form, with drop down menus can be a big help in keeping everything consistent. Some tools also incorporate the metadata creation steps directly into your ETL process. For example, if working in a statistical package like R or a programming language like Python, consider adding extra code that also produces a summary data file that includes all fields, labels, allowable values (for categorical fields) or descriptive statistics like mean, min, and max for continuous numeric fields.

PUBLISHING DATA (LOAD)

Data intermediaries use their data for different types of projects. Some produce neighborhood or community profiles with indicators summarized at different geographic levels (census tract, neighborhood, etc.). Others release cleaned data to the public through an open data portal. Data intermediaries also use the data from their repository to produce analytic reports or conduct research projects. Organizations should think about what the primary uses for their data will be and design the ETL process around it, but try to keep the process flexible enough to allow for additional uses of the data.

A direct link between the central database or data repository and web-based tools is an efficient way to publish data. For example, community profiles that pull summary data directly from an application programming interface (API) or SQL database do not require any staff time to refresh them whenever new source data are obtained and cleaned. Updates to the data can also be published more quickly with a direct link.

CONCLUSION

Taking the time to establish good data management practices that are suitable for your organization will pay off in data quality, efficiency, and sustainability of your systems. Your ETL processes may be imperfect, but can be improved over time as staff members gain skills and learn from experience. Periodic reviews of your procedures and technology solutions will also ensure that your practices stay up-to-date as technology and your organization evolves. With a foundation of high quality data in place, local data intermediary staff will be better positioned to offer insights from the data on important issues and assist local stakeholders in using the data to improve their community.

SECTION 2

CASE STUDIES

SECTION 2: CASE STUDIES

The case studies in this guide present a range of processes from which you can draw insights, but each approach has pros and cons that you should evaluate before choosing your solution. NNIP member organizations are diverse in terms of their institutional type and staff size. Some are nonprofit organizations with small staffs, while others are part of larger university centers or government agencies with staff and resources provided by the broader organization. We have noted some key information about each organization to help put their ETL process into context.

DATA DRIVEN DETROIT

Data Driven Detroit (D3) in Detroit, Michigan recently hired an IT director who brought skills and techniques from his previous career in the corporate world. The tool that D3 uses to manage its ETL process is Pentaho, a software program developed primarily for business intelligence.

DATA DRIVEN DETROIT

Institutional type: Low-profit limited liability company (L3C)
Number of staff: 10
Data management software: Pentaho, Python

The advantage to using a software tool with a visual user interface is that users do not need to know programming languages or statistical packages to clean and transform data. It also allows users to save their workflows so that data that are regularly updated can be quickly and efficiently processed.

D3 receives data from local agencies and stores them on a central server. Source data and any documentation that comes with them are organized by project. Within each project directory, data are further organized into folders for 1) source data, 2) processed data and 3) metadata created during the ETL process.

D3 staff use Pentaho to check the quality of the source data. For example, there is a process for standardizing geographic ID fields, such as parcel IDs, which are not always in a consistent format in the source data that D3 receives. At the end of the process, Pentaho exports cleaned data to a SQL database. Depending on project needs, a spreadsheet or similar flat data file can also be produced. D3 runs an open data portal based on the Esri Open Data Portal technology, which is linked to this SQL database. Data that include addresses are geocoded when loaded to the Esri portal automatically. Users can view that data on online in a map, or download a spreadsheet, KML file, or shapefile from the portal.

D3 is shifting its strategy from using programming languages like Python to software like Pentaho. Before hiring their current IT director, a significant amount of data work at D3 was performed using Python. However, given the specialized skill set needed to program in this language, D3 was concerned about the sustainability of their ETL processes. Pentaho is now the primary ETL tool, but Python is still used at D3 for specialized tasks, such as web scraping and PDF scraping, tasks that are simply not possible in an ETL tool. D3's IT director is using Pentaho to demonstrate the long-term value that tools like it can add to the organization in the long term.

COMMUNITY RESEARCH INSTITUTE

Community Research Institute (CRI) at Grand Valley State University in Grand Rapids, Michigan developed an ETL process by starting with the final use of the data in mind and then working backward to get there. The backbone of CRI's system is an Oracle database where all their data are stored and processed. CRI organizes this database by individual project.

COMMUNITY RESEARCH INSTITUTE

Institutional type: University center

Number of staff: 28

Data management software: Oracle, PostgreSQL, Python

Source data are obtained either as files provided directly from data partners or indirectly using scripts that engage in web scraping. For very frequent updates, CRI uses a script to automatically download CSV files from the data providers' Secure File Transfer Protocol (SFTP) and pull them directly into the database. For less frequent data, files are downloaded from a SFTP server by CRI staff and then uploaded manually into the database. In some cases, web scraping scripts go out and pull data from the web directly into the database.

CRI prioritizes the creation of metadata and data dictionaries. They start thinking about metadata as soon as a new relationship is established with a data provider. The staff creates specifications for the data file before a request for data is made, and the data provider is asked to return data and metadata that meet these specifications. This process makes it more likely that the data and format are exactly what they need. This also reduces the number of interactions with the data provider during the cleaning process since everything is clearly defined up front.

A standard set of steps is run on every dataset that is loaded into the database, including basic quality checks, such as making sure the fields and number of observations matches the numbers expected. After that, the process ensures that string variables are all converted to uppercase and dates are formatted as dates rather than strings. The exact cleaning depends on the project need. For record level data, data are checked to make sure every record has a unique ID. If the data are longitudinal, timestamps are checked and new fields are created if necessary to mark records during certain time periods.

Version control is built into the database so that any SQL scripts developed are versioned. Since source data are saved in the database and scripts are all versioned, CRI staff can easily roll back to any previous version of the data if necessary. The database itself is also regularly backed up. A transactional backup occurs nightly, which saves log files that act as restore points. Full backups are performed on the weekends and backed up to servers in a separate physical location.

CRI is fortunate to have a very powerful database as well as full-time programmers and database managers to help run it. Even though they use a propriety Oracle database, they are considering switching to an open source PostGRES database because it handles geospatial data better. CRI staff emphasize that the specific technology used is less important than the processes they have developed.

UNIVERSITY OF PITTSBURGH CENTER FOR URBAN AND SOCIAL RESEARCH

The University of Pittsburgh Center for Urban and Social Research (UCSUR) in Pittsburgh, Pennsylvania, along with partners at the City of Pittsburgh and Allegheny County, manages and operates the open data portal for the Western Pennsylvania Regional Data Center (WPRDC), an unusual role for NNIP Partners. UCSUR developed their ETL process to take source data and transform them into the best format for the WPRDC open data portal users. UCSUR also handles confidential data but those data are handled in a separate ETL process.

At the beginning of their partnership for the WPRDC, UCSUR hired a consultant to develop an ETL framework. They chose CKAN as their open data technology and carefully documented their selection process to help others. CKAN is a system like Esri Open Data or Socrata, but it is open source. UCSUR's entire ETL process is built around taking source data and standardizing it for CKAN.

UNIVERSITY OF PITTSBURGH CENTER FOR URBAN AND SOCIAL RESEARCH

Institutional type: University center
Number of staff: 40
Data management software: Python, CKAN

UCSUR receives source data primarily from their city and county partners via asecure FTP server. They ask for data in CSV or text format to avoid any encoding problems. In cases where data are received on a frequent basis, an API or data feed might be established to automate the process.

WPRDC has strict rules about confidential data so UCSUR asks that agencies never transfer confidential data to them, unless it's for a specific project that won't send data to the portal. In cases where confidentiality is unclear, they work directly with the data providers to help them remove any sensitive information before transferring the data.

A metadata record is provided to the data partner to complete before any data are transferred. The data provider gives information about field names, descriptions, and allowable values of the data. UCSUR will not proceed further with data cleaning and processing until this metadata record is complete. WPRDC uses the Data Catalog Vocabulary (DCAT) standard and WPRDC's metadata system is well-documented for data providers. They also make sure that the metadata record is compliant with the Federal Government's Project Open Data. This allows data to be harvested directly from the WPRDC into Data.gov.

Source data are processed using Python scripts and cleaned to make it consistent with other data in the WPRDC portal. For example, dates stored as strings are reformatted; yes/no fields are converted to 1/0; and IDs like parcel numbers are checked to make sure they can be linked to other parcel files. Data that includes addresses are geocoded using a master address file published by Allegheny County. Right now, a staff member manually loads data into the geocoder, but UCSUR intends to automate geocoding in the future so that any data that have columns that look like addresses are geocoded during the cleaning process.

UCSUR creates summary data for the select data that are used in their community profiles. In these cases, summary indicators are produced for geographies including census tract, neighborhood, and school district.

Open source tools like CKAN offer some advantages in terms of cost and flexibility. However, UCSUR staff and their consultant did have to spend time developing their ETL process to revolve around that particular tool. At the same time, UCSUR staff acknowledges that it might not be the best solution for everyone because it is specific to uploading data to CKAN as the end goal.

METROPOLITAN AREA PLANNING COUNCIL

The Metropolitan Area Planning Council (MAPC) in Boston, Massachusetts is currently undergoing a major change in how they handle their ETL process. Currently they have two departments, Data Services and Digital Services. The Data Services department stores source data in folders on an internal network, and processed data in an internal PostgreSQL database. The Digital Services department stores source data on a cloud-based database hosted by Amazon Web Services and processed data in a cloud-based SQL database. They are working on having the Data Services staff shift to the Digital Services' approach.

METROPOLITAN AREA PLANNING COUNCIL

Institutional type: Regional planning agency

Number of staff: 80 total; 11 in the data services department

Data management software: PostgreSQL, Python, R, ArcGIS, Mapzen, Carto

MAPC can take advantage of many different tools and technology in their ETL process because they have programmers on their staff. Whenever possible, MAPC tries to use APIs or direct feeds to obtain source data. For example, Census and some state agency data are automatically pulled from an API directly into MAPC's cloud-based server. For data without an API, source data files are sent over SFTP and then saved on MAPC's side.

MAPC programmers use many different technology tools to process and clean data. Their programmers have personal preferences for tools, but in other cases the technology is selected because it is best suited for the task at hand. For example, both Python and R are used for data cleaning, and the code is saved and re-used whenever an update is needed for a given dataset. Version control is handled by committing code on to GitHub. For geocoding, MAPC uses ArcGIS and Mapzen, a tool with state-wide address data for Massachusetts.

For publishing data, MAPC creates APIs, typically using a Ruby on Rails web application, but they also use the mapping tool Carto's API for spatial data. APIs power MAPC's web applications. For example, their community profiles are populated by pulling data directly from an API. Metadata are part of the ETL process and are updated every time the data are updated. Metadata are typically stored as a table in a database or as a spreadsheet and created by the staff member doing the data update.

NEIGHBORHOODINFO DC

NeighborhoodInfo DC is the data intermediary for the Washington, DC region based at the Urban Institute. Their ETL process is primarily based in SAS, a proprietary statistical analysis software. They take advantage of institution-wide software licenses, like SAS, which might be cost-prohibitive for small nonprofits. They have also started exploring the possibility of using R, a free, open-source statistical analysis software.

NEIGHBORHOODINFO DC

Institutional type: Non-profit research organization

Number of staff: 4

Data management software: SAS

NeighborhoodInfo DC receives source data mostly from the District of Columbia government. Many datasets are now available through Open Data DC, such as real property sales and reported crimes. Other data, such as records on Temporary Assistance for Needy Families and Supplemental Nutrition Assistance Program recipients, are provided to NeighborhoodInfo DC through an agreement with the different city agencies.

Source and processed data are all stored on a central server at the Urban Institute in directories organized by data source. A series of SAS programs convert source data into SAS datasets and perform several cleaning operations before saving the processed data and geographic summary files back onto the central server. Summary files refer to those where data are aggregated to custom Washington, DC geographies such as Ward and Neighborhood Clusters.

As an example of the type of data cleaning performed, NeighborhoodInfo DC cleans crime data by reformatting date fields to a consistent format and transposing the data from long to a wide format so that it can be analyzed longitudinally. They also use fields on the type of crime and weapon to re-code the incident as Part I and Part II offenses. For data with street addresses, such as real property sales, NeighborhoodInfo DC uses the Proc Geocode procedure in SAS, in conjunction with an address file provided by the DC government, to geocode addresses as part of in the main ETL process.

The ETL process includes labeling every variable in a dataset and creating SAS formats, which identify the field values. Using those labels and formats, NeighborhoodInfo DC produces metadata using SAS and exports the results as HTML pages that can be viewed in a web browser. The HTML metadata pages list every variable in a dataset, its label, formats (for categorical variable) and statistics such as minimum, maximum and mean (for continuous variables).

Github is used for version control of the code. It is also used as a task manager, to keep track of issues or data tasks to be performed, as well as to assign those tasks to staff members. Backups are done monthly, to an external hard drive. NeighborhoodInfo DC is in the process of evaluating a strategy for more frequent backups to a secure location.

Select indicators from many of the cleaned data sources are uploaded to the Neighborhood Profiles section of their website. Data are also used by other researchers at the Urban Institute for specific research studies.

RESOURCE LIST

Resources referenced in sections one and two are listed here.

Checklist for a Data Management Plan

http://www.dcc.ac.uk/resources/data-management-plans/checklist

Data Management Plans General Resources

http://www.dcc.ac.uk/resources/data-management-plans

Frequently Asked Questions—Cloud Computing

http://ptac.ed.gov/sites/default/files/cloud-computing.pdf

NNIP's Guide to Starting a Local Data Intermediary

http://www.urban.org/sites/default/files/publication/80901/2000798-NNIP%27s-Guide-to-Starting-a-Local-Data-Intermediary.pdf

University of Pittsburgh Software Selection Process and Criteria

http://www.neighborhoodindicators.org/node/5320

Software and Technology

- SAS - https://www.sas.com
- Stata - http://www.stata.com/
- R - https://www.r-project.org/
- MySQL - https://www.mysql.com/
- PostgreSQL - https://www.postgresql.org/
- Oracle - https://www.oracle.com
- Talend - https://www.talend.com/
- Pentaho - http://www.pentaho.com/
- Python - https://www.python.org/
- Ruby - https://www.ruby-lang.org/en/
- Esri - http://www.esri.com/
- QGIS - http://www.qgis.org/
- Carto - https://carto.com/

Managing Your Data: Assign Descriptive File Names

http://commons.esipfed.org/node/702

Avoiding Garbage In – Garbage Out: Improving Administrative Data Quality for Research
https://www.linkedin.com/pulse/avoiding-garbage-out-improving-administrative-data-quality-layne

Github
https://github.com/

Tip Sheet: Geocoding Software and Services
https://www.neighborhoodindicators.org/node/5434

Metadata

- Dublin Core Metadata Initiative -
- Federal Geographic Data Committee - https://www.fgdc.gov/
- Project Open Data - https://project-open-data.cio.gov/v1.1/schema/

Western Pennsylvania Regional Data Center
http://www.wprdc.org/

WPRDC: Inside Our Data Portal Selection Process
https://www.wprdc.org/news/inside-our-data-portal-selection-process/

Data Catalog Vocabulary (DCAT)
https://www.w3.org/TR/vocab-dcat/

Open Data DC
http://opendata.dc.gov/

National Neighborhood Indicators Partnership organizations mentioned:

- Data Driven Detroit - https://datadrivendetroit.org/
- Community Research Institute - http://gis.cridata.org/
- University of Pittsburgh Center for Urban and Social Research - https://www.ucsur.org/
- Metropolitan Area Planning Council - http://www.mapc.org/
- NeighborhoodInfo DC - http://neighborhoodinfodc.org/

NNIP is a collaboration between the Urban Institute and partner organizations in more than 30 American cities. NNIP partners democratize data: they make it accessible and easy to understand and then help local stakeholders apply it to solve problems in their communities.

NNIP NATIONAL NEIGHBORHOOD INDICATORS PARTNERSHIP

For more information about NNIP, go to
www.neighborhoodindicators.org or email nnip@urban.org.