Return to EnviroSensing Cluster main page

Contents

Overview

Archiving data snapshots and using appropriate metadata and packaging standards can increase the longevity and discovery of data immensely. However, these local curation techniques are still susceptible to threats to the projects or institutions that maintain the local archive. People in critical technology positions that maintain archives may change careers or retire, projects can lose funding, and institutions that seem solid can dissolve due to changes in political climate. For these reasons, partnering across institutions to provide archival services of data can greatly increase the probability that data will remain accessible for decades or into the next century.

In this chapter, we discuss techniques and issues involved with archiving data on a multi-decadal scale. For sensor data, we promote the use of periodic data snapshots, persistent identifiers, versioning of data and metadata, and data storage formats and strategies that can increase the likelihood that data will not only be accessible into the future, but will also be understandable to future researchers.

Introduction

A data archive is a location that has a reasonable assurance that data and the contextual information needed to interpret the data can be recovered and accessed after significant events, and ultimately after decades. Data archives should be maintained through backup strategies such as redundancy and offsite backup, in multiple locations and through institutional partnerships. Archiving activities should have institutional commitment, and ideally cross-institutional commitment. Archives may be locally maintained, may be part of a national or network-wide archive initiative, or both. For raw data, an archive can be a local or regional facility, whereas quality controlled, ‘published’ data should be archived in a community-supported network archive and available online.

Environmental research scientists are in need of accessing streaming data from sensor networks both provisionally in near real-time, after QA/QC processing, and in final form for long-term studies. Without appropriate archiving strategies, data are at great risk of total loss over time due to institutional memory loss, institutional funding loss, natural disasters, and other accidents. These typically include near-term accidents and long-term data entropy due to career and life changes for the original investigator(s) [Michener 1997]. Data, and the methods used to generate and process them, are often insufficiently documented, which may result in misinterpretation of the data or may render the data unusable in later research. Likewise, lack of version control or use of persistent identifiers for all files causes downstream

confusion, and hinders reproducible science.

Data managers are increasingly asked to both preserve raw data streams and to additionally provide automated, near real-time quality control and access to provisional data from these sensors. Typically, these provisional data streams undergo further visual and other quality checking and final data sets are published. Commonly, further interpretation occurs where some missing data are gap-filled through imputation procedures, or faulty data are removed. There is a strong need to archive these data streams and provide continued access, which ultimately safeguards the investment of both time and money dedicated to collect the data in the first place. There are a number of organizational, storage, formatting, and delivery issues to consider. However, four main archiving strategies should be used: creating well documented data snapshots, assigning unique, persistent identifiers, maintaining data and metadata versioning, and storing data in text-based formats. These practices, described below, will increase the longevity and interoperability of the data, and will promote their usefulness to current and future researchers.

Methods

Publishing of Snapshots

Generating periodic snapshots of near real-time sensor streams allows the data to be stored and described in a deterministic manner. The rate that snapshots are produced depends on the needs of the community using the data, but typically snapshot files are organized using hourly, daily, weekly, monthly, or annual datasets. It also depends on the sample rate and sample size. Producing thousands of tiny data files, or one file with gigabytes of data, would decrease the usefulness of the data from a transfer and handling perspective. Make it easy on the researchers using the data, and size the snapshots appropriately.

Without detailed documentation of the contextual information needed to interpret individual measurements, even well-archived data will be rendered unusable. Develop metadata files to accompany the data using a machine-readable metadata standard appropriate to the community using the data. Common standards include the ISO 19115 Geographic Information Metadata [ISO/TC 211, 2003], the Content Standard for Digital Geospatial Metadata (CSDGM) [FGDC, 1998], the Biological Profile of the CSDGM (FGDC, 1999], and the Ecological Metadata Language [Fegraus et al., 2001]. Also consider documenting sensor detailed deployment settings and processes with SensorML [OGC, 2000].

Likewise, snapshots of data that represent a time-series should be documented and packaged appropriately such that the relationships among files are clear. Many of the above metadata standards have their own means of linking data with metadata, however they are all implemented differently. Federated archiving efforts such as DataONE have adopted ‘resource maps’ [Lagoze, 2008] to describe relationships between metadata and data files in a language-agnostic manner. (See DataONE packaging in the Resources section, and the Open Archives Initiative ORE primer). Consider publishing resource maps of your data and metadata relationships to improve interoperability across archive repositories.

Once data collections are sufficiently described, delivery can also be a challenge. While providing resolvable links directly to the metadata and data files is a good practice, scientists often would like to be able to download full collections. Providing a service that packages files into a downloadable zip file is commonplace, but relationships between data and metadata can be lost. Consider using the BagIt specification (see BagIt in the Resources section) [Boyko, 2009], which provides simple additions to zip files such as a manifest file that maintains the machine-readable relationships between the items in the collection, while still allowing researchers to download data packages directly to their desktop.

Persistent Identifiers

The above snapshot archiving strategies hinge on the ability to uniquely identify each file or component of a package in an unambiguous manner. File names can often collide, particularly across unrelated projects. So, assigning unique, persistent identifiers to each file, and the originating sensor stream, is paramount to successful archiving. A persistent identifier is usually a text-based string that represents an unchanging set of bytes stored on a computer. Persistent identifiers should be assigned to science data objects, science

metadata objects, and other files that associate the data and metadata together, such as resource maps. Opaque identifiers tend to be best for persistence and uniqueness (like UUIDs), but can be less memorable. Systems such as the Digital Object Identifier service (DOI) and EZID can help in maintaining unique, resolvable identifiers (see UUIDs, DOIs, and EZID in the Resources Section). Each version of a file or products derived from files (see versioning below) should also have a persistent identifier. If snapshots of data are being extended with new data, a new version of the dataset needs to be published. Shorter identifiers are best, and avoid using spaces and other special characters in identifiers to increase compatibility in file systems and URLs. Ultimately, the use of persistent identifiers allows associated metadata to track the provenance of cleaned, quality assured data or other derived products, and promotes reproducible science and citable data.

Versioning

Data from sensor streams are usually considered 'provisional' until they have been processed for quality control, and multiple versions of the data may be generated. However, provisional data are often used in publications and are cited as such. That said, in order to support reproducible science using sensor data, each version should be maintained with it's own citable identifier. Overwriting files or database records with new values or with annotated flags will ultimately change the underlying bytes, and effectively break the 'persistence' of the identifier pointing to the data. This applies to metadata or packaging versions as well, and so care must be taken to plan in versioning within your storage system. Your versioning strategies of raw data will be dependent on your snapshot strategies (e.g. appending to hourly files, then snapshotting and updating metadata files, or alternatively, say, producing daily, weekly, monthly, or annual packages that include data files and metadata files for the time period of covered). However, by making citable versions, researchers will be able to access the exact bytes that were used in a journal publication, and peer review of studies involving sensor data streams will be more robust and deterministic.

Data Storage Formats

Sensor data may be stored in different structures, each with its own advantages and disadvantages. A suite of variables from one station and collected at the same temporal resolution may be stored within one wide table with a column for each variable, each time being one record of several variables. Alternatives might be a table for each variable or one table of the format of [time, location, variable, value]. This latter system may be value centric with metadata attached to each value or series centric with metadata attached to a certain time interval for one variable (e.g., a time series of air temperature between calibrations). No matter how you organize your data, long-term, archival storage file formats need to be considered. In the digital age, thousands of file formats exist that are readable by current software applications. However, some formats will be more readable into the future than others. As an example, Microsoft Excel 1.0 files (circa 1985), are not readable by Microsoft Excel 2012 since the binary format has changed over time in a backward-incompatible manner. Therefore, unless these files are continually updated year after year, they will be rendered unusable. The same is true for database system files (.dbf) that hold the relational table structures in commonly used databases such as Microsoft SQL Server, Oracle, PostgreSQL, and MySQL. Database files must be upgraded with every new database version so they do not become obsolete. A good rule of thumb is to archive data in formats that are ubiquitous, and are not tied to a given company's software. Archive data in ASCII (or UTF-8) text files preferably, since this format is universally readable across operating systems and software applications. If ASCII encoding would cause major increases in file sizes for massive data sets, consider using a binary format that is community-supported such as NetCDF. NetCDF is an open binary specification developed at the University Cooperative for Atmospheric Research (UCAR), and provides open programming interfaces in multiple languages (C, Java, Python, etc.) that are supported by many scientific analysis packages (Matlab, IDL, R, etc.). By choosing an archive storage format that isn't tied to a specific vendor, data files will be readable in decades to come even when institutional support for maintaining more complex database systems falls short.

Data Storage Strategies

For local archives, the most common storage strategy is to just directly store files in a hierarchical manner

on a filesystem. Many data managers use a mix of location, instrument, and time-based hierarchy to store files into folders (e.g. /Data/LakeOneida/CTD01/2013/06/22/file.txt). This is a very simple, reliable strategy, and may be employed by groups with little resource for installing and managing database software. However, filesystem-based archives may be difficult to manage when volumes are large, or the number of instruments or variables are growing and don't fit a straight hierarchical model.

Many groups use relational databases to manage sensor data as they stream into a site's acquisition system. Well known vendor-based solutions like Oracle Database and Microsoft SQL Server are often used, as well as open source solutions such MySQL and PostgreSQL. These systems provide a means of data organization that can promote good quality control and fast searching and subsetting based on many factors, beyond the typical location/instrument/date hierarchy described above. Databases also provide standardized programming interfaces in order to access the stored data in standard ways across multiple programming languages. From an archive perspective, the use of databases may help in managing data locally, but should be seen as one component of a workflow to get data into archival formats.

Local databases used for managing data should be backed-up regularly, ideally to an offsite location, and the underlying binary database file formats should regularly be upgraded to the newest, supported versions of the database software. Ultimately, data stored in databases should be periodically snapshotted and stored in archival file formats, described above, with complete metadata descriptions to enhance their longevity.

Although local filesystems and databases are the most practical means of managing data, they are often at risk of being destroyed or unmaintained over decadal scales. Natural disasters, computer failure, staff turnover, lack of continued program funding, and other risks should be addressed when deciding how to archive data for the long term. One of the strategies for best data protection is cross-institution collaborations that provide storage services for their participants. These sorts of arrangements can guard against institutional or program disollution, lack of funding, etc. Consider partnering with community supported archives such as the LTER NIS, or federated archive initiatives such as DataONE to archive snapshots of streaming sensor data (see both in the Resource Section).

Best Practices

The following list of best practices are taken from the above recommendations, as well as additional considerations when archiving sensor-derived data.

- Develop and maintain an archival data management plan such that personnel changes don't compromise access to or interpretation of data archives (potentially through University Library programs)
- Employ a sound data backup plan. Archived data should be backed up to at least two spatially different locations, far enough apart that they won't be affected by the same destructive events (natural disasters, power or infrastructure issues). Perform daily incremental backups and weekly complete backups that may be replaced periodically, and annual backups that won't change. (Crashplan, acronis)
- Generate periodic snapshots of near real-time sensor streams (acronis)
- Develop metadata files to accompany the data using a machine-readable metadata standard
- Assign persistent identifiers to science data objects, science metadata objects, and other files that associate the data and metadata together
- Maintain versioned files with their own citable identifier
- Preferably archive data in ASCII (or UTF-8) for text files, or community supported formats like NetCDF for binary format

- Archive all raw data, but all raw data do not necessarily need to be available online.However, assign a persistent identifier to each raw data file to be able to document provenance of the published, quality controlled data.
- Partner across institutions to provide archival services to mitigate programmatic losses
- Preferably make data publicly available that have appropriate QA/QC procedures applied.
- Assign a different persistent identifier for published datasets of different QC levels in an archive. In the methods metadata, document the provenance and quality control procedures applied.
- Document contextual information for each data point. i.e., in addition to assigning a quality flag, assign a methods flag which documents field events like calibrations, small changes, sensor maintenance, sensor changes etc. Include notes that handle unusual field events (e.g., animal disturbance etc.) Encode metadata for sensor-derived data using community and or nationally accepted standards.
- Ensure the timezone for all time stamps is captured. Datalogger are being manually set to a certain time. Consider daylight savings.
- Establish the meaningful naming conventions for your variables taking into account the type of observation that is archived, adjectives describing the location, instrument type, and other necessary variable determinants.
- Determine the precision for your observation values in advance
- Preferably follow a naming convention or controlled vocabulary for variables (See the Resources Section)
- Avoid using databases for archival storage, but use them for management and quality control. However, if databases are used for managing sensor data then periodic snapshots into ASCII or open binary data formats are recommended.
- Track changes to data files within metadata files to maintain an audit trail

Case Studies

1. LTER NIS

The NSF Long-Term Ecological Research Network Information System (LTER NIS) is the central data archive for all data generated by LTER research and related projects. All data including sensor data are submitted with metadata in the Ecological Metadata Language (EML). Data are publicly available through this portal and through DataONE, of which the LTER NIS is a member. Specific approaches to archive streaming sensor data are following the best practice recommendations given in this document: Datasets are submitted as snapshots in time and it is up to the site information managers to decide the length of time in each snapshot, i.e., how frequently a new dataset is submitted. Minimally quality controlled data sets are submitted, while the raw data are archived at each site. As of this writing no standards for quality control levels or data flagging have been adopted by the LTER community.

Features of the LTER NIS include:

- Public availability of data and metadata
- Congruence check - quality check of how well the metadata describe the structure of the data
- Use of persistent identifiers
- Strong versioning of metadata and data files in the system
- Member Node of DataONE
- Support of LTER and related projects data storage using access control rules

- Replication of data and metadata across geographically dispersed servers

2. KNB

The Knowledge Network for Biocomplexity (KNB) is an international network that facilitates ecological and environmental research on biocomplexity. For scientists, the KNB is an efficient way to discover, access, interpret, integrate and analyze complex ecological data from a highly-distributed set of field stations, laboratories, research sites, and individual researchers. The KNB repository has been storing and serving data for over a decade, and stores over 25,000 data sets.

Features of the KNB include:

- Public availability
- Metacat, an open source data management system
- Morpho, an open source, desktop metadata editor
- Support for any XML-based metadata language, but optimized for the Ecological Metadata Language
- Use of persistent identifiers
- Strong versioning of all files in the system
- Support for cross-metadata packaging using resource maps
- Cross-institutional partnering with the LTER and DataONE
- Support for both public and private data storage using access control rules
- Replication of data and metadata across geographically dispersed servers
- International participation, and support for multi-language metadata descriptions

Recent developments of the KNB include support for the DataONE programming interface (API) in both the Metacat and Morpho software products. This API promotes interoperability of archival repositories, and enables federated access to environmental data. Since the KNB products support this open API, anyone can create their own web or desktop applications that are optimized for their research community.

3. GIWS, University of Saskatchewan WISKI data archive

Global Institute for Water Security (GIWS) at the University of Saskatchewan is directly involved in the collection of the field data from different research areas including Rocky Mountains, Boreal Forest, Prairie, and others.

In addition to the “in-house” managed data, GIWS uses external data sets from organizations such as Environment Canada and Alberta Environment. Data management platform on which GIWS currently operates is the Water Information System Kisters (WISKI). This system is used together with Campbell Scientific LoggerNet software and custom .NET modules in automated tasks that handle data collection, centralized data processing, storing, and reporting. After processing, the environmental data sets are published and made available to specific groups of users through the Kisters WISKI Web Pro web interface and KiWIS web service. Both applications can query the centralized database and return data in the formats that are used for visualization or further processing and dissemination purposes.

Features of the GIWS system include public availability, use of persistent identifiers, support for cross-institutional partnering, data access control for different groups of users, support for OGC WaterML2 data format. [See GIWS in the Resources Section]

Resources

BagIt Zip file format: https://wiki.ucop.edu/display/Curation/BagIt

DataONE: http://www.dataone.org/what-dataone and http://www.dataone.org/participate

DataONE Packaging: http://mule1.dataone.org/ArchitectureDocs-current/design/DataPackage.html

Digital Object Identifier (DOI) System: http://doi.org

Ecological Metadata Language: http://knb.ecoinformatics.org/software/eml/

FGDC Metadata Standards: http://www.fgdc.gov/metadata/geospatial-metadata-standards

GIWS: http://giws.usask.ca/documentation/system/GIWS_WISKI.pdf

ISO 19115 metadata Standard - Geographic Information
http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=53798

EZID Identifier Service: http://n2t.net/ezid

LTER Network Information System: http://nis.lternet.edu

Open Archives Intiative Object Reuse and Exchange (Resource Maps) Primer:
http://www.openarchives.org/ore/1.0/primer.html

Universally Unique Identifiers (UUID): http://en.wikipedia.org/wiki/Universally_unique_identifier

CF Metadata http://cf-convention.github.io/

References

Biological Data Working Group, Federal Geographic Data Committee and USGS Biological Resources Division. 1999. CONTENT STANDARD FOR DIGITAL GEOSPATIAL METADATA, PART 1: BIOLOGICAL DATA PROFILE. FGDC-STD-001.1-1999
https://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/biometadata/biodatap.pdf

Boyko, A., Kunze, J., Littman, J., Madden, L., Vargas, B. (2009). The BagIt File Packaging Format (V0.96). Retrieved May 8, 2013, from http://www.ietf.org/id/draft-kunze-bagit-09.txt

Fegraus, E. H., Andelman, S., Jones, M. B., & Schildhauer, M. (2005). Maximizing the value of ecological data with structured metadata: an introduction to Ecological Metadata Language (EML) and principles for metadata creation. Bulletin of the Ecological Society of America, 86(3), 158-168.
http://dx.doi.org/10.1890/0012-9623(2005)86%5B158:MTVOED%5D2.0.CO;2

Lagoze, Carl; Van de Sompel, Herbert; Nelson, Michael L.; Warner, Simeon; Sanderson, Robert; Johnston, Pete (2008-04-14), "Object Re-Use & Exchange: A Resource-Centric Approach", arXiv:0804.2273v1 [cs.DL], arXiv:0804.2273

Metadata Ad Hoc Working Group, Federal Geographic Data Committee. 1998. CONTENT STANDARD FOR DIGITAL GEOSPATIAL METADATA. FGDC-STD-001-1998.
http://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/base-metadata/v2_0698.pdf

Michener, William K., James W. Brunt, John J. Helly, Thomas B. Kirchner, and Susan G. Stafford. 1997. NONGEOSPATIAL METADATA FOR THE ECOLOGICAL SCIENCES. Ecological Applications 7:330–342. http://dx.doi.org/10.1890/1051-0761(1997)007[0330:NMFTES]2.0.CO;2

Open Geospatial Consortium, Inc. (OGC). Url: http://www.opengeospatial.org (2010).