Lex Van Horn

Matthew Mayernik

LIS 545 B: Data Curation I

Term Project Final Report [DRAFT COPY]

# Data and Metadata Profile

For this term project, I will be focusing on the data set “[Average Date When Lilacs Bloom in Idaho](https://catalog.data.gov/dataset/average-date-when-lilacs-bloom-in-idaho),” published by the University of Idaho and harvested from the Idaho Geospatial Data Clearinghouse by Data.gov. Key stakeholders for this data include the University of Idaho and its faculty and any researchers focused on phenological data, especially those focused on Idaho. According to the Data.gov site for this dataset, phenological data is “information on dates when plants and animals reach various stages in their development.” The data set does not appear to come with any usage restrictions. It’s listed on Data.gov as “public access,” but the link to the license type information is no longer active.

This data set is available on ArcGIS Hub and the Esri Rest API, and as GeoJSON, CSV, KML, and Shapefile downloads. Each of the downloads, other than the Shapefile download, include one file. The Shapefile download includes six files, one each of the following types: CPG, DBF, PRJ, SHP, SHX, and XML. The GeoJSON, CSV, and KML files can be opened and edited by several software types, but the Shapefile files can only be used in ESRI software, such as ArcGIS. Each version of the dataset contains an averaged date for the first lilac bloom based on records from 1957--1966 for designated regions within Idaho.

The metadata available with the data set varies based on which page or file is being observed. The websites where the files can be downloaded from have the most comprehensive metadata, although each has a different focus and purpose. The Data.gov site, following the [Project Open Data Metadata Schema v1.1](https://project-open-data.cio.gov/v1.1/schema/), includes fields for resource type, metadata creation and update dates, publisher, maintainer, identifier, data publish and modification dates, category, access level, metadata context, schema version, catalog describedby, harvest object id, harvest source id, harvest source title, homepage url, license, metadata type, source datajson identifier, source hash, source schema version, and spatial. This site seems focused on metadata related to where the data set originated. The ArcGIS page, which does not list which (if any) schema it is following, includes fields for feature layer, information updated date, next update date, published date, records number, access, copyright, and relevant area. This page seems focused on metadata related to the recency of the data set. The Esri Rest API page, which does not list which schema it follows, includes fields for name, display field, type, geometry type, description, copyright text, min scale, max scale, default visibility, max record, count, supported query formats, use standardized queries, extent, drawing info, hasZ, hasM, has attachments, has geometry properties, HTML popup type, object ID field, unique ID field, global ID field, type ID field, fields, types, is data versioned, has contingent values, supports rollback on failure parameter, last edit date, schema last edit date, data last edit date, and supported operations. This page seems focused on metadata related to how the data interacts with the API. The file downloads, in comparison, have barely any metadata. The GeoJSON file includes type, name, and crs in the file header. The KML file has the document id, schema name, and the field names in the file header. These metadata in these two files seem focused on making the file understandable and usable to future users. The CSV file doesn’t have any metadata in the file text itself, but does include the file size visible in the Information tab of Excel. Since I do not have access to ArcGIS software, I was unable to open the Shapefile files. None of the metadata found in the files themselves appear to follow a specific metadata standard other than what is expected to be included in the filetypes themselves based on code formatting.

Including more detailed information related to the individuals who participated in the creation of the data set within the Data.gov metadata could be beneficial for users searching data sets from specific researchers. In addition, the titles of the metadata fields could be named more clearly for users less familiar with the Project Open Data schema. Adding basic metadata from the Data.gov page (such as a link to the license information, publishing institution, and the initial publishing date) or fields crediting the original researchers and clarifying the “Shape Area” portions of the data to the downloadable files would be especially helpful for users unfamiliar with the data understand the data itself and reference it in future work.

There were no publications provided with the data set and searches of the data set’s title to both the University of Idaho library website and Google Scholar elicited only one publication: [Geospatial Data on Your Own Terms](https://arc.lib.montana.edu/ojs/index.php/pnla/article/view/406), which does not cite the data set, but references it on page 37.

# Repository Profile

After reviewing the available repositories, I determined that the “[Average Date When Lilacs Bloom in Idaho](https://catalog.data.gov/dataset/average-date-when-lilacs-bloom-in-idaho)” data set would best fit in the [USA National Phenology Network](https://www.re3data.org/repository/r3d100010362) (USA-NPN) repository. While the USA-NPN repository is not yet open for submissions of projects outside those it runs, its website states that it is planning to expand its capacity in the future to encompass existing phenology data sets not created by USA-NPN. I decided this would be the best spot for the data set I selected because both the repository focuses on the same country as that which the data originates from and the focuses on the same type of data (phenological data) in its collection scope.

As of now, the repository is not yet open for data submissions, but once the process is finalized, the repository seems like it will accept data submissions from anyone. The [Share a Dataset](https://www.usanpn.org/data/share) page states that it USA-NPN will accept data sets which do not utilize the USA-NPN protocols, but the methods via which these submissions could occur are not yet available on the website. A “coming soon” message is listed near the submission instructions.

Once these submissions open, it appears that the repository will accept observational and gridded phenology data. USA-NPN’s [Data Quality](https://www.usanpn.org/data/quality) page states that observational data in the National Phenology Database undergo quality assurance processes “which target species identification, phenophase status evaluation, and data entry” and quality control measures related to “plausibility, validity, and reliability.” The observational data currently available is collected by “naturalists, professional researchers, high school students and park visitors,” among others, according to the Data Quality page. The gridded data offered by USA-NPN consists of the Spring Indices and Accumulated Growing Degree Day maps produced by consolidating multiple sources of information. Once the repository opens to submissions, it seems that it will accept data types which fall within these categories. It’s unclear if additional categories will be added when submissions open.

So far, the repository only provides guidance for the Submission Information Package (SIP) in the form of requesting details regarding access rights, metadata, and a recommended citation for the data set. Metadata is not required to be submitted in a specific schema, but the submissions page requests that the schema used by the submitter is identified. The submissions page does not clarify what to do if there is no existing metadata for the data set or if it’s not based on an existing public schema. The USA-NPN repository’s existing data sets utilize [FGDC-compliant metadata](https://data.usanpn.org:3002/pop/fgdc), according to the USA-NPN [Observational Data Documentation](https://pubs.usgs.gov/of/2018/1060/ofr20181060.pdf). Those with feedback are encouraged to reach out to a designated support email, but no further detail is provided on human assistance or consulting regarding the submission process.

A log-in is not required to download data from the USA-NPN repository. Website visitors can go to the [Explore Phenology Data](https://www.usanpn.org/data) page and click on the link for “Download Observed Data” to visit the [Phenology Observation Portal](https://data.usanpn.org/observations/get-started), which allows visitors to select criteria in order to download a data set based on what is available in the USA-NPN repository which meets the visitor’s needs. The Dissemination Information Package (DIP) from downloading observational information automatically includes site ID, latitude, longitude, elevation in meters, state, species ID, genus, species, common name, kingdom, individual ID, phenophase ID, phenophase description, first yes year, first yes month, first yes day, first yes DOY, first yes julian, date, numdays since prior no, last yes year, last yes month, last yes day, last yes DOY, last yes julian date, and numdays until next no. The DIP can optionally include the dataset ID, observedby person ID, partner group, site name, species functional type, species category, lifecycle duration, growth habit, USDA PLANTS symbol, ITIS number, plant nickname, patch, phenophase category, numYs in series, numDays in series, multiple observers, multiple firstY, observed status conflict flag, AGDD, AGDD in F, Tmax winter, Tmax spring, Tmax summer, Tmax fall, Tmin winter, Tmin spring, Tmin summer, Tmin fall, Prcp winter, Prcp spring, Prcp summer, Prcp fall, accum prcp, or daylength.

Gridded data downloads are available via the [USA-NPN Geoserver Request Builder](https://data.usanpn.org/geoserver-request-builder/), which can be found by exploring the [Accumulated Growing Degree Days](https://www.usanpn.org/data/maps/AGDD) or [Spring Indices](https://www.usanpn.org/data/maps/spring) pages. I could not find specific DIP-related information related to this download request form.

It appears that only one access mechanism is available for observational data, as it can only be accessed or downloaded via the portal, but there are multiple access options for gridded data. Those options include the [Visualization Tool](https://data.usanpn.org/vis-tool/#/) and gridded data and related product downloads (available when individual options are clicked on the [Models and Maps](https://www.usanpn.org/data/maps) page). There is no method to query the database outright.

# Recommended Citation (APA 7th Edition)

Landscape Dynamics Lab (2006). *Average Date When Lilacs Bloom in Idaho.* [Data set]. Idaho Geospatial Data Clearinghouse. <https://geocatalog-uidaho.hub.arcgis.com/datasets/fcf608df8f4649838bd34527af8823d0_0>

# Considerations for Long-Term Preservation

Many of the file types used in this dataset rely on the continuation of Esri Rest and ArcGIS. While these are commonly utilized software now, they are proprietary and this dataset should be monitored in case these platforms cease to provide support in the future. Some of the other file formats – GeoJSON, CSV, and KML – are not proprietary and therefore not reliant on one specific piece of software. Shapefile file types, on the other hand, require ArcGIS in order to open.

# Copyright Statement

The copyright license used by this dataset is the CC By-NC-SA 4.0, provided by Creative Commons. This license allows the public to share and adapt the dataset as long as it is used for noncommercial purposes, adapted versions of the dataset are distributed under the same license as the original, and appropriate credit is attributed to the creators of the dataset. Based on the type of data being provided and the types of rights necessary to utilize datasets in future research, this seems like an appropriate copyright license for this dataset.

# Human Subject Considerations

This data set does not include personally identifiable data about people and therefore no extra steps will need to be taken to anonymize or adjust the data set to abide by best practices for privacy and human subject ethics.