# ENVIRONMENTAL SCIENCE / HERBIVORY – DATA MANAGEMENT PLAN

## DESCRIPTION

This is a fictional exemplar data management plan developed in May 2020 by Danielle Dennie (Research Data Librarian, Concordia University) for educational and guidance purposes. It is mostly based on information taken from a data curation profile which addresses a "data set generated from the study of the mechanisms and ecological consequences of plants' induced responses to herbivore damage" (Wright, 2012).

Wright, S. J. (2012). Environmental Science / Herbivory - Cornell University. *Data Curation Profiles Directory*, *4*(3). https://doi.org/10.5703/1288284315002

## DATA COLLECTION

*WHAT TYPES OF DATA WILL YOU COLLECT, CREATE, LINK TO, ACQUIRE AND/OR RECORD?*
The data collected will be field survey and bioassay data recorded in tabular format. There will also be Gas Chromatography – Mass Spectrophotometry (GC-MS) instrumentation data.

*WHAT FILE FORMATS WILL YOUR DATA BE COLLECTED IN? WILL THESE FORMATS ALLOW FOR DATA RE-USE, SHARING AND LONG-TERM ACCESS TO THE DATA?*
The field survey and bioassay data as well as the analyzed GC-MS data will be saved as .xls files in MS Excel 2007. However, the files will be converted into open format .csv files. Total expected file size is approximately 0.55 MB. The GC-MS data will be saved as proprietary instrument specific Saturn 2000 GC-MS files. The files can be then be converted into mzML, an open format for mass spectrometry files. Total expected file size is 38 MB. Statistical analysis files from Statview 5.01 will be saved as proprietary .svd, but these can be converted to .xls and subsequently .csv files. Expected file size is approximately 0.45 MB. Figures prepared for publication will be saved as .opj files from the OriginLab 8.5. These can be converted to the open form .tif. Expected file size is approximately 0.35 MB.

*WHAT CONVENTIONS AND PROCEDURES WILL YOU USE TO STRUCTURE, NAME AND VERSION-CONTROL YOUR FILES TO HELP YOU AND OTHERS BETTER UNDERSTAND HOW YOUR DATA ARE ORGANIZED?*
Files will be kept in separate folders: raw, analyzed, finalized. Files will be named according to a specified format which has been recorded in a readme file that is kept in the root project folder. The format follows the convention of ProjectAcronym_TypeOfData_FileCreatorInitials_Date. The date will be used as a version control. A version control table will be kept in the root project folder to indicate which versions were updated when, by who, and what was changed.

## DOCUMENTATION AND METADATA

*WHAT DOCUMENTATION WILL BE NEEDED FOR THE DATA TO BE READ AND INTERPRETED CORRECTLY IN THE FUTURE?*
The data is organized in Excel tables linked by a meta-table. In the meta-table, every experiment is described in detail and notes have been added to column headings describing specific treatments in detail.

*HOW WILL YOU MAKE SURE THAT DOCUMENTATION IS CREATED OR CAPTURED CONSISTENTLY THROUGHOUT YOUR PROJECT?*
Lab notebooks will be kept to document the research project and the experiments undertaken. Protocols for file naming, file versioning, folder structure, as well as how to capture and structure the experimental data within the Excel tables, will be written and stored in the project's root folder that all lab staff have access to.

*IF YOU ARE USING A METADATA STANDARD AND/OR TOOLS TO DOCUMENT AND DESCRIBE YOUR DATA, PLEASE LIST HERE.*
No specific metadata standards are being used, however, the meta-table described previously includes rich description of each experiment according to a set of defined inputs. The data should also be accompanied by basic study-level metadata such as the when and where of data collection, citation information (author, title, publication date, etc.), keywords, species, and geography. The Ecological Markup Language (EML), commonly used in ecology, includes many of these metadata elements, and we plan to explore the use of this metadata standard.

**STORAGE AND BACKUP**

*WHAT ARE THE ANTICIPATED STORAGE REQUIREMENTS FOR YOUR PROJECT, IN TERMS OF STORAGE SPACE (IN MEGABYTES, GIGABYTES, TERABYTES, ETC.) AND THE LENGTH OF TIME YOU WILL BE STORING IT?*
The anticipated storage requirement should be no more than 50 MB, which represents about 40 files. Given the small amount of space requirements, we plan on keeping our datasets indefinitely.

*HOW AND WHERE WILL YOUR DATA BE STORED AND BACKED UP DURING YOUR RESEARCH PROJECT?*
The active data is stored on a University networked drive which is backed up nightly. We also make backups every three months onto a hard drive which is stored in a different geographical location in case of fire or other local disaster.

*HOW WILL THE RESEARCH TEAM AND OTHER COLLABORATORS ACCESS, MODIFY, AND CONTRIBUTE DATA THROUGHOUT THE PROJECT?*
The networked drive where all active data are kept is accessible to all lab staff. The project folders are password protected. These passwords are kept by the PI and shared with the team members involved in the project.

**PRESERVATION**

*WHERE WILL YOU DEPOSIT YOUR DATA FOR LONG-TERM PRESERVATION AND ACCESS AT THE END OF YOUR RESEARCH PROJECT?*
We would like to deposit all data associated with our publications into a repository and to develop a workflow to continue the practice with future publications. Although we are aware of some repositories in ecology, such as the Knowledge Network for Biocomplexity, a consultation with the University's Research Data Librarian will help identify other possible repository options for our research data.

*INDICATE HOW YOU WILL ENSURE YOUR DATA IS PRESERVATION READY. CONSIDER PRESERVATION-FRIENDLY FILE FORMATS, ENSURING FILE INTEGRITY, ANONYMIZATION AND DE-IDENTIFICATION, INCLUSION OF SUPPORTING DOCUMENTATION.*
Most of the data is organized in MS Excel, and will be saved as .csv files for long term preservation. The laboratory chemical analysis data is generated using several different instruments which export the raw data in proprietary formats. The software necessary to use the chemical analysis data is standard among researchers in our field. Unfortunately, however, backwards compatibility is not a priority for the companies that produce the instruments and accompanying software. We will therefore export the data in the open.mzML format, however, it is not clear how much information in the original proprietary data-rich format will be lost.

**SHARING AND REUSE**

*WHAT DATA WILL YOU BE SHARING AND IN WHAT FORM? (E.G. RAW, PROCESSED, ANALYZED, FINAL).*
The raw data, including tables of field experiment data and the raw files from chemical analyses, are the most important parts of the data to share. However, the analyzed GC-MS data, the statistical analyses, and the finalized graphs for publication will also be shared.

*HAVE YOU CONSIDERED WHAT TYPE OF END-USER LICENSE TO INCLUDE WITH YOUR DATA?*
The data will be available to anyone without restrictions, however, data citation is important and will be a requirement of using the data. Therefore, the data will receive a CC-BY license.

*WHAT STEPS WILL BE TAKEN TO HELP THE RESEARCH COMMUNITY KNOW THAT YOUR DATA EXISTS?*
Ideally, the journal should provide a link between the data and the related publication(s). If the data is deposited in the KNB repository (https://knb.ecoinformatics.org/), a persistent digital object identifier (DOI) will be minted for the dataset thus improving chances of discoverability. This repository also generates a citation for the dataset, helping researchers give attribution to the work. Datasets in KNB are also harvested by Google as well as DataONE (https://search.dataone.org/data) and searchable/discoverable on those platform. Researchers will also find the dataset via databases like Web of Science, linked from the reference list in the original publication.

**RESPONSIBILITIES AND RESOURCES**

*IDENTIFY WHO WILL BE RESPONSIBLE FOR MANAGING THIS PROJECT'S DATA DURING AND AFTER THE PROJECT AND THE MAJOR DATA MANAGEMENT TASKS FOR WHICH THEY WILL BE RESPONSIBLE.*
Research data management will be a shared responsibility and will involve the Principal Investigator, one graduate student, and one research staff. Policies and procedures relating to research data management will be documented and everyone in the lab will be trained on these procedures prior to working on the project. The student will be responsible for collecting the data, maintaining a detailed lab notebook, and filling out the Excel spreadsheets with the data. The research staff will assist with running GC-MS samples, saving the raw data files in the appropriate folders, and making regular backups of all the research data. The Principal Investigator will oversee the research during the active phase of the project and ensure the safekeeping of the data after the project closure.

*HOW WILL RESPONSIBILITIES FOR MANAGING DATA ACTIVITIES BE HANDLED IF SUBSTANTIVE CHANGES HAPPEN IN THE PERSONNEL OVERSEEING THE PROJECT'S DATA, INCLUDING A CHANGE OF PRINCIPAL INVESTIGATOR?*
When the graduate student leaves, all the data, including the lab notebooks, are kept by the Principal Investigator. If the Principal investigator leaves the project, the department will assume responsibility over the project's data.

*WHAT RESOURCES WILL YOU REQUIRE TO IMPLEMENT YOUR DATA MANAGEMENT PLAN? WHAT DO YOU ESTIMATE THE OVERALL COST FOR DATA MANAGEMENT TO BE?*
Storage space on the university network drive and a personal hard drive. The entire project would also fit onto a 4-GB USB key with space to spare. Minimal long-term costs would be expected as long as the University maintains the network drive and the cost of replacing hard drives every 5 years is reasonable. The cost for preparing the data for deposit will also be negligible as the raw files and the master spreadsheet to organize the data by experiment are prepared and maintained during the active phase of the research by the Principle investigator.

**ETHICS AND LEGAL COMPLIANCE**

*IF YOUR RESEARCH PROJECT INCLUDES SENSITIVE DATA, HOW WILL YOU ENSURE THAT IT IS SECURELY MANAGED AND ACCESSIBLE ONLY TO APPROVED MEMBERS OF THE PROJECT?*
No sensitive data are included in this project.

*IF APPLICABLE, WHAT STRATEGIES WILL YOU UNDERTAKE TO ADDRESS SECONDARY USES OF SENSITIVE DATA?*
N/A

*HOW WILL YOU MANAGE LEGAL, ETHICAL, AND INTELLECTUAL PROPERTY ISSUES?*
The data will be available to anyone without restrictions, however, data citation is important and will be a requirement of using the data. Therefore, the data will receive a CC-BY license.