October 5, 2023

Data Management Plan for the Green SkEye project: Greenhouse Imaging and Analytics for Omics Innovation in Agriculture

A Data Management Plan created with DMP Assistant

Data Management Planning Expert Group

Information

Abstract: This exemplar DMP was created for a greenhouse plant imaging project to deliver analytics for plant images as well as phenomic and genomic data at the University of Saskatchewan. It also involves the development of computer software to collect, extract, and analyze plant genomic and phenomic data. In addition to general best practices in research data management, this DMP describes the various resources (such as GenBank, BioProject, GitHub, Zenodo, local computing facilities) used to manage the large scale of data generated in the project. It also provides an example on how to handle end-user agreements for sharing and reusing purposes when there are other parties involved in a project.

Creator: Lingling Jin

Affiliation: University of Saskatchewan

Funder: CFI JELF, Global Institute for Food Security

Template: Portage Template

ORCID iD: 0000-0002-4586-2347

Project abstract: Data-driven processes and decisions are increasingly important in the agriculture sector. However, while large amounts of farm and plant data are being collected by growers and scientists, the ability to create actionable information from these large datasets remains a key challenge. For plant breeding, it is critical to match genomic with phenomic data concerning the health and resilience of specific crop varieties. For growers, the rapid and sustainable management of weeds, insects and diseases in the field can be facilitated by crop imaging combined with novel data analysis techniques. In this project, the Green SkEye platform will be developed for detailed greenhouse plant imaging combined with cutting-edge computing infrastructure to deliver analytics for plant images as well as phenomic and genomic data. The ability to measure plant traits rapidly and precisely, and associate them with genomic and evolutionary data, will play a pivotal role in the development of new computational tools to support plant breeders worldwide, in turn creating innovations for on-farm decision making.

Identifier: 8989

Principal Investigator(s): Lingling Jin, Ian Stavness, Leon Kochian

Start date: 02-05-2022

End date: 30-04-2025

Last modified: 24-07-2022

Grant number / URL: 41499

Data Management Plan for the Green SkEye project: Greenhouse Imaging and Analytics for Omics Innovation in Agriculture

Data Collection

What types of data will you collect, create, link to, acquire and/or record?

There will be four types of data collected or created during the research process:

1) image and sensor data of the plants collected by the greenhouse imaging systems. For example, the imaging system can take RGB images, NDVI images or 3D reconstruction datasets from top or side of the plants to extract leaf number, plant height phenotypes, or track colour change to infer disease progression of the examined plants.
2) plant genomic data (i.e. SNP's identified from DNA sequencing) using the OPAL sequencing lab.
3) phenomic data extracted from the image and sensor data of the plants.
4) software code created by the research group to analyze plant image, sensor data, and phenomic and genomic data. Specific programming languages will be determined according to the purpose of different types of analysis.

> Research data can come in many different forms including such things as both quantitative and qualitative information, software code, audio-visual files, as well as discipline and instrument specific outputs.

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?

Each greenhouse imaging device will generate 4 data streams on the plants: a 12-megapixel colour image, a 12-megapixel near-infrared (NIR) image, a 2-megapixel colour image and a 2-megapixel depth image from the depth sensor. At this stage, we are considering various options for the file format of the images, such as PNG and TIFF for the colour images because these file formats are non-proprietary and retains data quality when compressed. The exact

format will be determined as the project proceeds based on the requirement of the analyses and the data storage size available.

The DNA sequencing lab (OPAL or Omics and Precision Agriculture Laboratory) at the Global Food Institute for Food Security operates a number of state-of-the-art DNA sequencers that will generate genomics data in FASTQ, a text-based format for storing both biological sequence and its corresponding quality scores.

Software code is also text-based format.

As stated above, the image file format (PNG or TIFF), and sequencing data and software code format (txt) are non-proprietary, thus can be re-used by and shared with other projects and allow for long-term access.

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?

Because of the large number of data files generated in the project, directories will be structured hierarchically to make it easier to navigate and locate these files.

The main directory of the image and sensor data files will be organized by the image type such as RGB, NIR and depth. Certain experiments will require high temporal data capture in order to characterize diurnal patterns of plant growth, therefore, these data will be organized into sub-directories by capturing date.

For the format of date and time, we will follow the recommendations of International Standards ISO 8601[1], and use YYYYMMDD format for date, and hhmmss for time. This format makes it easier to manage and disambiguate the large volume of files. It is also easily readable and writable by computer software.

For example, images for plant IDs “AC001” to “AC101” captured on June 1, 2022 will be grouped into a sub-directory named “20220601”.

Image and sensor data will be named according to plant identification number (ID), sensor ID, image type (RGB, NIR, depth, etc) and the date and time of image capture. For example, an NIR image of the plant ID AC001 captured by sensor RPI01 at 10:12:59, June 1, 2022 will be named as “AC001_RPI01_NIR_20220601_101259.png”.

Below is an example of a file path for an image file placed in a main directory and a sub-directory:

NIR > 20220601 > AC001_RPI01_NIR_20220601_101259.png

Processed images will be saved into a separate subdirectory under the main directory, with original file name and a brief description of processing action. For example, a cropped version of image AC001_RPI01_NIR_20220601_101259.png will be saved in a subdirectory named “NIR Processed”, with a new file name AC001_ RPI01_NIR_20220601_101259_crop.png

[1] ISO 8601 is an internationally recognized format for representing dates and times.

Genomic data and phenomic data will be structured by directories of different species and genotypes of plants, such as plant “AC001” to “AC101”.

Software development will be facilitated by using git (https://github.com/), a free open-source code hosting platform that provides functions for version control and file change tracking to coordinate work among group members developing source code collaboratively.

Documentation and Metadata

What documentation will be needed for the data to be read and interpreted correctly in the future?

A document will be created to describe the detailed methodology of the project and how each experiment will be conducted (e.g., soil condition, moisture, lights, disease control).

For each experiment, a spreadsheet will be used to document metadata of genotypes and corresponding images. For example, each row in the spreadsheet points to the file paths and names of genomic and different types of image data of the same plant so that software can locate associated data of the same plant automatically. Data is organized in self-explanatory sub-directories in most experiments.

Software development will have README files to help others understand and use the code in the future. The README files will include the project title, a brief description of the project, instructions on how to install the software, visual examples of demo data, usage instruction, contributors of the project, and license information.

Each software code will also have comments at the beginning of each file, which include information such as a brief description of the script, creation date, version, contributors, input and output to help others to understand it.

How will you make sure that documentation is created or captured consistently throughout your project?

We will develop an application to be used with each imaging device in the greenhouses. The application will read the QR code of plants being captured by the device and record the plant ID and date and time of capture automatically, thus allowing us to consistently document image data generated throughout the project.

Training will be provided to researchers involved in the project to make sure they are familiar with the data collection process. Researchers working in the greenhouses will be required to keep detailed records of plant selection, experimental conditions, etc.

As stated above, git will be used to facilitate software development, which ensures that all the documents and changes will be captured consistently throughout the project.

If you are using a metadata standard and/or tools to document and describe your data, please list here.

Due to the uniqueness of the project, we have not found any existing metadata standards to incorporate; therefore, the metadata for documenting and describing the data generated in the project is not based on any single standard but rather has been custom developed to support this project.

There are many established metadata standards spanning across research disciplines and methodologies to describe research data, such as Dublin Core, MODS (Metadata Object Description Schema), and DDI (Data Documentation Initiative). However, there is not always an 'out of the box' metadata standard that will address the needs of a given research project, and so custom solutions are at times warranted. Researchers may connect with metadata experts/librarians available to discuss the specific details of the metadata to describe the data.

For metadata of image and sensor data, an application/software is being developed to read plant QR code which documents and describes the plant, such as Plant ID, scientific name, grower, date and time captured, etc., as described above.

Metadata of genomic and phenomic data will include project information, organism information, sequence information (such as sequencing lab, sequencing platform, etc), and phenotype information (such as temperature range, disease, antimicrobial resistance, etc).

Metadata for software code will be generated in git, as stated above.

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?

It is estimated that the project will generate several terabytes of data, including images, genomics, phenomics, and software code. The exact amount of storage needed will depend on the number of plants and duration of experiments.

The data will be stored locally on the Copernicus Cluster storage (https://wiki.usask.ca/display/ARC/Copernicus+Cluster), a high-performance computing infrastructure at the University of Saskatchewan. We plan to retain the data for a minimum of 5 years to 2027, the year when there is guaranteed funding for Copernicus.

How and where will your data be stored and backed up during your research project?

Raw data will be transferred from imaging equipment or sequencing lab using Globus, a software provided by University of Saskatchewan IT unit for fast and reliable transferring of many files and/or large files, to the Copernicus Cluster storage.

> Research data generated within a research project can at times be very large in magnitude, as well as distributed across systems and institutions. Consider as early as possible how you will transfer the research data efficiently, securely, and reliably across systems when planning the data management.

Files stored on Copernicus are automatically backed up on a daily basis by the IT unit of the University of Saskatchewan. The IT unit also has data protection in place for the long-term storage on Copernicus, which preserves the last two copies of all files on long term storage to tape backup media.

How will the research team and other collaborators access, modify, and contribute data throughout the project?

Members in the research team will be able to get approval from Co-PIs to access, modify and contribute data throughout the project on Copernicus with their university user ID (USask NSID) and password. Other collaborators will be able to access data by applying for a guest USask NSID and password from the university IT unit and they will have read-only permission, also upon the approval of Co-PIs.

Preservation

Where will you deposit your data for long-term preservation and access at the end of your research project?

After the project is completed, a final version of the data will be preserved in and accessible by the research group from long-term backup storage on Copernicus within the next five years. Because of the large file size, we will also apply for long-term backup storage from Compute Canada, which preserves the data for long-term access.

Raw genomic sequencing data will be deposited to the GenBank database, a genetic sequence database of the National Institute of Health, which provides long-term preservation and allows public access . GenBank assigns a unique permanent accession number to each complete record, which facilitates sharing and discovering the genomic data. A sample GenBank Record with accession number can be found here.

> Research disciplines may have different expectations in terms of where to deposit research data. For example, in this case, nucleotide sequences are often expected to be deposited in GenBank. Researchers may wish to connect with experts, including institutional Librarians, in learning about and selecting an appropriate repository.

We will archive software code developed in git in Zenodo, an open data repository. Zenodo will also issue DOIs for the software code to make it easier to discover and reuse. This process can be completed by logging into Zenodo using Github credentials which will enable access to the repository; detailed steps are described here.

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.

Raw and processed data will be cleaned up from local storage and turned into a final version for preservation. Raw data files (PNG or TIFF, and text files) will be archived and compressed using lossless methods so that they are ready for preservation. We choose lossless compression to allow other researchers to reproduce the research, particularly from the image data.

As stated above, PNG, TIFF, and txt file formats are non-proprietary, preservation-ready formats.

Sharing and Reuse

What data will you be sharing and in what form? (e.g. raw, processed, analyzed, final).

Analyzed image data will be shared as a public dataset to facilitate other researchers' projects.

Raw sequencing data deposited in GenBank with unique accession numbers will have long-term and public access to the community.

Have you considered what type of end-user license to include with your data?

Because this project involves industry partners, the research group and industry partners will have governing data agreements for specific experiments, some of which may have restrictions on end-user license due to commercial reasons. We will consider a Creative Commons license (CC BY) for the data, which allows others to use our data with attribution to us, if permitted by the governing data agreement for specific experiments between the research group and industry partners. Otherwise, the type of end-user license will be based on the terms and conditions as specified in the agreements.

What steps will be taken to help the research community know that your data exists?

Processed and labelled high-quality image data will be published on platforms such as Kaggle (https://www.kaggle.com/), an online community platform of data scientists, to share with the research community.

We will register the project in the BioProject database (https://www.ncbi.nlm.nih.gov/bioproject) of the National Library of Medicine. BioProject is a collection of biological data related to a single project from an organization. It assigns each project with a unique Project ID and provides users a single place to find links to the various biodata generated from the project and deposited into the archival databases such as GenBank.

The Project ID for our research and the accession numbers for the genomic data deposited in GenBank will be published in corresponding manuscripts. Software pipelines will be deposited to GitHub and have DOIs through Zenodo. These efforts will make it easier for others to find and retrieve the data, and to reproduce the results.

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.

Co-PIs of this project will be responsible for managing the project data during and after the project. Project collaborators who run experiments and generate data using the greenhouse and sequencing facilities will be responsible for managing the data of their own projects.

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?

There are 3 co-PIs for this project. If one co-PI leaves the project, the team will have thorough discussions regarding the procedure of handing over and managing data before he/she leaves. The details of the procedure will be well documented for other group members to follow.

What resources will you require to implement your data management plan? What do you estimate the overall cost for data management to be?

The USask Copernicus data center already has sufficient facilities and storage to implement the data management plan.

The only other cost is hiring students to develop the software, which will be supported by researchers' grants.

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?

Not applicable.

If applicable, what strategies will you undertake to address secondary uses of sensitive data?

Secondary uses of the commercially sensitive data would be determined by the agreements with industry partners.

How will you manage legal, ethical, and intellectual property issues?

Data generated in the project will be used based on terms and conditions as specified in the agreements between the research team and the industry partners.