ADMIN DETAILS
Project Name: Professor Chartreuse Plan
Principal Investigator / Researcher: Julie Timm, Yichao Liang, and Roberto Abarca Mero
Description: Professor Chartreuse's work on the science of science, which tries to better understand how researchers and scientists work together to generate new knowledge.
Institution: JCU

DATA COLLECTION

WHAT TYPES OF DATA WILL YOU COLLECT, CREATE, LINK TO, ACQUIRE AND/OR RECORD?
Medical records consist of text files and transcripts. Instrumentation data will also be collected with some of the research, which will also be number and text 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?
Because the data being collected is mostly open source medical research, it would be best to create an XML database to convert all the JSON records. XML is easy to transform into Excel if needed and it is easy to maintain in the long term. All the existing Excel files should be converted into CSV and added to the new database. Normal naming and filing conventions should be followed when creating the records. The keywords that Professor Chartreuse has already used for searching will be added as metadata to the XML database to make for easier searching and recall. As well, dates and key words will be added to make the data searchable and retrievable. This database will be created in the cloud storage for JCU so that all collaborators will have easy access to it. XML is also the preferred format of PubMed.

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?
In order to name and organize the data files already existing and the ones that will be generated, a naming convention has been designed. The main guidelines are the following:
● ● There will be a clear classification of the types of files in record (transcript, photo, etc.).
● No spaces or special characters.
● Underscores will be used instead of periods or spaces.
● File names must be descriptive
● Dates of entry must be included (international standard for date notation YYYY_MM_DD or YYYYMMDD)
In order to assure that no information would be lost and make sure of the reliability of the data for this project, the data management plan has established the following version control program:
● Every week there will be a data migration to the data storage. Therefore, there must be a version control file for each week's final file.
● Each version will be organized through a software such as GitHub to simplify the data management process.
● Every time a new file is created or updated there will be a record / comment on the latest update to identify the recent changes.

DOCUMENTATION AND METADATA

WHAT DOCUMENTATION WILL BE NEEDED FOR THE DATA TO BE READ AND INTERPRETED CORRECTLY IN THE FUTURE?

There are different types of data, some needs documentation to make the data usable by other researchers and other do not need documentation because it is understood from the title. For the data that required documentation such as the science of science, the documentation must include: research methodology used, sample size, variable definitions, assumptions made, format and file type of the data, a description of the data capture and collection methods, explanation of data coding and analysis performed (including syntax files- if available).

The PubMed data will also require an explanation of how the data was collected and how it can and cannot be used, even though it is open source.

As part of the research, it is advised to follow the DDI standard. As part of it, an XML file will be included containing all of the necessary information of the structure. Thus, making it easier to understand and for machines to read at the same time.

HOW WILL YOU MAKE SURE THAT DOCUMENTATION IS CREATED OR CAPTURED CONSISTENTLY THROUGHOUT YOUR PROJECT?

To ensure that the measurements and data creation are consistent, each contributor must follow the format established when all the old data has been imported. Meaning that all data and information must be dated chronologically, and the contributor must use the standard terminology already established. And enter the following information: date, name, keywords, publications that used the data, and authors.

IF YOU ARE USING A METADATA STANDARD AND/OR TOOLS TO DOCUMENT AND DESCRIBE YOUR DATA, PLEASE LIST HERE.

As mentioned before, to maintain the relevancy and reputation of the research, DDI standard has been selected as metadata standard. The DDI is in XML format, which will enable the effective exchange of information

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 data accumulation has been an average of 5GB per year over the course of 4 years. If this trend continues storage will need to be able to grow about 5GB per year for the life of the research which is estimated to last at least another 30 years. The way the data will be stored will be the following:

- Data generation (JCU Data servers) 200 GB
- Backup server (GitHub) 200 GB
- Backup Hard Drive 200 GB

The lifespan of medical data is estimated at 15-20 years, however, because of the extra storage estimated into the storage locations no data will have to be deleted.

HOW AND WHERE WILL YOUR DATA BE STORED AND BACKED UP DURING YOUR RESEARCH PROJECT?
The data will be saved on the server (JCU), external hard disks (in Chartreuse's possession) and internet cloud (GitHub). After the end of the project, funds will have to be allocated towards the maintenance of the archived records.

HOW WILL THE RESEARCH TEAM AND OTHER COLLABORATORS ACCESS, MODIFY, AND CONTRIBUTE DATA THROUGHOUT THE PROJECT?
Each team and collaborators can access it through internet and web application. Each team will have permission to edit the relevant parts. Once that is done, the information will be passed onto the respective server and backup systems. The easiest way to collaborate will to have all researchers involved sign up for a GitHub account.

PRESERVATION

WHERE WILL YOU DEPOSIT YOUR DATA FOR LONG-TERM PRESERVATION AND ACCESS AT THE END OF YOUR RESEARCH PROJECT?
Besides distributing research data on the research project website, a copy of the research data will be deposited in repository Open Science Framework for long-term preservation. The research community will have access to the repository through Archivematica, digital preservation system.

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.
The data will be migrated to new formats. Excel files will be converted into XML format, which is more preservation-friendly. Metadata and documentation that will be deposited alongside the data to make data discoverable and reusable. Metadata will include keywords, types of data, created dates about each file. Related information including references, research reports, the original research proposal will be deposited with the research data.
Normalization is necessary when preparing data for preservation.

WHAT DATA WILL YOU BE SHARING AND IN WHAT FORM? (E.G. RAW, PROCESSED, ANALYZED, FINAL).
All the PubMed data and the transformed visualizations of the data will be made available to graduate students and to other researchers. Thus, is they need access to the raw files or the analyzed data it will be granted.

HAVE YOU CONSIDERED WHAT TYPE OF END-USER LICENSE TO INCLUDE WITH YOUR DATA?
The end-user license should be: cc by-nc-nd. Other researchers will need to be able to publish the data is they use it. however, Chartreuse has put considerable effort into gathering and modifying the data so he should be acknowledged and should be the only one to adapt what he has already worked on.

WHAT STEPS WILL BE TAKEN TO HELP THE RESEARCH COMMUNITY KNOW THAT YOUR DATA EXISTS?
JCU has a dedicated webpage for researchers. The research will be announced on this page as well as through word-of-mouth and publications that use this data/research.

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.

The responsibility for managing the data management project will be the team in charge of the creation of it. As part of the responsibility, the data management team has not only developed the data management plan but has also created the main milestones for its implementation. Once the data management implementation is completed the data management will fall under Professor Chartreuse and whoever he designates to be his data management consultant.

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?

Professor Chartreuse should designate two other colleagues to help manage his research in the event that he is no longer able to. A copy of this plan should be kept as a guide to future researchers involved in this project.

WHAT RESOURCES WILL YOU REQUIRE TO IMPLEMENT YOUR DATA MANAGEMENT PLAN? WHAT DO YOU ESTIMATE THE OVERALL COST FOR DATA MANAGEMENT TO BE?

GitHub Organization accounts will cost $9 monthly per user. The first five Organization account users will cost a flat fee of $25 per month. A hard drive with 1TB of storage will cost $50-120. And the JCU cloud storage comes with the research grant proposal at no extra cost and will include funding to cover the cost of GitHub and a hard drive.

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?

The database will be accessible only with a username and password through a GitHub account or through the cloud at JCU. This data is not generally sensitive, as it is already found in another location as open data, but the data will be double checked for sensitive information. The ‘sensitive’ information will not be available to the public. The data should only be shared via GitHub, the JCU cloud, or Chartreuse's hard drive. The data should never be shared via email or other non-secure methods.

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

An agreement for intellectual property rights will be prepared, that researchers will have to agree to before they have access to the data. The research standards at JCU should also be consulted.

HOW WILL YOU MANAGE LEGAL, ETHICAL, AND INTELLECTUAL PROPERTY ISSUES?

The research does not contain sensitive information. As part of the ownership of the research, the property rights for this research are: CC BY-NC-ND. This type of licensing allows the findings of the research to be used, if they're attributed to Professor Chartreuse directly and only in a research environment. This copyright will not allow anyone use it for commercial purposes. However, Once the new research is done, the researcher can set up a new type of licensing for their new findings.