# Professor Chartreuse Plan

## DMP title

### Admin Details

**Project Name:** Professor Chartreuse Plan

**Principal Investigator / Researcher:** Julie Timm

**Description:** Professor Chartreuse is expanding his personal research to include more collaborative work with colleagues both here at JCU and at other institutions. His work focuses on the science of science, which is a very meta area of research that tries to better understand how researchers and scientists work together to generate new knowledge. The data that Professor Chartreuse collects is mostly available from public domain resources. His most important collection comes from records published in a database called PubMed that indexes a wide variety of medical research. These records are available via a web API and are served in JSON. Thus far, Professor Chartreuse has used keyword searching to find the records of interest to him, and has transformed those JSON records into a series of large and complex Excel files. This has worked fine for Professor Chartreuse up to this point, but has already caused some headaches as he has tried to share his data with potential collaborators. The large volume of excel files has led to confusion as to which contains which type of data, and which is the most up to date. One of Professor Chartreuse’s colleagues hopes to expand their dataset to include both records from other databases (such as Wikidata) as well as survey and interview data with scientists. The data collected is used to write papers and create exercises for graduate classes. Professor Chartreuse also regularly visualizes portions of his data, both for substantive research reasons, and also in visualization demonstrations for his students. Professor Chartreuse has a few specific issues that he mentioned would be helpful to include in a data management plan: ● He said he’s interested in maybe consolidating his data, and has heard that something called an ER diagram or a schema might be helpful. ● He is concerned about how to make sure his data is consistent and protected. He once lost an excel file and all of the data associated with it, forcing him to go back and re-do a month of data collection. ● Professor Chartreuse has been gathering the public data that he uses by hand. He searches by keyword, and then types details from the returned results into his excel spreadsheets. He wonders if there’s a better way for him to collect this public data. ● He is worried about how best to both share his data with collaborators, and also with other science of science researchers who might benefit from his curated datasets.

**Institution:** Portage

### 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.

#### 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. 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?

Question not answered.

### Documentation and Metadata

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

Question not answered.

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

Question not answered.

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

Question not answered.

### 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?

Question not answered.

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

Question not answered.

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

Question not answered.

### 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 CSV 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. The digital preservation system used is Archivematica. Archivematica uses normalization as its technique. Archivematica will prepare digital files and upload them to the repository automatically. The process can be easily monitored and controlled since Archivematica provides a GUI interface.

### Sharing and Reuse

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

Question not answered.

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

Question not answered.

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

Question not answered.

### 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.

Question not answered.

#### 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?

Question not answered.

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

Question not answered.

### 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?

Question not answered.

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

Question not answered.

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

Question not answered.