Data Management Consultation Report for Professor Chartreuse

        This project generates science of science research data that will be gathered by Professor Chartreuse, distributed to his colleagues for collaborative purposes and to his graduate students at JCU for educational purposes. The data is predominantly from public domain resources, and will maintain the licensing of PubMed, which is the database Professor Chartreuse deems most valuable to his research. The data will require different levels of accessibility due to their sensitive nature. Professor Chartreuse and his colleagues require full access to the data, and students should be able to download the data without altering the original dataset. The data generated by the research team will be a revisable dataset. Professor Chartreuse is solely responsible for data storage and providing access to other parties. At the moment, Professor Chartreuse stores all of his data in large Excel files, which creates an obstacle as Excel files have a large margin of error within the dataset and are difficult to use for sharing with his team.  Our data management team suggests that the 20GB of previously generated data will be converted from JSON to CSV files, while new data that is produced will be saved as CSV in order for the files to be machine readable. Prior to converting the 20GB of previously accumulated data, we suggest working with the research team to assess what data is needed, what data should be disposed of, and the data that should be archived. We will also work to anonymize any data to ensure they do not contain any identifiers. Up until now, Professor Chartreuse has gathered public data by hand by searching for keywords. Our team suggests using Perl as the programming language. Professor Chartreuse will be able to extract from the dataset using a graphical user interface. All saved data files will be deposited in cloud storage using Open Science Framework.

        Open Science Framework will allow for secure access to the dataset. Professor Chartreuse will have the ability to control who has access to the data by adding and removing contributors to the project, allowing him to tailor the privacy to his team of collaborators while sharing data with his graduate students. OSF will allow Professor Chartreuse to fork a project to share with his graduate students. Open Science Framework is open source, free to use, and has version control capabilities which will reduce the risk of losing valuable data. OSF allows contributors to tag the research data using keywords and controlled vocabularies to increase its findability and to be able to locate data quickly. This function will provide ease of data sharing between the research team and assist with reproducibility amongst other researchers using OSF.

Professor Chartreuse stipulated that his data have the date, name, keywords, publications that used the data, and authors as a minimum requirement for metadata standards. In addition to having the ability to search by these queries, the team will have the ability to add a license, description of the project, and an affiliation using OSF. We recommend that Professor Chartreuse and his contributors input information in all of the metadata fields on OSF to ensure all of the metadata attributed to their projects is standardized, and therefore easy to query in OSF. In addition, it is crucial that the team uses a standard naming convention for saving all of the data

files to improve searching. These steps will all contribute to ensuring quality data which will comply with the guidelines of PubMed.

We suggest that our team works with Professor Chartreuse to write ontologies that will create relationships between the concepts he is studying within the domain of SciSci. We also recommend that he installs OWL2Perl on his devices. We can work with Professor Chartreuse and his collaborators to train them to have an understanding of the ontologies created and Perl’s ability to parse the information from the schema, which assists in effectively extracting value from the data. This step will help to find patterns within the dataset. Lastly, should the team require a citation software, Zotero and Mendeley are both compatible with OSF. Our data management team believes that this feature will be beneficial with sharing research within the academic community in PubMed.

While OSF’s proprietary storage, called OSF Storage, allows for 5GB per individual file to be uploaded, Professor Chartreuse can choose to upgrade to an add-on cloud storage provider if his files surpass that size limit. In this case, we recommend that the team uses G Suite for Higher Education as it is free unlimited storage and will help to implement open collaboration between JCU and other institutions and is compatible with OSF. Both OSF and G Suite for Higher Education encrypt the data which will provide further security. Professor Chartreuse stated that he would like to maintain his data until he no longer works at JCU, but he may also receive up to 15GB of storage for free through Google Drive when he does decide to leave or retire.

If there is data that the research team does not need quick access to, we recommend registering the project in OSF. Once the project is registered, it is archived to OSF and will be

protected as it will no longer be able to be edited or deleted. The data will be preserved by OSF; they use Rackspace as a primary server and Amazon Glacier as a backup. OSF also has a fund for preservation which will allow data to have read-access for approximately 50 years should OSF close. The research team will be producing data and disseminating it at a later date, so long-term storage is essential for reproducibility.

In conclusion, we feel that maintaining the steps of the data lifecycle can easily be accomplished with Professor Chartreuse and his research team. We want to place emphasis in providing access to the data, which is why we suggest using Open Science Framework, as the research will contribute to the work of other researchers and will ultimately benefit science of science as a whole by enabling the data to be reused. We feel that the cloud storage option lends itself to being able to easily keep track of various contributions to the projects, while maintaining security and privacy measures. Lastly, preserving the data is a fundamental step in the data lifecycle and we feel that Open Science Framework provides enough backup servers and access control for this project.

References

Google cloud for higher education (2018). Retrieved from <https://edu.google.com/intl/en_ca/higher-ed-solutions/>

Open Science Framework (2018). Retrieved from https://osf.io/