Database Management Systems – Group Project

(MGMT 4540 / BUSI 6516 / INFO 6540)

Jay Jacobson, David Lanteigne, Alexander Sakellis & Nicholas Thompson

Data Management Consultation Reports

Periwinkle

Green

&

Pinkerton

Data Management Plan

Professor Periwinkle – Marine Wildlife Research

Executive Summary:

Outlined in this report is a data management plan, which includes recommendations for database management, data storage, and data sharing. It is understood that there has been a considerable amount of data collected and shared amongst research team members and other interested parties since 1998.

Best practice calls for data to be made available as *open data*, and follow the FAIR (findable, accessible, interoperable and reusable) guiding principles for data management. This is more than just a best practice, it is also a requirement for publicly-funded research, meets the university's policy on research data, is often required by peer reviewed journals, and shows accountability for public funds.

We have prepared this report based on requirements outlined by the university librarian and input from the research project manager, Dr. Periwinkle, to be thorough and of benefit to the research team.

1. Project Description

Through funding from Innovation Canada, and the Research Excellence Fund, Professor Periwinkle and her team have been monitoring marine wildlife using a variety of sensor techniques, including tags and arrays. This research collects copious amounts of raw sensor data which is converted to Network Common Data Form (NetCDF) format. This research is also supplemented with the generation of complex simulations run from the data, saved as csv files. Looking forward, the intent is to be able to share this research with other research teams, as well as provide a means of sharing contributions between members. The envisioned collaborative space would be similar to the Integrated Ocean Observing System (IOOS) managed by the National Ocean and Atmospheric Administration (NOAA) in the United States, or the European Marine Observation and Data Network (EMODnet). Additionally, it is the hope of researchers to share data through the Ocean Biogeographic Information System (OBIS), one of the world's leading databases for ocean biodiversity and biogeographic data (OBIS, n.d.).

2. Documentation, Organization, and Storage

Based on consultations with the research team, it is understood that along with raw sensor and array data, researchers record field notes using *Darwin Core* and then convert these file to *NetCDF* format, as well simulations are run using this data and saved in csv file format. That being said, other contributors are not required to use this format. It is also estimated from the information provided that approximately 705 gigabytes (GBs) of data has been collected in the past and this will continue to grow at approximately 800 megabytes (MBs) per day or 24 GBs per month. This value is displayed monthly as it has been explained that data is collected on a monthly basis.

In order to properly store and backup this valuable data, it is proposed that a local storage server be created in addition to a cloud-based storage solution. The local storage server should be a Redundant Array of Independent Disks (RAID) system. This system adds a layer of data redundancy by storing the files across multiple drives which protects against failure (Jacobi, 2012). For an added layer of security, it is recommended that the drives be organized using RAID 5 and that three 5 Terabyte (TB) hard drives be used, this ensures that data is backed up across all of the drives.

This configuration will allow for 5 TBs of storage as it utilizes data parity across all of the drives (Jacobi, 2012). With the price of storage continually declining, this storage solution is expected to cost $150 per hard drive and $100 for the hard drive enclosure. This would bring it to a grand total of an estimated $550.

In terms of a cloud storage solution, this will be required so that citizen scientists and global research teams are able to collaborate and share data easily. It is proposed that Sync.com be used by the team for cloud storage. This service provides users with end to end encryption, and ample storage for users all at an affordable price (Sync.com, 2018). Per user, each can store up to 10 terabytes of data, which will allow researchers to continue to add data at the same rate for some years to come. Each user license costs $15 per month and members are able to set up a free account (with limited functionality) which would be ideal for other researchers looking to download the data (Sync.com, 2018). This places the cost of online storage at $540 per year for three licenses.

Once the RAID equipment and cloud storage service has been acquired, it will be important that all existing data stored on floppy disks, CDs, and DVDs, be uploaded to the cloud.

Going beyond the equipment and cloud service to store the data, it is important that data standards and file naming conventions be adhered to. As mentioned through conversations, it is the goal of researchers to share data with OBIS at some point in the future. With that in mind, it is recommended that OBIS data standards be implemented at this time. This is particularly important with regards to three pieces: the species occurrence data, dataset metadata, and sampling specific data (OBIS, n.d.).

With regard to species occurrence data, the use of Darwin Core should continue to be used. Dataset Metadata however should be configured to follow Ecological Metadata Language (EML). This metadata standard was developed in cooperation between OBIS and the Ecological Society of America and associated efforts, and has a minimum requirement of four terms; title, citation, contact, and abstract (OBIS, n.d.).

Finally, for sampling specific data, like organism, geographic feature, depth, or other metrics, Darwin Core Archive and OBIS-ENV-DATA standards should be used (OBIS, n.d.). This archive system acts as a star schema with each of the sample data pieces, like organism, as ID columns. Using this archival standard will make searching the data very easy and highly standardized as you will be able to search for specific ID columns.

When applying this information with the FAIR model of research data management, having these storage solutions and data standards insures that the data is both findable and interoperable.

3. Access, Sharing and Re-use

The ultimate goal of using the recommended data storage and data standards is to both ensure that a backup of the data is created, and provide a means of collaboration and sharing within teams and external parties. Although citizen scientists' data follow open data policies, appropriate licenses should be applied to publicly funded research. This is done to protect the intellectual property of the researchers and adds accountability to the research.

The Canadian Government provides funding to researchers through several funding agencies. It is understood that the research and infrastructure of Dr. Periwinkle's team is funded by Innovation Canada and the Research Excellence Fund. Both of these agencies provide funding through the Government of Canada and thus have policies in place for the data sharing and licensing.

In accordance to the rules of The Natural Sciences and Engineering Research Council of Canada (NSERC), “grant recipients are required to ensure that any peer-reviewed journal publications arising from Agency-supported research are freely accessible within 12 months of publication, either through the publisher's website (Option #1) or an online repository (Option #2)” (NSERC. 2014).

To best protect the intellectual property of research teams, it is recommended that Creative Commons licences be applied to the work of Dr. Periwinkle and her team. Specifically, the license should be a Creative Commons attribution requiring Share Alike (CC-BY-SA). This license requires that users of the data give appropriate credit, provide a link to the license, and indicate any changes that were made to the data. Additionally, the ‘Share Alike’ element of this license ensures that if another person were to transform or build upon the original material they must also distribute this material using the same license. By applying this license to the research being conducted it ensures that the intellectual property of the data is not lost while still ensuring that it is available.

Currently, research assistants and students that have access to the team’s data are not required to sign a contract indicating that the data and findings are owned by Dr. Periwinkle. In addition to applying a license to the data, it is also recommended that a contract be instituted to ensure that ownership of the data is clear and legally binding. Assuming that local storage and cloud storage repositories will be created for the data, this ensures that both team members and interested parties will have available access to the data.

Following the data lifecycle, data starts at the ‘creation’ phase and ends at the ‘reuse’ phase after the life of the project has ended and data is no longer being added. In order to ensure that this ‘reuse’ phase is reached, it is important that the data be properly preserved and able to be accessed. It is assumed that all of the data will retain its value after the life of the project given that there are many insights that can be obtained through sensor array data and field notes. It is imperative that this research be preserved so that future research can be conducted using the datasets and that it can be used to inform new research endeavours.

4. Archiving

As previously mentioned, it is the goal of the Dr. Periwinkle to eventually share the data through OBIS’ database archives. This will require that data follow the Darwin Core Archive format as well as the other data standards set by the organization. Additionally, this archived data can be stored using OBIS' infrastructure which will free up space in the research team's storage and provide an easy means of re-use for future research.

5. Conclusion

Based on the information provided, it is understood that the sharing of data is important. Through this holistic database plan it has been recommended that new practices and policies be put in place to ensure data is properly managed. This includes the local backup of data through a local RAID storage server as well as implementing the cloud based storage solution of Sync.com.

Protection of intellectual property is also a focus for the plan and it is advised that two changes occur; the institution of a Creative Commons license, specifically a CC-BY-SA license, and that a contract be put in place with students and other researchers that have access to the data.

Finally, data should be prepared so that it meets the data standards of other ocean biodiversity and biogeographic research such as OBIS. This will allow for easier analysis by other researchers that access the data, and it provides the ability share the data in their Archives.

References

Canadian Science Publishing. (2016). What is my funding agency's requirement for open access? *NRC Research Press.* Retrieved from

Jacobi, J.L. (2012) RAID made easy. *PCWorld.com*. Retrieved from www.pcworld.com/article/194360/raid-made-easy.html

Natural Sciences and Engineering Research Council of Canada. (2014). *Policy Statement: Peer-reviewed Journal Publications.* Retrieved from www.nserc-crsng.gc.ca/ NSERC-CRSNG/policies-politiques/Tri-OA-Policy-Politique-LA-Trois_eng.aspc

Ocean Biogeographic Information System (n.d.). *About OBIS, Vision & Mission*. Retrieved from www.iobis.org/about/

Ocean Biogeographic Information System (n.d.). *Ecological Metadata Language.* Retrieved fromwww.iobis.org/manual/eml/

Ocean Biogeographic Information System (n.d.). *How to Contribute data to OBIS?* Retrieved from www.iobis.org/manual/#contribute

Sync.com. Inc. (2018). *Why Sync?* Retrieved from www.sync.com/your-privacy/

Sync.com. Inc. (2018). *Pricing.* Retrieved from www.sync.com/pricing/

Treasury Board Secretariat. (2016). *Government of Canada Cloud Adoption Strategy*. Retrieved from www.canada.ca/en/treasury-board-secretariat/services/information-technology/ cloud-computing/government-canada-cloud-adoption-strategy.html

Data Management Plan

Dr. Green – Interdisciplinary Primary Care Teams Research

Executive Summary

Outlined in this report is a data management plan, which includes recommendations for database management, data storage and data security.

Best practice calls for *non-personal and unprotected data* to be made available as *open data* if possible and follow the FAIR (findable, accessible, interoperable and reusable) guiding principles for data management. This is more than just a best practice; it is a requirement for publicly-funded research, meets the university's policy on research data, is often required by peer reviewed journals, and shows accountability for public funds. Notwithstanding, provisions for the protection of privacy of a researcher's participants are paramount as an ethical issue in research.

We have prepared this report based on requirements outlined by the university librarian and input from the research project manager, Dr. Green, to be thorough and of benefit to him and his graduate student researchers.

1. Project Description

Dr. Green is focusing his research on evaluating how teams perform in high-pressure situations. He has selected hospital workers in primary care units as a case study for his analysis. Essentially, the professor wants to know whether the high-pressure environment in which these individuals operate affect how the team functions as a cohesive unit.

The primary data collection methodology includes interviews and written analysis. While Dr. Green has only two streams of data collection, the manner in which his information is stored varies greatly, and the file type is rather diverse, including Word, PDF, Excel and MP3 files. In terms of volume the real burden will come from the text files, such as Word, PDF and plain text, as these documents number in the hundreds. The audio MP3 files of interviews are relatively few in comparison.

Dr. Green has sole ownership of the data collected through his research. Although graduate students sometimes help him, he is wary of disclosing data to others. For this project Dr. Green is employing two graduate students who will need to have access to some of the data.

The primary objective of the data management plan should focus on Dr. Green's requirements for organization and security. Organization is needed for the classification and sorting of his text files. Additionally security is paramount to his data needs as there are confidentiality requirements for his audio files. Lastly, it is important to note that Dr. Green does not want his data to be open source.

2. Documentation, Organization and Storage

As mentioned previously, Dr. Green's data is spread across a variety of file formats. While the text related documents outnumber the audio files, it can be assumed that the latter will be a higher burden in terms of file size. An MP3 file containing an hour of content, depending on bitrate quality, can range anywhere from four megabits to 144 megabits. In comparison, an average PDF/Word file is under one megabit.

According to the follow-up questions with the professor, currently about 24 gigabytes of data is storage is on his USB drive. He also predicts that his future needs will triple in size over the next 10 years. As such, the professor will be looking for a storage platform that is scalable in capacity to parallel with the needs of his research.

Dr. Green is currently using three different types of storage platforms for each of the different types of files he deals with. Text and Excel documents are presently stored using Zotero, audio files with DropSync.com and Google Docs for transcripts. Best practices of data management go against what the professor is currently doing. Instead Dr. Green should consider consolidating his files.

The first task should be converting any Word documents to PDF files. Not only would this reduce the number of file types but it would also prevent other people from editing the document. Secondly the professor could benefit from the use of file containers such as ZIP files. These containers can store similar-in-content files, thereby drastically reducing the number of files the professor has to sort through. An example of this would be grouping together a MP3 file, its associated transcript and any other supporting documents pertaining to the interview into a single ZIP file. Lastly ZIP files would also reduce the file sizes of the professor's MP3 files, reducing his total storage needs.

The use of ZIP files will also have a secondary benefit for Dr. Green. By using this simple categorization method, the professor will be making use of metadata. Given that the ZIP files will contain directly linked content, each file makes use of metadata in the sense that the files will be categorized by subject. The use of ZIP files, and by extension metadata, will ensure that Dr. Green can easily identify and reuse the data at a later date.

Theoretically, using file containers will reduce the volume of files, however Dr. Green will still need to house these files somewhere that is secure but also accessible. We recommend that Dr. Green stores his file containers on a single cloud platform, rather than the three that he is currently using.

We recommend the cloud storage service from Sync.com. The company's Personal Pro Plan offers five hundred gigabytes of storage for an affordable $49 per year (Sync.com, 2018). Occasionally the professor has a need to share his data with his research team. Sync.com would allow him to do this via shareable URLs to the professor's platform. Sync.com also enables version control, something that will be important to Dr. Green as his team starts to sift through and work with the data. Security is paramount to Dr. Green as he is the only one who can view sensitive interview data. Sync.com would allow him to implement access control measures. This would ensure that his research team has access to the data but not to the files containing sensitive information which may identify research participants.

Another feature the professor is looking for from his storage solution is accessibility. Presently he keeps a USB drive with him at all times as a fail-safe option. The use of cloud storage would ensure that the professor has access to his data anywhere he goes that has a secure Wi-Fi connection. Sync.com also allows users to access their files via mobile devices as well as a mobile application. If the professor uses a smartphone he can have access to his files at any time regardless of a Wi-Fi connection. This would remove the need for carrying around a USB drive.

3. Access, Sharing and Re-use

The primary user, contributor and manipulator of the data will be Dr. Green. At times the professor will need his research team to access the project's files. However, there are certain audio files which will be restricted for the professor's exclusive access. Other than a couple of graduate students, Dr. Green does not wish to share his data with anyone else at this given time. This means the project's files will have to be protected with the appropriate Creative Commons licensing agreement.

In order to best protect his intellectual property the professor should pursue a CC-BY-NC-ND license. This type of agreement dictates that users of Dr. Green's data cannot use, modify or commercialize his data without his permission. Specifically this type of Creative Commons license is focused on maximizing the integrity and privacy of the creator's content. Dr. Green will need this type of license since a primary concern is the proprietary of his data.

In terms of sharing, using Sync.com will allow the professor to act as a gatekeeper for his data. Specifically he will have sole authority to share his project data with whomever he chooses. Two-factor authentication and device lockout functionality provided by the platform will further increase his sense of security of access.

It is important to note that while Dr. Green wishes for his data to be proprietary he may run into a policy conflict with the university. The university employs and funds the professor so the issue of who actually owns Dr. Green's data may arise in the future. There is a very real possibility that the university may lay claim to the professor's project, depending on how his employment contract was constructed.

4. Archiving

Since the professor's project will span a number of years, he will need an archive system to easily find and refer to his previous work. Dr. Green should employ a few best practices of archiving. First, he should ensure that he is regularly backing up his files that he has stored on his laptop/desktop. As new data comes in it can become an afterthought to store it in the cloud rather than just on his computer. Essentially Dr. Green will have to get into the habit of saving files on the single cloud server rather than his previous practices.

Another recommendation is to have a member of his research team digitize his spreadsheets from his work prior to 2002. It would be worth archiving this information as the professor could then easily refer back to his earlier work. He could store these digitized spreadsheets with Sync.com as his cloud drive will have ample space to support them.

The final aspect of archiving concerns duration of the data's usefulness or specifically its lifecycle. There are six stages in the data lifecycle; creating, processing, analyzing, preserving, accessing and re-using data. The stages most relevant to Dr. Green's work are preserving, access and re-using the data. Preservation of data pertains to the proper backup and storage of files. Following best practices of constant cloud-based backups would ensure this step is adhered to. Secondly, access to data and the concerns around who will have access is important to the professor's work. Due to confidentiality issues, the professor needs to restrict access to his files. Lastly, re-using data will be important for the purposes of study recreation. If the professor ever decides to release his study, he will have to make sure the associated data is within arms' reach for the purposes of re-use. This will be important as colleagues look to validate Dr. Green's study by recreating his project.

5. Conclusion

Dr. Green's research will need a unique solution as he has many different streams of data; open-source healthcare stats, interviews and quantitative excel sheets all have different information pertaining to his research. The professor should look to consolidate files by categorizing them within container files such as ZIP files. This would reduce the volume of files, organize them by content and reduce his overall storage needs.

Once his files are consolidated he should then use a cloud-based storage platform for housing them. Sync.com is a cloud-based storage company which offers a competitively priced personal plan that includes five hundred gigabytes of data. Most importantly, this platform offers version and access control. Both of these are paramount to the professor given the level of confidentiality of the data. As the professor's research begins to grow there will be a greater need to adhere to archiving best practices. The professor is thus encouraged to follow through with the recommendations given in this report throughout the duration of the project.

References

Sync.com. Inc. (2018). *Pricing.* Retrieved from www.sync.com/pricing/

Data Management Plan

Professor Pinkerton – Management Information Systems Research

Executive Summary

Outlined in this report is a data management plan, which includes recommendations for database management, data storage and data sharing. It is understood that the data created and collected is both significant in amount and varied in sources.

Best practice calls for data to be made available as *open data*, and follow the FAIR (findable, accessible, interoperable and reusable) guiding principles for data management. This is more than just a best practice; it is a requirement for publicly-funded research, meets the university's policy on research data, is often required by peer reviewed journals, and shows accountability for public funds.

We have prepared this report based on requirements outlined by the university librarian, and input from the research project manager, Professor Pinkerton, to be thorough and of benefit to the research team.

1. Project Description

While most of the research in question is related to management information systems, we understand the professor has many interests and an insatiable appetite for data. Currently, Professor Pinkerton has a significant number of Excel spreadsheets (17,384 in total) saved to a laptop. The median size of the spreadsheets is 1,000 rows of data by 10 columns. This amount of data requires approximately 60 GB of storage. The professor also gathers data from literature reviews, job descriptions, student performance evaluations, several files from other professors, as well as data sets from open data portals.

As a resource to her peers, Professor Pinkerton is often asked to share data files, but the number of requests for data exceeds the professor's available time to reply. The willingness to share the collection of data will be aided by cloud-based storage with a clear folder structure currently being designed by postdoctoral student Neil Gaiman.

2. Documentation, Organization, and Storage

Professor Pinkerton has roughly 60 GB of data worth of Excel files (XML), and very little data in other formats. While the professor is constantly adding to her data set, all of the data eventually ends up in Excel (XML), usually in numerical values, which does not require much storage. We recommend the university's cloud service which should provide enough storage to cover the storage needs for a long time to come if data continues to be added at the current rate. Once the data storage limit is exceeded, additional storage can be acquired at a relatively inexpensive cost.

The university's cloud also allows data sharing with anyone the researchers choose to share it with through an emailed link. Considering Prof. Pinkerton wishes only to share her data with certain parties, university policies will not create any barriers to this sharing of data.

It is understood that the plan is to continue to acquire data and add it to the collection at an increasing rate. While it may not be necessary at the current time, it is important that we give an alternative to the university's cloud in the event that additional storage is required. A cloud storage solution will be required so that citizen scientists and global research teams are able to collaborate and share data easily.

It is proposed that the cloud storage service Sync.com be used by the team. This service provides users with end to end encryption, and ample storage for users all at an affordable price. Sync.com also allows you to share any kind of file with anyone you would like quickly and easily (Sync.com, 2018). Each user can store up to one terabyte of data for $5 per month, and members are able to set up a free account with limited functionality, which would be ideal for other researchers looking to download the data (Sync.com, 2018).

While the size of the collection is not currently large enough to warrant purchasing cloud services, the documentation and organization of Professor Pinkerton's collection is very important. In order to maximize the societal benefit of the collected data, it is important that it be organized in a way that the professor and her fellow academics can quickly and easily locate. For this reason, we recommend all the data sets be tagged with metadata – essentially, data about data. There are three types of metadata; descriptive, structural and administrative, and we suggest employing all three in order to organize the data efficiently.

Descriptive metadata helps discover and identify the data; this includes the title, abstract, author, and keywords that can help readers locate the correct file. Structural metadata describes the type, version, and other important information. Finally, administrative metadata provides information such as how, why, when, and where the data was generated as well as the file type and who can access the file.

3. Access, Sharing and Re-use

It is important to have open access to data (in this case your collection). Ideally, data would be open to everyone to read and be shared. This is because academic research can have a great impact on the world as the spread of knowledge can improve processes, and help gain a better understanding of why certain events occur. Communication of the results that researchers identify is important so that other academics can verify and build on the outcomes. As research is often funded as a public good, the sharing of that data will maximize the value of the investment. By saving costs and sharing expertise, researchers across Canada can collectively benefit.

Please see below for a chart of our recommendation for the applicable licenses for the different types of data in Professor Pinkerton's files. The licenses that apply to the data all conform to the definition of open data. These licenses are widely used and considered a best practice. We have also provided brief descriptions of the relevant licenses for your information, with a chart contain the applicable license we recommend:

- **Creative Commons (CC0)**: This work is dedicated to the public domain all rights to the work are waived worldwide under copyright law. You can modify, distribute, and perform the work without asking permission.
- **Creative Commons Attribution License (CC-BY 4.0):** With this license, you are free to share and adapt the work as long as you give appropriate credit, provide a link to the license, and indicate if changes were made, however the you cannot infer that the licensor endorses you or your use.
- **Open Data Commons Attribution License (CC BY-SA 4.0):** With this license, you are free to share, and adapt works for any purpose from this work as long as you provide a link to the license, give appropriate credit, and indicate if changes were made, however you cannot infer that the licensor endorses you or your use. Also, if you adapt the work, you must distribute the contributions under the same license as the original. (Opendefinition.org. n.d.)

Type of data	Literature Review	Job Description	Student Performance	Open Portal	From Other Researchers	Personal Research
License	CC-BY-SA	Proprietary	Proprietary	CC0	Retain License	CC-BY-SA

(opendefinition.org. n.d.)

Our recommendation for access, sharing and reuse of this collection of data is to have two separate folders in the cloud, one of which contains Professor Pinkerton's proprietary information and is restricted to her use only. The reason for this is that the student performance data is restricted from being distributed, however if Professor Pinkerton were to anonymize the data she could have it under the license CC-BY-SA. Also, the job description data is not very useful outside of Professor Pinkerton's context. The second folder would contain all of the open data, and can only be accessed following a request to Professor Pinkerton. This will eliminate the need for Professor Pinkerton to search for files for her peers, but will also restrict access to the files as she requested.

4. Archiving

Archiving data is a cheaper way to store data which usually comes with the downside of being slower to access. It is also cheaper to backup data offline as opposed to in the cloud. Professor Pinkerton has made it explicitly clear that she fully intends to use all of the spreadsheets that she has collected at one point or another. Also, she is a well-known resource within her department. Researchers in her field are often requesting to use her spreadsheets. Any of the files in her collection could be used at any moment in time. Ideally, we should avoid archiving files in order to keep them readily accessible to those researchers who are looking to access them. We should also ensure that if they are archived, they are done so online. It is also assumed that all data will retain value after the life of the project and therefore no files will be deleted. Given the fact that the amount of data in Professor Pinkerton's possession is small enough to be stored for free, it seems counterproductive to archive any data at this time. Considering that it took 10 years for Professor Pinkerton to accumulate 60GB worth of data that is constantly in use, and the cost of storing data in cloud services is continually becoming cheaper, it is unlikely that a system of archiving Pinkerton's data will ever be necessary.

5. Conclusion

The above report contains our Data Management Report to suit your research needs. We have outlined best practises and assessed the cheapest and most viable options for your circumstance. Our plan includes the continued use of your university's cloud services until the amount of data reaches the point where a paid service is required, however this will likely not be for a while. Once you've reached that point, we recommend using *Sync.com* at a cost of only $5 per month.

Reference

Opendefinition.org (n.d.). *Open Definition 2.1*. Retrieved from https://opendefinition.org/od/2.1/en/

Opendefinition.org (n.d.). *Open Licenses.* Retrieved from https://opendefinition.org/od/2.1/en/

Sync.com. Inc. (2018). *Why Sync?* Retrieved from www.sync.com/your-privacy/

Sync.com. Inc. (2018). *Pricing.* Retrieved from www.sync.com/pricing/