What data will be created or collected (type, size, format, etc.)

Week 11 - Data acquisition

Week 10 - Data collection

Week 8 - DML & DDL

**(DML** is abbreviation of Data Manipulation Language. It is used to retrieve, store, modify, delete, insert and update data in database. **DDL** is abbreviation of Data Definition Language. It is used to create and modify the structure of database objects in database.)

|  |  |  |
| --- | --- | --- |
| Professor Periwinkle | Professor Green | Professor Pinkerton |
| - I want to be able to search by organism, geographical feature, depth, environmental conditions. I would be interested to see if you have any other ideas.(Discussion)  - All of my data is relevant and has value!(Discussion)  - Current data is stored in various ways, locally, hard drives, DVDs, CDs, etc. And a few files are in a Dropbox account. (Discussion)  - CSV files in uncompressed and raw data (200GB) (Discussion)  - size is all of your TSV files in uncompressed and raw data (500 GB) (Discussion)  - wildlife sightings reported by citizen scientists are downloaded as tsv files (Discussion) | - Currently he is using two primary methods of collection, Textual and content analysis. (Case)  - Textual is 383 individual documents describing - Team, Team outcomes, Team Practices.  (Case)  - Data is obtained from healthcare organization.(Case)  - Currently it is in the form of PDF, Word, plain text, and other documents all in digital formats (Case)  - Currently it is stored in Zerotero database (Size unknown) (Case)  - Qualitative data is stored in excel spreadsheet. (Case)  -Interviews recorded in mp3, then transcribed in Word. (Case)  # Challenge to keep the anonymity of the interviewee | 95% are from external sources (other researchers, Various governments, private corporations, etc.)  All in Spreadsheet format (excel or CSV)  Management information systems work (what the data relates to) |

What licenses apply to the data

<http://pitt.libguides.com/copyright/licenses>

Week 11 Slides are on Licensing

|  |  |  |
| --- | --- | --- |
| Professor Periwinkle | Professor Green | Professor Pinkerton |
| -Never bothered to get specific licenses (Discussion)  -Their data is open data and I own the data I produce. (Discussion)  -There is not procedure or contract in place that the students (incoming or leaving) have to sign. Do you think that it is necessary to have that written down somewhere? (Discussion) |  | -95% external resource |

Data Collection•

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

•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?

•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

**Professor Periwinkle**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Data(type) | Format | Frequency | Size | Total Size | Structure | Naming | Version Control | For Data Reuse | Owner | Licence |
| Sensor data (ROMV & static buoys) | RAW | day | 300mb |  |  | Data Type + TimeStamp data created+format | GitHub,SVN,Source Control | For Analytics data needs to be in same format such as csv or json | Prof |  |
| Uncompressed formatted sensor data | netCDF | day | 500mb |  |  |  |  | NETCDF can be easily converted to csv with general purpose programming language like python |  |  |
| Citizen sightings Report | TSV |  |  | 500 GB |  |  |  | Having data in csv makes it easy to have all data visualized in same visualization tools and can also be used easily for ML purposes | Open Data |  |
| Field Notes | Darwin core Format |  | 2GB |  |  |  |  |  |  |  |
| Simulation data | csv |  | Gigabytes of data | 200 GB |  |  |  |  |  |  |

RECOMMENDATION

1. Analytic purpose streamline data (CSV or JSON)
2. If the data is confidential - Cloud
3. No issue of scalability
4. No issue for back up
5. If data sharing is going to be on country level scaler where people provide you data and you create a platform to share it then a special web application along with dedicated server will be required
6. Use cloud server like amazon  like amazon
7. Need to have a contract with student  ( Use for academic purpose? ) why?

**Questions**

1. **Raw data is in which format - ( IMPORTANT because this will indicate the ease of converting the data to any required format)**

**Professor Green**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Data(type) | Format | Frequency | Size | Total Size | Structure | Naming | Version Control | For Data Reuse | Owner | Licence |
| audio recordings | mp3 | 15 | 57.6mb | 1 GB |  |  |  |  |  |  |
| Document files | text,pdf,word |  |  | 24 GB |  |  |  |  |  |  |
| Open source datasets | Text,pdf,word |  |  | 1 GB |  |  |  |  |  |  |
| Transcribed documents on drive |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |

**Questions**

1. **Is storing at different places intentional for data security? If not, would it be feasible for you to have all data in one place.**

# **Professor Pinkerton**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Data(type) | Format | Total Size | Structure | Naming | Version Control | For Data Reuse | Owner | Licence |
| 12 years of student performance data | xlsx | 60GB | Done by the post doc/ and very well structured by professor | Dataset name + open/non open  +timestamp |  | A portal to access all data. Open data can be immediately downloaded and for accessing non open data a ticket is created on website with requester details and sent to professor for approval. If approved then the requestor information for that dataset is added in master excel sheet and the requestor is granted the data set |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |

**Points:**

**She wants the control over who the data is shared with**

**What does she mean by ad hoc, is there any analysis that she provides in access with the data?**

Discussion points/Questions

|  |  |  |
| --- | --- | --- |
| Professor Periwinkle | Professor Green | Professor Pinkerton |
| Recommendation?  For analytic can we propose streamlining the data in the format CSV or JSON  Are we uploading in a cloud?  Since there is no issue of confidentiality or scalability?  I guess if we upload data on cloud, back up is not required since the cloud service would automatically do that for us right?    If data sharing is going to be on country level scaler where people provide you data and you create a platform to share it then a special web application along with dedicated server will be required  Use cloud server like amazon like amazon  Need to have a contract with student (Use for academic purpose?) why? |  |  |