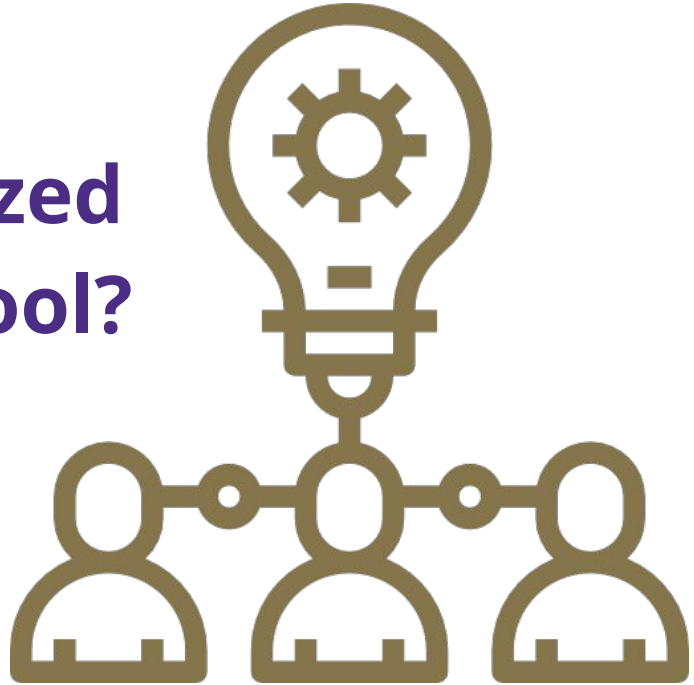
Is There a Need for Centralized Data Repository in the iSchool?



Meet the Team

GROUP 7

YANJIE NIU

Yanjie was primarily responsible for developing research questions, doing interviews, conducting overall qualitative analysis, and providing recommendations.

PRAKIRN KUMAR

Prakirn was primarily responsible for initiating the quantitative research by designing the survey, doing interviews, and did the quantitative analysis and data visualization.

ISHITA BHANDARI

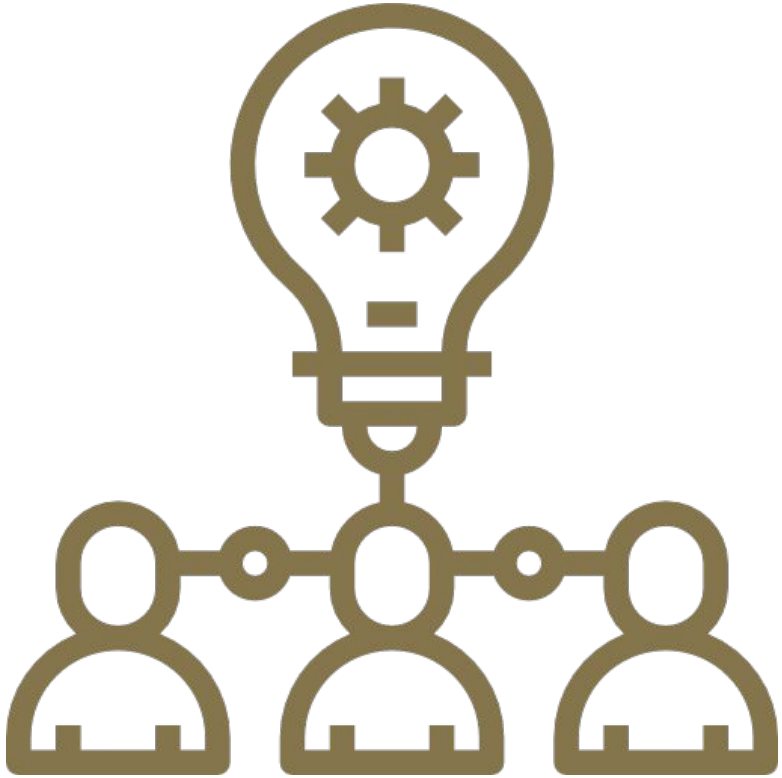
Ishita was primarily responsible for defining aims and objectives for this research. She also conducted interviews and helped in qualitative analysis.

SEJAL KHATRI

Sejal was primarily responsible for defining approach for data collection. She also conducted interviews and helped in qualitative analysis and data visualization.

Table of Contents

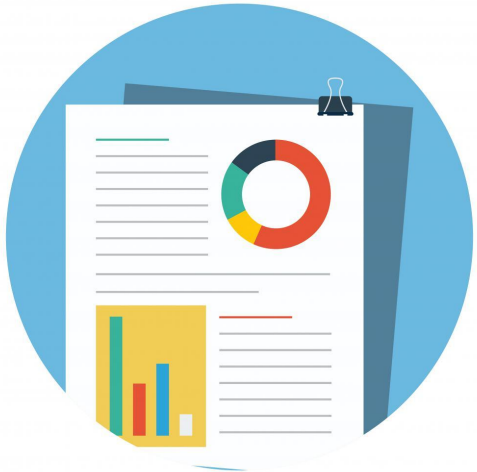
INTRODUCTION



Executive Summary

This research is aimed at identifying whether or not there is a need for a centralized data repository at the iSchool, where our community can publish and access research as well as course-related data. We did this by interviewing lecturers, PhD students, and professors and surveying students.

We managed to get a general perspective of students who rated the repository 4.1, based on its usefulness, on a scale of 1 to 5 where 5 being the highest. Hence, we found that the such a repository is beneficial for the students as well as needed. On the other hand, our interviewees believe that there is no need for such a repository.



We tried to understand the needs and concerns of our interviewees when it came to sharing their data on a centralized portal. Sensitive nature of the data and requirement of time and efforts were the major deterring factors. Whereas, security and persistent nature of the portal were some of the major requirements.

Among our top recommendations is a further research to explore the students need in more depth. iSchool can also have a data management portal plan based on the research methodologies (qualitative or quantitative) that are used. We also recommend iSchool to collaborate with other departments and take this repository to the university level.

Motivation and Significance

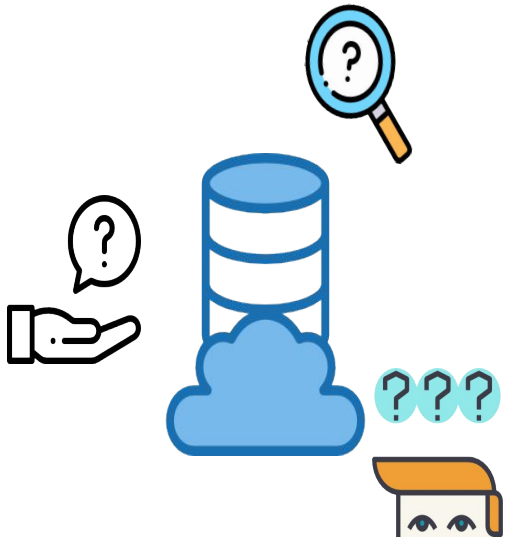
A centralized data repository is a common platform for instructors, researchers and students to collaborate and access the academic data like, but not limited to, research materials, dissertations or coarse materials. With universities like Harvard and Georgia Tech joining the open data movement and creating their “dataverses - open source web applications for sharing, preserving, citing, exploring, and analyzing research data”, one questions as to why the Information School at the University of Washington is hesitating from joining the wagon. Currently, the iSchool doesn’t provide a unified data portal for its community to store and share their data on a single platform. As a researcher, if one wants to know about the research being done in the university and wants to access it, then one has to jump through a few hoops and try to find relevant materials posted on different portals. This is exactly what sparked a series of questions in our mind - What are the expectations and major concerns associated with educators when it comes to sharing their data on these platforms and feasibility of such a portal. The scope of this research is limited to the Information School only due to time and various other constraints.

Aim and Objectives

Is There a Need for Centralized Data Repository in the iSchool?

AIM

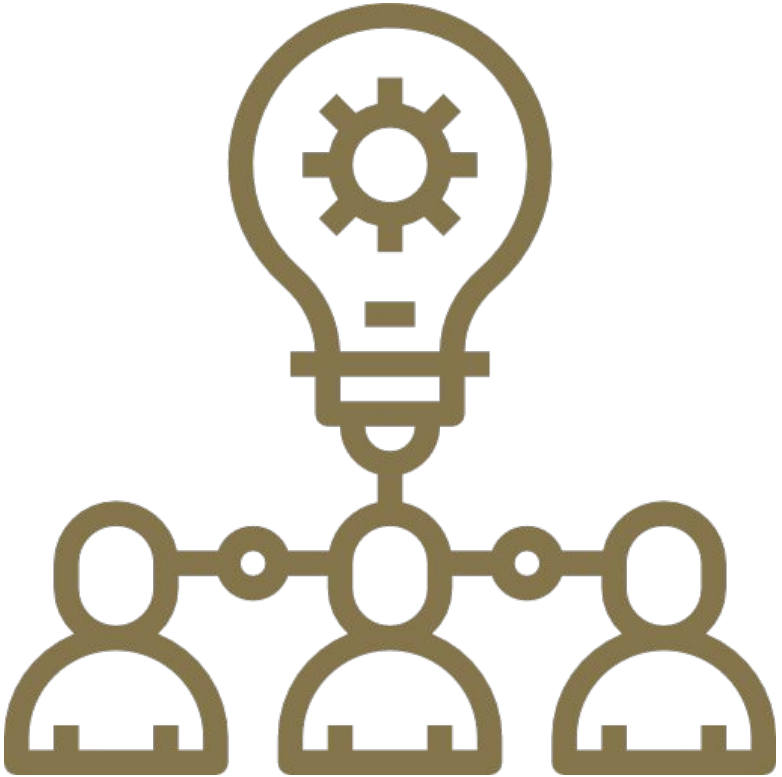
We aim to analyze the need of a centralized repository which serves as a unifying platform for publishing and sharing data related to the research and other academic activities going on in the Information School (iSchool). The emphasis is on unearthing the underlying reasoning behind people's response to the need of a centralized repository.



OBJECTIVES

- To explore the reasons behind people's responses.
- To obtain exact needs and expectations of the people regarding the topic.
- To check the feasibility of creating a repository to meet those needs.
- To provide meaningful recommendations to the iSchool.

METHODOLOGY

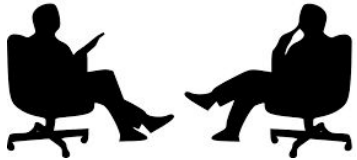


Approach for Data Collection

We are following the **mixed approach** for data collection, taking advantage of both Qualitative and Quantitative research methods, as this will help us in obtaining data that can be measured and analyzed and data that cannot be measured.

Secondary Literature Research

To understand the research that is already being done in the field of open data and data sharing, we studied some of the secondary research literature, which helped us form a base structure for our research.



Some of the sources that we used for secondary research are:

1. “Open Research Data Repositories: Challenges and Opportunities for Libraries” by Abdul Azeez.
2. “The State of Open Data Limits of Current Open Data Platforms” by Katrin Braunschweig, et al.
3. Researcher-Library Collaborations: Data Repositories as a Service for Researchers Andrew S. Gordon, et al.

Primary Research Methods

Our primary data collection method are interviews and surveys. This will help us understand the current scenario of the open data availability and usage at the iSchool and the opinion and needs of several professors and students who are involved with such data on daily basis.

Sampling Methods

Snowball Sampling: Snowball Sampling method is the method in which participants refer the researcher to others who may be able to potentially contribute or participate in the study. Since we are unaware of the research interests of the professors, we can use this method to help professors introduce us to researchers who are working in the same area.

Purposeful Sampling: It is the method in which participants are selected or sought after based on pre-selected criteria based on the research question.

Since, our research problem consists of the term 'Open Data', we feel not everyone will be able to understand it properly. Therefore, we have decided to use purposeful sampling to target PhD students and professors working in the field of data.

Quota Sampling: It is the method in which we would gather data from a certain number of participants that meet certain characteristics.

Data Collection

Qualitative Data Collection:

Our selected participants for Interviews are iSchool professors who we see as the potential contributors and users of open data repository along with some professors who might have different opinions to get a holistic view. We have also interviewed PhD students who are doing research in the field of open data and taking initiatives for open data movement. PhD students helped us understand the complexities of our research and were of great resource for directing us to professors with similar interests. We have interviewed 4 professors and 5 PhD students for our research.

Quantitative Data Collection:

The interviews conducted by us, of various professors and students helped us in collecting a vast array of qualitative data, but in order to check the credibility of the data and insights that we obtain from qualitative data, we need to resort to quantitative data. For this we decided to send out surveys to Master's students in order to understand the need of the students.

Research Considerations

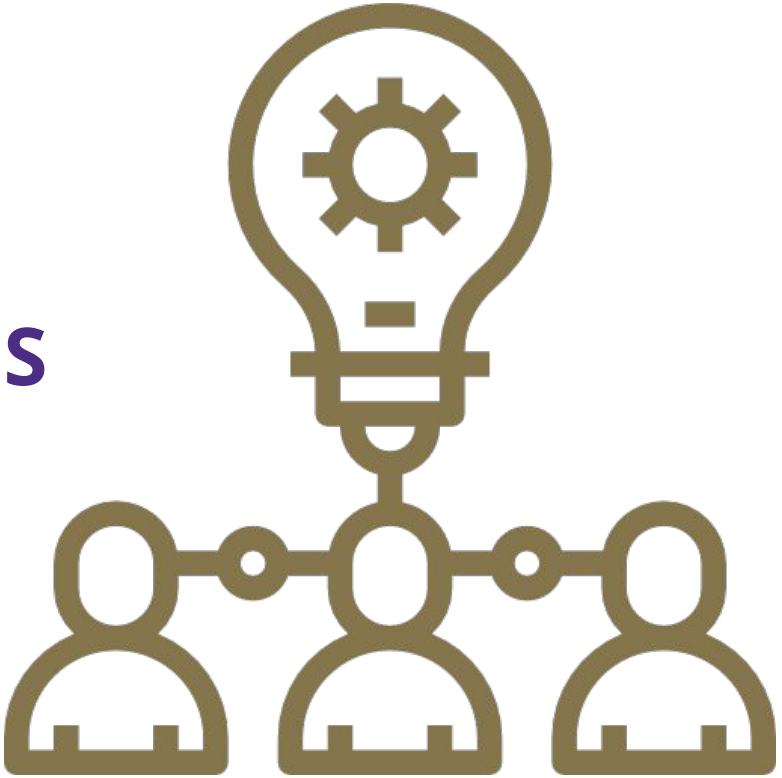
Validity and Reliability

- For qualitative analysis we used the Data Triangulation approach for reliability testing.
- Thus, following these testing approaches we maintained the validity and reliability of our research
- We evaluated the credibility of qualitative data by using the quantitative data.

Ethical Considerations

- Maintaining anonymity of our interviewee's responses
- Respecting privacy of our respondents
- Refraining from introducing our bais into the analysis
- Being mindful of our research consequences

FINDINGS AND ANALYSIS



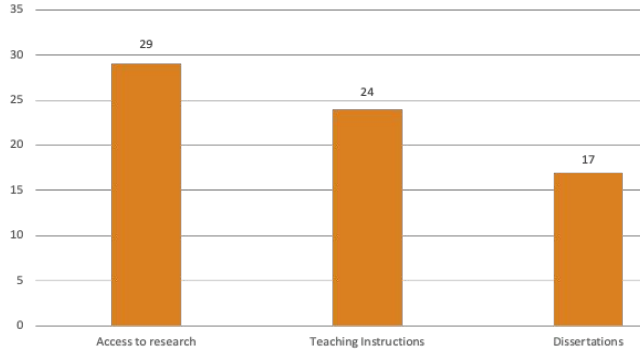
Quantitative Analysis

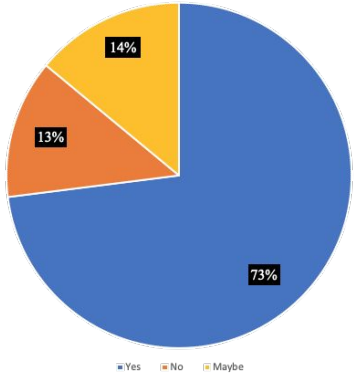
Method of analysis: For quantitative analysis, we chose the descriptive design approach of analysis, which seeks to describe the current status of a variable or phenomenon. We chose this method because we wanted to seek participants' answers or opinions on an idea which correlation to any other variable.

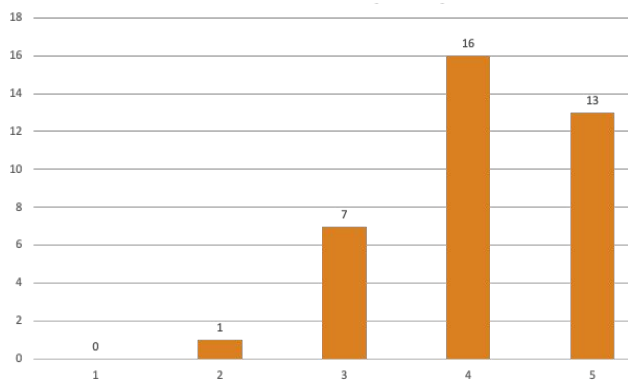
Data Screening: In order to know the opinions of students about the idea of having an open data repository, the goals of having such repository and its usefulness we made a short survey. Moreover, we chose to send surveys to the students only because we felt that it is better to interview the professors in order to deeply understand their opinion on our research question.

The survey questions were designed by carefully keeping in mind to avoid any biases or leading questions. We got 37 responses for the survey through our fellow MSIM students and PhD students in the iSchool. The response received from the survey very insightful and gave us the direction for our further research. For reliability testing, we used **Homogeneity approach**.

Quantitative Analysis







Question 1: GOALS OF THE REPOSITORY

The responses revealed that nearly 78% of the students would like an access to the past as well as ongoing researches in the iSchool and about 65% of the participants want the feature of teaching instructions (class slides, data sets, etc.) of all the courses offered by the iSchool in the Open Data repository.

Question 2: WILLINGNESS TO SHARE DATA

To our surprise, 3/4th group of participants were ready to share their research paper or projects on an open data repository of the iSchool.

Question 3: RATING OF USEFULNESS OF SUCH REPOSITORY

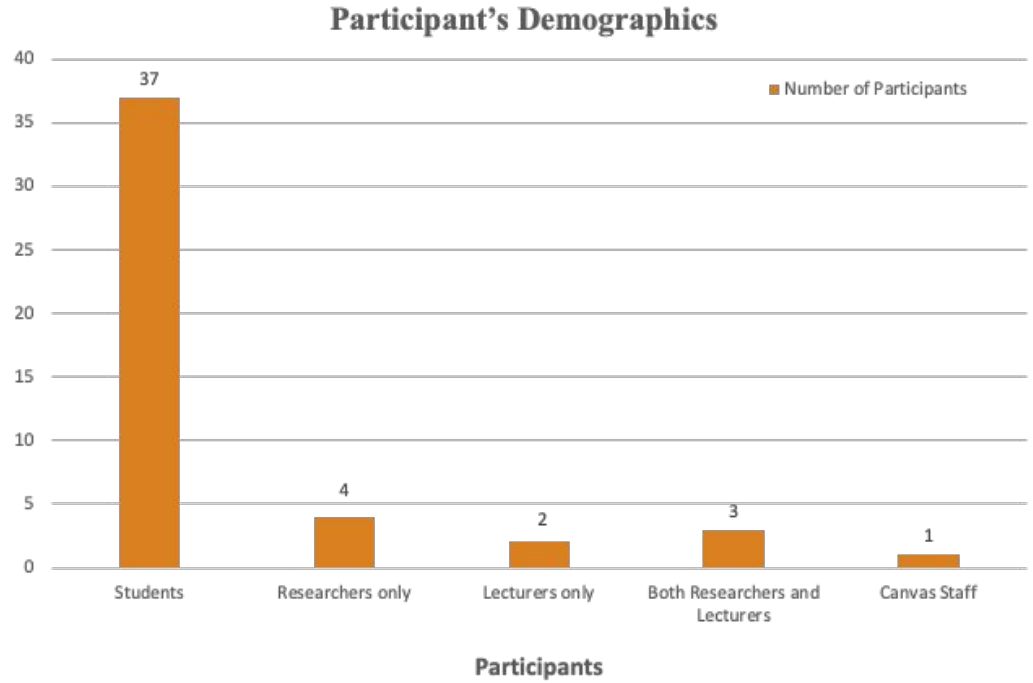
The survey participants rated the usefulness of having such open data repository for academic purposes at the iSchool as 4.1 on a scale of 5 (5 being most useful), on an average.

Qualitative Analysis

Analysis Method

We used **Grounded theory method** for analysing our qualitative data. Grounded theory is a systematic methodology in the social sciences involving the construction of theories through methodical gathering and analysis of data. It is a research methodology which operates inductively. generating new theory emerging from the research data.

Since we were unaware of the research interests of the professors and current sources provided by the iSchool, we used this theory to build on the data we collected and set a direction for our research.



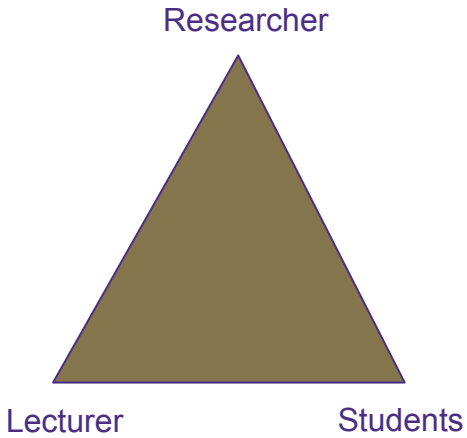
Qualitative Analysis

Reliability testing

For qualitative analysis we used the Data Triangulation approach. It is a method to check and establish validity in our study by analysing our research question from multiple perspectives and to arrive at consistency across data sources.

We used this method to understand requirements from all the people in the iSchool. Also, because the target audience for iSchool centralized repository is varied.

Data Triangulation Approach



Findings

Lecturers

We interviewed 3 lecturers who do not do research. When we asked if they would publish data on the centralized repository, 2 of them said yes. Most of them said they themselves do not need a centralised repository.

Some Findings:

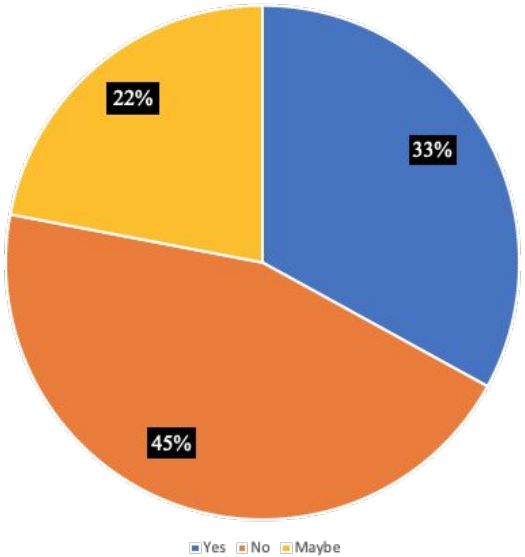
- Did not find the need for the repository.
- They were satisfied by the functionalities provided by **Canvas.**

Researchers

We interviewed 7 professors and PhD students who do research.

Some Findings:

- 4/7 participants are willing to share their data.
- No Persistent portal exists for putting extra datasets for students.
- Discipline based data portal exists but not institutional based like iSchool data portal
- Data generated from interviews is hard to anonymize and can lose context



Interviewees' willingness to share data on iSchool's Data Repository

Conclusion

Students feel having such Data Portal is:

HELPFUL and NEEDED

Lecturers & Researchers feel having such Data Portal is:

HELPFUL but NOT NEEDED

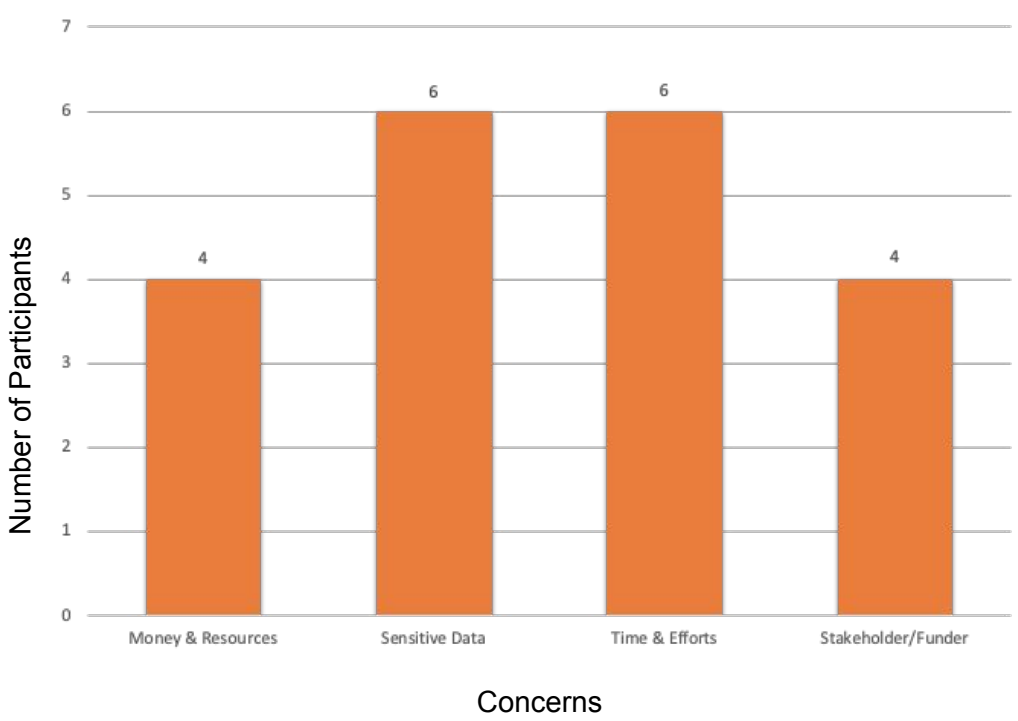
Concerns

Lecturers

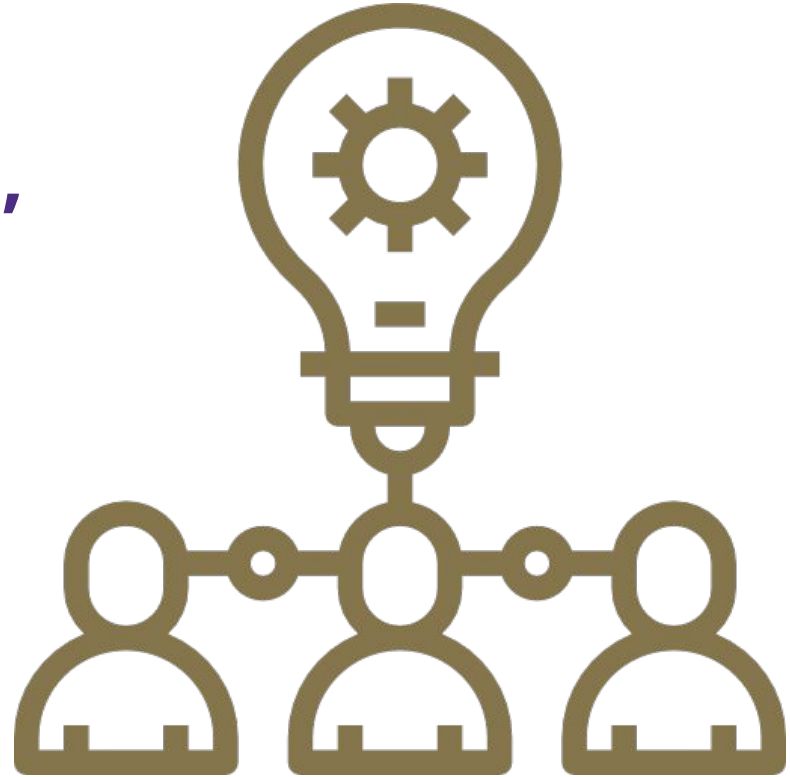
- It should not end up being a dumping ground.
- There is an important consideration of constant maintenance.

Researchers

- Researchers who study human subjects, social science, anthropology or sensitive studies, do not want to share, because of data sensitivity, intellectual property, copyright, extra time and effort, lack of motivation, and loss of context after anonymization.
- There are external factors like stakeholders which have authority over the data, so data scientists cannot publish the data



RECOMMENDATION, LIMITATIONS & FUTURE ACTIONS



Recommendations

Based on our qualitative and quantitative data analysis on needs, wishes, and concerns about building an open centralized repository in iSchool for research and instruction data sharing, we come up with actionable recommendations accordingly.

1. Identify Students' Needs (3 months).
ISchool could organize more interviews and focus groups to understand why students need a data portal, and what they need from such a portal. Then, iSchool could come up with how to meet students' various needs of the data portal accordingly. For example, iSchool could contact faculty members and PhD students in the areas of interest, based on students' responses, to ask which data they are willing to share with students and which existing data portals they know would be helpful for students.

Reasons:

- In our surveys, iSchool students indicate that they have the need to practice data processing and analyzing for learning, job-searching, and research. They hope iSchool could provide them with a data repository to accumulate hands-on data use experience.
- Six teaching faculties think that they put course-related instruction materials in Canvas, but they also have some data at hand that they do not use anymore. They are willing to share it with students to let them learn and practice data processing.

Impacts:
ISchool would connect students, and faculty members and PhD researchers, to efficiently develop actionable plans regarding the data portal to benefit both sides.

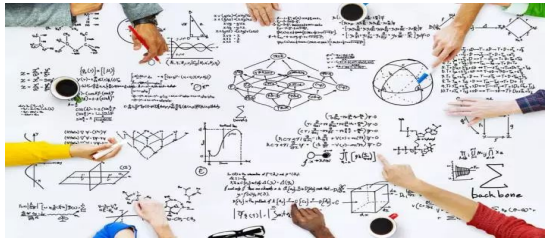
Costs and Challenges:
One challenge is to get students involved during their busy schedules. To hold events to let students know their involvement would provide them with more possibilities to do projects and make their resumes stand out would be a good solution.

Recommendations

2. Adjust Data Management Portal Plan, Based on Methodologies.

The need of data repository also depends on the ideas of data providers, which differ by research methodologies. To address this issue, thus, we have to consider these differences and develop recommendations accordingly.

For quantitative study areas, since a lot of researchers already have labs and data sharing is normalized, they can try to recruit more students to join the labs, clean and then use data. Especially, iSchool could have a newsletter to share opportunities to do data research at iSchool and even UW in general (3 months to prepare and coordinate).



For qualitative study areas (such as history and anthropology), since more identifiable information and copyrights issues are involved, and context might not be succinctly summarized in an annotation (Snyder, 2014) or context might be lost after anonymization, it is more like voluntary work for researchers to spend more time and efforts to clean and share data, rather than required. Professor Megan Finn in iSchool faced many challenges to manage data in her project and ended up depositing Zotero citation records and notes (Finn, 2017).

After doing research, we find there is the Qualitative Data Repository (QDR) by Zotero to store and share digital data (and accompanying documentation). So iSchol could encourage qualitative researchers to share some of their data in this data portal and then encourage students to use data in this portal. It needs pilot tests. (2 years to prepare and do pilot tests).

QDR
QUALITATIVE DATA REPOSITORY

Recommendations

2. Adjust Data Management Portal Plan, Based on Methodologies.

Reasons:

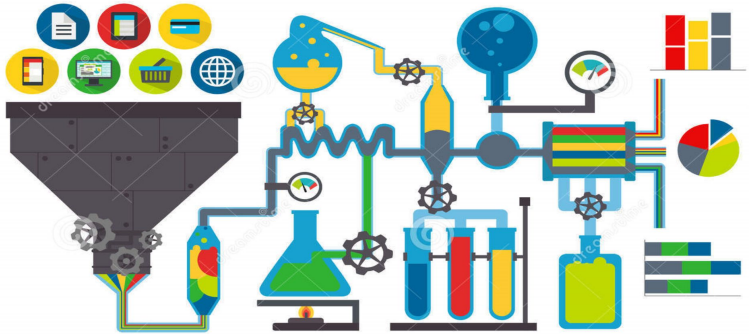
- We chatted with iSchool MSIM students and they were very willing to even do volunteer work in labs to learn, but they knew few chances.
- Most of our interviewees mentioned the differences in data sharing, between qualitative and quantitative research.

Impacts:

- Speed up data sharing in quantitative study through students' engagement.
- Tag potential values of qualitative data through providing access to it.
- Promote a healthy and active ecosystem in research communities regarding data sharing and use.

Costs and Challenges:

One challenge might be the willingness of faculties and PhD researchers to share their qualitative data in the Qualitative Data Repository, since it is a new exploration. Also, they might feel lack of motivation to do extra amount of work in their already busy schedules. ISchool could do some pilot tests to help decide whether to adopt it more.



Recommendations

3. Promote the Implementation of Data Portal at the University Level through Departmental Collaboration (3 years).

From the aspect of platform, the issue of data portal is always linked to organizational structure. The trend would be to have a united data portal to provide data access to more students and researchers at UW. Specifically, staff in charge of IT at both the iSchool and the university level should make long-term plan and collaborate to build a data portal.

Reasons:

- Most of our interviewees discussed that their vision about data portal would be a unified data repository on a larger scale.
- There are less quantitative researchers in iSchool than qualitative ones. Thus, data available to be shared in iSchool might not be enough to meet the needs of students.
- Data sharing among different departments at UW would be easier approved through IRB, compared with cross-institutional data sharing.

Impacts:

- Cost-effective solution for iSchool.
- Data sharing on a larger scale to benefit more people.
- This branding plan would make UW have a higher reputation, because of students' various hands-on experiences of transforming data to information in such a portal.

Costs and Challenges:

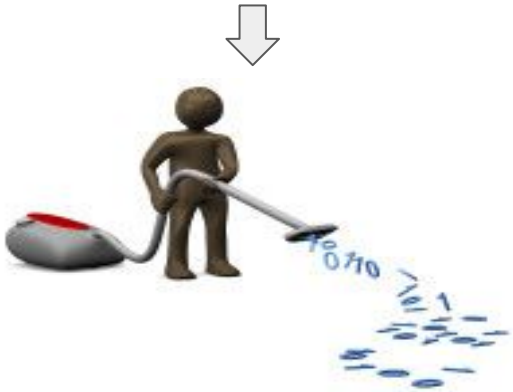
UW has a decentralized structure. It would take long time to communicate and make actionable plans, let alone implementation.

Limitations and Future Actions

Due to time limit, we have not done enough interviews about students. We would collect more qualitative data from diverse student groups (informations, MLIS, and MSIM) with different specializations interest and understand their specific needs of the data portal in order to make more plans accordingly.

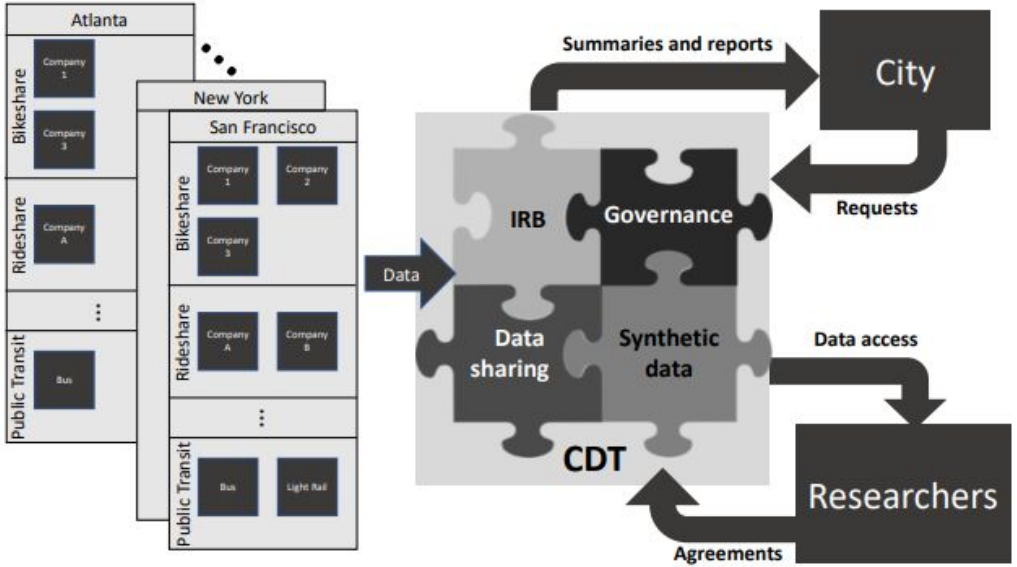
Also, since a lot of our interviewees hoped UW could be the one to have such a unified data portal, we should have interviewed more IT staff at the university level to study whether the need of a unified data portal by iSchool students could be met through a centralized data portal at UW. If we have more time, it would be an important aspect to further explore.

Further studies on open data repository need to address the following issues. Technically, right now, there are no systematic methods of cleaning data and assuring analytic properties and privacy (Dwork & Yekhanin, 2008; Narayanan & Shmatikov, 2008). For data-providers, they still need to mainly rely on manually clean data, which means their concern about lack of time, energy, and motivation still exist. Thus, to address this technical bottleneck, people interested in CS could do more research on the relation between microdata and metadata, the process of re-identification, and the possible solutions to stop working backwards toward reproduction. To develop accountable algorithms and softwares and help do data cleaning efficiently would be a challenging but also impactful way to explore.



Limitations and Future Actions

Moreover, apart from technical solutions, an inspiring pathway to develop could be an integrated legal-technical approach to deal with sensitive data (Young et al., 2018). A third-party data trust could be introduced to balance competing interests of the stakeholders in the process of data-sharing. This would provide low-risk data access and sharing. However, one challenge might be the selection of a third-party data trust. What factors to consider when selecting a third-party data trust, what qualifications a third-party data trust should have, how to formulate new laws to regulate a third-party data trust and help both data providers and receivers reach an agreement would be the issues to explore in practice. A local project “UW Transportation Data Collaborative” is an inspiring data governance initiative that we can learn from.



Example of the Collaborative Data Trust (Young et al., 2018)

References

1. Castro, D. (2015). *How Open is University Data.* Retrieved from: http://www.govtech.com/data/How-Open-Is-University-Data.html
2. Dwork, C., & Yekhanin, S. (2008). New Efficient Attacks on Statistical Disclosure Control Mechanisms, *Advances in Cryptology—CRYPTO.* Retrieved from: *2008*, http://research.microsoft.com/research/sv/DatabasePrivacy/dy08.pdf.
3. Finn, M. (2014). Data Management: A Mandated Opportunity Collaborative Research. In the author's possession.
4. Narayanan, A., & Shmatikov, V. (2008). Robust De-anonymization of Large Sparse Datasets, *Proceedings of the 2008 IEEE Symposium on Security and Privacy*, 111-125.
5. Snyder, J. (2014). Active Citation: In Search of Smoking Guns or Meaningful Context? *Security Studies, 23* (4), 708–14. Retrieved from: https://doi.org/10.1080/09636412.2014.970409.
6. Young, M., Rodriguez, L., Keller, E., Sun, F., Sa, B., Whittington, J., & Howe, B. (2019). Beyond Open vs. Closed: Balancing Individual Privacy and Public Accountability in Data Sharing. In the author's' possession. Retrieved from: https://homes.cs.washington.edu/~billhowe/publications/pdfs/young_open_v_closed_semi_synthetic_data.pdf

Appendix

1. **Affinity Diagram:**

Mapping of the data to make an affinity diagram with four clusters under which lie several keywords with similar themes. These keywords were generated from the qualitative data that we collected in the span of last 1-2 weeks.

Appendix

2. Survey Link:
https://goo.gl/forms/CHSoOIZ6S2s3NvFp1

3. Interview Questions:

- **Publishing data questions-**
  - Think about the recent research you conducted or the course material you structured.
  - Did you publish that data? Where?
  - Would you publish the data you curated for research/ teaching purposes on the open data portal provided by the iSchool?
  - What type of data would you publish if such portal exists?
  - Are you interested and excited about the idea of iSchool having an open data repository?
  - Are you willing to share your course instructions and resources on an open data repository of the iSchool?
  - Are you willing to share your research paper or projects on an open data repository of the iSchool?
  - If Ischool decides to have an open data repository, what all would you need?
  - If compared to canvas -
    - Additional features ?

Appendix

- **Accessing Data Questions-**
  - Do you have the need to access the data related to iSchool?‘
  - How do you access data related to iSchool?

- **Significance of Open Data Repository-**
  - Going back to our research question, do you think there is a need to have an open data repository in iSchool? Who might benefit from it? Why?
  - If you are decision marker about this issue, what you would do to decide whether there is a need to have an open data repository? Why?
  - Also, if you are a decision marker, what you would recommend ischool to do to address/implement the open data repository issues? Do you have some suggestions? Why?
  - What you might concern and obstacles? Costs? Technology? Privacy? Administrative? How do you suggest to address them?
  - Last but not least, as a information science and data researcher, what is your vision about the issue of open data repository and data governance in the future, in the field of education and the society as a whole? Which other research projects no matter in academia or industry you think are setting a good direction toward your vision and you recommend us to read more?