Advances in Engineering Education

SPRING 2016

Perceptions and Practices of Data Sharing in Engineering Education

ADITYA JOHRI
George Mason University
Fairfax, VA

SEUNGWON YANG
Louisiana State University
Baton Rouge, LA

AND

MIHAELA VORVOREANU

AND

KRISHNA MADHAVAN
Purdue University
West Lafayette, IN

ABSTRACT

As part of our NSF funded collaborative project on Data Sharing within Engineering Education Community, we conducted an empirical study to better understand the current climate of data sharing and participants' future expectations of the field. We present findings of this mixed method study and discuss implications. Overall, we found strong support for sharing research data within the community but participants were cautious about ethical and privacy implications, as well as issues around ownership of data. Participants expressed the need for an easy to use system to support sharing research data.

INTRODUCTION

We conducted a small mixed-method study consisting of a survey and interviews to better understand the culture of data sharing within the engineering education community. Our goals were to get some basic idea of what engineering education community members think about data sharing, what are the existing practices, and what expectations do community members have for the future. Results from such a study has significant implications for the design and development of a data sharing infrastructure for engineering education that places the requirements of the larger community at the center of the infrastructure design process. Our findings confirm many commonly held notions in the community about data sharing. For example, there is recognition that there is very little data sharing in our community. There is also a consensus that it is hard to share qualitative data. Surprisingly, we found quite a bit of support for more data sharing.

In the interviews we conducted, members of the engineering education community expressed a lot of interest in sharing their data and using others' data as well. Results also indicate a great interest in helping the community move towards more data sharing. From our conversations with community members and from our own experiences, we were aware that data sharing is minimal within the community, especially at larger scales, even though some data is shared among collaborators. Our aim with the research study was to get better insights into the current state of data sharing in engineering education and what needs to be done if data sharing is to be supported.

DATA COLLECTION

Survey

We adopted a survey instrument from Tenopir et al. (2011)[1]. We included most of the items they used, but also adapted many for the engineering education community. It is important to note that this survey was originally designed for use in the natural sciences. We publicized the survey instrument primarily to one sub-community of American Society for Engineering Education (ASEE) – Educational Research and Methods (ERM) division – through their mailing list. We had a total of 32 responses of which 27 responses were usable. We discuss details of the survey demographics in the next section on survey findings.

Interviews

In their survey responses we asked respondents to provide us with an email address if they wish to be contacted for a follow-up interview. 15 respondents provided their email addresses and we conducted interviews with 6 participants. Since the survey responses were stored separately from the email addresses, we did not correlate the survey and interview responses.

1 Link to survey instrument: doi:10.1371/journal.pone.0021101.s001

	Frequency	Percent
Academic	27	96
Government	0	0
Commercial	0	0
Non-profit	1	4
Other	0	0
Total	28	100

Table 1. Primary work sector.

SURVEY FINDINGS

Demographic Information

The majority of our respondents were academics (*Table 1*). On average, respondents reported that almost 40% of their time was spent on research and around 30% on teaching (*Table 2*). Other prominent work activities included administration (11%) and service (6.97%) (6 respondents reported in open-ended response to the "Other" category). In terms of age, the sample was spread from 25 years of age to over 67 years (*Table 3*). The respondents were relatively evenly spread from around 30 years to 50 years of age. The sample had almost equal number of male and female respondents (13 & 14 respectively) (*Table 4*).

Research Affiliation

For 3/4th of the respondents, engineering education was the primary area of research (*Table 5*). 10% of the respondents considered it their secondary area of research. The rest worked primarily in other fields but had an interest in engineering education. We had a mix of academic positions

	Average	Standard Deviation
Administration	11.00	15.78
Outreach	3.70	8.69
Policy support	0.33	1.83
Research	38.67	25.43
Teaching	29.33	22.27
Other	6.97	18.97

Table 2. Allocation of work time.

	Frequency	Percent
18–24	0	0
25–31	4	15
32–38	5	19
39–45	6	22
46–52	6	22
53–59	1	4
60–66	3	11
67 or older	2	7
Total	27	100

Table 3. Age group.

within the sample. The largest subgroup was Associate Professors at 29%, followed by Assistant Professors at 21%. A small subset of graduate students, 14% also responded to the survey (*Table 6*).

Research Funding

75% of the respondents reported that they were funded through federal/national agencies, the rest received funding from the State or Private foundations (*Table 7*). 70% of the respondents also indicated that they were required to submit a data management plan as part of the process for securing their research funding.. The rest were equally divided between those who were not required and those who did not know (*Table 8*).

Ownership of Data and Data Description

Only 22% of the respondents indicated that they had sole responsibility for approving access to all their research data, while almost 37% reported that they had sole responsibility for some of their data (*Table 9*). 41% reported that they did not have sole responsibility for approving access to the data. In the open-ended response to an item that asked “what additional approvals would be

	Frequency	Percent
Male	13	48
Female	14	52
Total	27	100

Table 4. Gender.

	Frequency	Percent
I consider engineering education my primary area of research	21	75
I consider engineering education my secondary area of research	3	11
Engineering education is tertiary to my research program	1	4
I serve primarily as an evaluator on engineering education research projects	0	0
Other	3	11
Total	28	100

Table 5. The level of engagement with engineering education research (choose one).

necessary", respondents indicated that they would need permission from fellow PIs/collaborators and/or from IRB. One respondent also raised the concern that unless sharing was explicitly requested in the informed consent, permission could not be granted after data were collected. When asked about the use of metadata, a useful tool for data sharing, 30% of respondents reported that they used metadata to describe their data; 70% said they did not (*Table 10*).

METHODOLOGICAL PREFERENCES OF SURVEY RESPONDENTS

The preference for research methods in our sample was diverse. 46% of the respondents used either quantitative or qualitative data, depending on their needs. 25% identified themselves as

	Frequency	Percent
Administrator	1	4
Assistant Professor	6	21
Associate Professor	8	29
Professor	4	14
Graduate student	4	14
Undergraduate	0	0
Lecturer	0	0
Post-doctoral fellow	0	0
Researcher	1	4
Other	4	14
Total	28	100

Table 6. Current position.

	Frequency	Percent
Federal/national government	21	75
State/regional government	1	4
Corporation	0	0
Private foundation	1	4
Other	5	18
Total	28	100

Table 7. Primary funding agency.

	Frequency	Percent
Yes	19	70
No	4	15
Don't know	4	15
Total	27	100

Table 8. Requirement for a data management plan by funding agencies.

	Frequency	Percent
Yes - for all my datasets	6	22
Yes - for some of my datasets	10	37
No	11	41
Total	27	100

Table 9. Sole responsibility for approving data access.

	Frequency	Percent
Yes	8	30
No	19	70
Total	27	100

Table 10. Use of metadata.

	Frequency	Percent
I am a qualitative researcher	7	25
I am a quantitative researcher	1	4
I am a mixed-methods researcher (I use both quant and qual in the same study)	5	18
I am a multiple methods researcher (I use both quant and qual depending on what I need)	13	46
Other	2	7
Total	28	100

Table 11. Qualitative vs. quantitative researchers.

qualitative researchers, and 18% considered themselves mixed-method researchers. A small subset, 4% identified quantitative methods as their primary technique (*Table 11*).

Surveys were the most common data collection method used by the respondents (86%), followed by interviews (71%), and focus groups (64%). Respondents also used observations (50%), archival data (36%) and experiments to (32%) commonly (*Table 12*). In their open-ended responses respondents also mentioned using concept inventories and other assessment instruments in their research.

CURRENT PRACTICES RELATED TO DATA SHARING

We asked respondents if some or all of their data were available to others and if data were available, how they were shared (*Table 13*). 77 of the respondents reported that none of their data were available to others on either the project or PI website; 19% reported that some data were available

	Frequency	Percent
Surveys	24	86
Focus groups	18	64
Data for simulations and modeling	0	0
Experimental (involving some degree of manipulation)	9	32
Interviews	20	71
Observational (no manipulation involved)	14	50
Archival data	10	36
Other	5	18

* Note: This is ‘check all that apply’, thus the percent does not sum to 100.

Table 12. Type of data collection method (check all that apply).

	None	Some	Most	All	Total
On my organization's or project website	20 (77%)	5 (19%)	1 (4%)	0 (0%)	**26 (100%)**
On the principal investigator's website	21 (84%)	4 (16%)	0 (0%)	0 (0%)	**25 (100%)**
Through a national network	22 (85%)	4 (15%)	0 (0%)	0 (0%)	**26 (100%)**
Through a regional network	24 (96%)	1 (4%)	0 (0%)	0 (0%)	**25 (100%)**
Through a global network	23 (92%)	2 (8%)	0 (0%)	0 (0%)	**25 (100%)**
On my personal website	21 (84%)	4 (16%)	0 (0%)	0 (0%)	**25 (100%)**
Through interpersonal exchange with another researcher	7 (25%)	11 (39%)	7 (25%)	3 (11%)	**28 (100%)**
Other	11 (79%)	2 (14%)	1 (7%)	0 (0%)	**14 (100%)**

Table 13. Availability and modality of data distribution.

on these websites. The numbers were similar for availability of some data either through a national (15%) or on their personal website (16%). Little data were available through either a regional or a global network. The respondents did report that some or most of the data were available through interpersonal exchange with another researcher (64%).

The majority of the respondents (54%) agreed that their organization or project has a formal established process for managing data in the short-term. However, only a smaller percentage (42%) of respondents reported that there was a process for storing data beyond the life of their projects (*Table 14*). Similarly 30% agreed that their organization provided tools to support data management on short-term projects and 23% reported this was the case for long-term project. Only 20% respondents agreed that their organization provided any training on data management issues. Respondents also reported that their organization provided minimal funds for data management in the short-term (31%) and only 12% of respondents said that they had long-term funding support for data management.

VIEWS ON DATA SHARING WITHIN RESPONDENTS' COMMUNITY

When asked to comment on the use of data in their research field, almost half of the respondents (48%) agreed that lack of access to data generated by other researchers for a major impediment to progress (*Table 15*). When it came to their own research fewer researchers (29%) reported that lack of data access restricted their ability to answer research questions. A large majority felt that data might be misinterpreted if shared - both due to complexity (74%) and poor quality of data (63%) and almost all respondents were concerned that shared data might be misused (89%).

My organization or project:	Agree strongly	Agree somewhat	Neither agree nor disagree	Disagree somewhat	Disagree strongly	Total
…has a formal established process for managing data during the life of the project (short-term).	6 (23%)	8 (31%)	2 (8%)	9 (35%)	1 (4%)	**26 (100%)**
…has a formal established process for storing data beyond the life of the project (long-term).	4 (15%)	7 (27%)	3 (12%)	10 (39%)	2 (8%)	**26 (100%)**
…provides the necessary tools and technical support for data management during the life of the project (short-term).	4 (15%)	4 (15%)	6 (23%)	9 (35%)	3 (12%)	**26 (100%)**
…provides the necessary tools and technical support for data management beyond the life of the project (long-term).	2 (8%)	4 (15%)	4 (15%)	12 (46%)	4 (15%)	**26 (100%)**
…provides training on best practices for data management.	2 (8%)	3 (12%)	5 (20%)	9 (36%)	6 (24%)	**25 (100%)**
…the necessary funds to support data management during the life of a research project (short-term).	5 (19%)	3 (12%)	2 (8%)	10 (39%)	6 (23%)	**26 (100%)**
…provides the necessary funds to support data management beyond the life of the project (long-term).	2 (8%)	1 (4%)	5 (19%)	9 (35%)	9 (35%)	**26 (100%)**

Table 14. Organizations' involvement with data.

	Agree strongly	Agree somewhat	Neither agree nor disagree	Disagree somewhat	Disagree strongly	Total
Lack of access to data generated by other researchers or institutions is a major impediment to progress.	3 (11%)	10 (37%)	6 (22%)	7 (26%)	1 (4%)	**27 (100%)**
Lack of access to data generated by other researchers or institutions has restricted my ability to answer research questions.	2 (7%)	6 (22%)	5 (19%)	9 (33%)	5 (19%)	**27 (100%)**
Data may be misinterpreted due to complexity of the data.	8 (30%)	12 (45%)	2 (7%)	4 (15%)	1 (4%)	**27 (100%)**
Data may be misinterpreted due to poor quality of the data.	6 (22%)	11 (41%)	7 (26%)	2 (8%)	1 (4%)	**27 (100%)**
Data may be used in other ways than intended.	13 (48%)	11 (41%)	1 (4%)	1 (4%)	1 (4%)	**27 (100%)**

Table 15. Views on the use of shared data.

	Frequency	Percent
Lack of funding	4	15
Lack of standards	4	15
People don't need them	2	8
There is insufficient time to make them available	10	38
There is no place to put them	5	19
They shouldn't be available	9	35
Sponsor doesn't require it	2	8
Don't have the rights to make the data public	12	46
Other	12	46

Table 16. Potential reasons for unavailability of data (check all that apply).

When asked why their research data were not available to other researchers (respondents could select more than one item) respondents stated that they do not have the rights to make the data public (46%) and also insufficient time to make them available (38%) (*Table 16*). Some respondents also reported that they thought data should not be available (35%). In their open-ended responses (*Table 17*), participants expressed a concern with protecting the identity of participants, and the difficulty of sharing data (for example, artifacts collected at a field site). One respondent also noted that they had never been asked to share data and another commented that they had never considered it.

Other (detail)
Never considered it, would have to look into ethics.
No one has ever asked
I'm at a small teaching institution, so I have no funding for my work.
No person on the team who is good at website design
It would be too difficult, if not impossible, to protect the identity of the participants.
Is made available through publication
potential for loss of confidentiality precludes sharing
Not a priority and therefore I do not make the time to organize this.
Data is often in the form of physical artifacts that lose much in translation to digital forms. Data is not translated to be shared until requested.
We have committed to making our data public, but we haven't done it. (We will start soon.)
The IRB Process and Rules are confusing and cumbersome. It is not worth my time to try to figure out how to share my successful practices or learned failures.
The data are available upon request

Table 17. Detail reasons for 'Other' category in Table 16.

EXPERIENCE WITH DATA COLLECTION AND MANAGEMENT

When we asked respondents about their experiences with collection and use of research data, most of them expressed satisfaction with their processes for collecting and storing research data for the large part. But participants were dissatisfied with the process of storing data for long-term use (*Table 18*). 40% of respondents largely disagreed with the statement "I share my data with others." However, almost the same number of respondents (44%) agreed with the same statement. 65% of the respondents reported that there was no procedure for others to easily access their data. Overall, we find that the community has not paid much attention to issues of data sharing and therefore existing practices are geared towards conducting research in small teams.

One of the crucial items in the survey gauged respondents' attitudes towards data sharing and their potential actions if they had the opportunity to share data (*Table 19*). 78% of respondents agreed that they would use other researchers' data if it were shared. 61% agreed that they would place at least some part of their own data in a repository without any restrictions. Respondents overwhelmingly disagreed that that they would place all of their data in a repository with no restrictions on it. About 78% of respondents said they were more likely to make data available if they could place some conditions on access. Finally, 85% respondents agreed that it was important to cite their data

	Agree strongly	Agree somewhat	Neither agree nor disagree	Disagree somewhat	Disagree strongly	Total
I am satisfied with the process for collecting my research data.	8 (30%)	17 (63%)	1 (4%)	1 (4%)	0 (0%)	**27 (100%)**
I am satisfied with the process for searching for my own data.	3 (11%)	15 (56%)	4 (15%)	3 (11%)	2 (7%)	**27 (100%)**
I am satisfied with the process for cataloging/ describing my data.	2 (7%)	17 (63%)	2 (7%)	4 (15%)	2 (7%)	**27 (100%)**
I am satisfied with the process for storing my data during the life of the project (short-term).	4 (15%)	14 (52%)	5 (19%)	3 (11%)	1 (4%)	**27 (100%)**
I am satisfied with the process for storing my data beyond the life of the project (long-term).	2 (7%)	7 (26%)	4 (15%)	11 (41%)	3 (11%)	**27 (100%)**
I am satisfied with the process for analyzing my data.	12 (45%)	12 (45%)	1 (4%)	1 (4%)	1 (4%)	**27 (100%)**
I share my data with others.	1 (4%)	10 (40%)	4 (16%)	7 (28%)	3 (12%)	**25 (100%)**
Others can access my data easily.	0 (0%)	4 (15%)	5 (20%)	6 (23%)	11 (42%)	**26 (100%)**
I am satisfied with the tools for preparing my documentation related to my data.	3 (12%)	11 (42%)	7 (27%)	4 (15%)	1 (4%)	**26 (100%)**

Table 18. Experience with collecting and using research data.

	Agree strongly	Agree somewhat	Neither agree nor disagree	Disagree somewhat	Disagree strongly	Total
I would use other researchers' datasets if their datasets were easily accessible.	8 (30%)	13 (48%)	4 (15%)	1 (4%)	1 (4%)	**27 (100%)**
I would be willing to place at least some of my data into a central data repository with no restrictions.	5 (19%)	11 (42%)	6 (23%)	2 (8%)	2 (8%)	**26 (100%)**
I would be willing to place all of my data into a central data repository with no restrictions.	1 (4%)	4 (15%)	5 (19%)	10 (39%)	6 (23%)	**26 (100%)**
I would be more likely to make my data available if I could place conditions on access.	8 (31%)	9 (35%)	6 (23%)	0 (0%)	3 (12%)	**26 (100%)**
I am satisfied with my ability to integrate data from disparate sources to address research questions.	2 (7%)	4 (15%)	9 (33%)	8 (30%)	4 (15%)	**27 (100%)**
I would be willing to share data across a broad group of researchers who use data in different ways.	4 (15%)	14 (54%)	5 (19%)	0 (0%)	3 (12%)	**26 (100%)**
It is important that my data are cited when used by other researchers.	14 (54%)	8 (31%)	4 (15%)	0 (0%)	0 (0%)	**26 (100%)**
It is appropriate to create new datasets from shared data.	5 (20%)	11 (44%)	8 (32%)	0 (0%)	1 (4%)	**25 (100%)**

Table 19. Viewpoints on sharing research data.

if it was used by other researchers. These responses clearly provide very strong guidelines for any data sharing infrastructure that would be useful to the engineering education community.

Assuming the plausibility of data sharing in the future, we asked respondents what they would consider as conditions for a fair exchange for data use. We split the question into two parts – their opinion of fair exchange if their data were being used (*Table 20*) and their opinion if they were using other people's data shared with them (*Table 21*). Overall, we did not find any significant differences across the items and overwhelmingly respondents that they *did not* think that it was fair to get or give co-authorship simply because data were being shared. However, the results indicate that it was fair to expect formal acknowledgement of data providers through citation of data providers and/or funding agencies when shared data were used. They also reported that it was fair to expect the opportunity to collaborate on a project if data were shared. Majority of respondents agreed that it was fair for the data provider to review findings based on their data but not for them to approve them prior to dissemination. Almost all respondents also agreed that the data provider must be given a list of all products that used the data (papers, articles, presentations, etc.) and there was also majority consensus among the respondents that legal permission for data use should be there and/or mutual agreement between the parties.

	Fair	Not fair	Total
Co-authorship on publications resulting from use of the data	9 (35%)	17 (65%)	**26 (100%)**
Formal acknowledgement of the data providers and/or funding agencies in all disseminated work making use of the data	27 (100%)	0 (0%)	**27 (100%)**
Formal citation of the data providers and/or funding agencies in all disseminated work making use of the data	26 (96%)	1 (4%)	**27 (100%)**
The opportunity to collaborate on the project (including, for example, consultation on analytic methods, interpretation of results, dissemination of research results, etc.)	26 (96%)	1 (4%)	**27 (100%)**
Results based (at least in part) on the data could not be disseminated in any format without the data provider's approval.	9 (35%)	17 (65%)	**26 (100%)**
At least part of the costs of data acquisition, retrieval or provision must be recovered.	12 (46%)	14 (54%)	**26 (100%)**
Results based (at least in part) on the data could not be disseminated without the data provider having the opportunity to review the results and make suggestions or comments, but approval not required.	12 (46%)	14 (54%)	**26 (100%)**
Reprints of articles that make use of the data must be provided to the data provider.	18 (69%)	8 (31%)	**26 (100%)**
The data provider is given a complete list of all products that make use of the data, including articles, presentations, educational materials, etc.	25 (96%)	1 (4%)	**26 (100%)**
Legal permission for data use is obtained.	18 (67%)	9 (33%)	**27 (100%)**
Mutual agreement on reciprocal sharing of data	18 (69%)	8 (31%)	**26 (100%)**
The data provider is given and agrees to a statement of uses to which the data will be put.	20 (77%)	6 (23%)	**26 (100%)**

Table 20. Conditions on using your data.

	Fair	Not fair	Total
Co-authorship on publications resulting from use of the data	8 (32%)	17 (68%)	**25 (100%)**
Formal acknowledgement of the data providers and/or funding agencies in all disseminated work making use of the data	23 (96%)	1 (4%)	**24 (100%)**
Formal citation of the data providers and/or funding agencies in all disseminated work making use of the data	23 (96%)	1 (4%)	**24 (100%)**
The opportunity to collaborate on the project (including, for example, consultation on analytic methods, interpretation of results, dissemination of research results, etc.)	21 (88%)	3 (12%)	**24 (100%)**
Results based (at least in part) on the data could not be disseminated in any format without the data provider's approval.	6 (26%)	17 (74%)	**23 (100%)**
At least part of the costs of data acquisition, retrieval or provision must be recovered.	9 (41%)	13 (59%)	**22 (100%)**
Results based (at least in part) on the data could not be disseminated without the data provider having the opportunity to review the results and make suggestions or comments, but approval not required.	13 (57%)	10 (43%)	**23 (100%)**
Reprints of articles that make use of the data must be provided to the data provider.	16 (70%)	7 (30%)	**23 (100%)**
The data provider is given a complete list of all products that make use of the data, including articles, presentations, educational materials, etc.	22 (96%)	1 (4%)	**23 (100%)**
Legal permission for data use is obtained.	15 (65%)	8 (35%)	**23 (100%)**
Mutual agreement on reciprocal sharing of data	17 (74%)	6 (26%)	**23 (100%)**
The data provider is given and agrees to a statement of uses to which the data will be put.	20 (87%)	3 (13%)	**23 (100%)**

Table 21. Conditions for using other people's data.

	Position	Research Method	Gender	Area
1	Associate Professor	Qualitative/Mixed	F	Motivation
2	Associate Professor	Qualitative	M	Conceptual Knowledge
3	Assistant Professor	Experimental/Mixed	F	Design
4	Assistant Professor	Quantitative/Mixed	M	Institutional Issues
5	Graduate Student	Quantitative	M	Student Performance
6	Graduate Student	Qualitative	F	Student Collaboration

Table 22. Interview Informant Demographic.

Interview Findings

We conducted interviews with a sub-section of the survey respondents (all interview participants had volunteered). Interviews ranged from 15 minutes to 40 minutes in length. The interview protocol was open-ended with three broad questions. All subjects were asked:

- What are your thoughts on sharing of research data?
- What are the potential barriers to data sharing?
- What resources or infrastructure will you need to share data?

The interview participants were selected to ensure a representative sample in terms of career trajectory, research methods, and gender. *Table 22* provides details.

Given the small sample, we did not code the interviews but present the findings under some broad headings. We include lengthy quotes from participants in our write-up to capture their voices. The quotes though are *not* verbatim as we had to revise them to preserve the confidentiality of subjects (given the small research community it would be easy to otherwise identify the respondents). The quotes are as close to the intended meaning as possible.

Thoughts on data sharing

The first question we asked subjects was to reflect on data sharing in engineering education. All interview subjects agreed that data sharing was beneficial and that the engineering education community did not do enough data sharing. Participants provided different justifications for why data sharing was important. Some were concerned with a lack of replication of research while others felt that there was too much duplication and therefore resources were not being utilized efficiently. Interview participants also raised some concerns with data sharing, such as interpretation of data, and lack of infrastructure to share data. Overall, all subjects expressed an interest in sharing their data and also using other researchers' data provided the details could be worked out.

"It is a good thing for us to do. Different perspectives on data and replication are important. Not just one person who does the study but somebody should go replicate the study; it's easy to mess things up by accident; we don't do enough replication in the engineering education community." [1]

"We should do more of it. We study the same thing. People get funding to study the same thing. We should be able to use existing data. I am doing a journal paper on women in engineering and the findings are the same as they have been for years. We need to stop duplicating our efforts. I think it's stupid to keep studying the same things again and again." [1]

"If it is publicly funded, it should be shared. There are limited resources and if more people look at it we will learn more. There have to be some safeguards for the data and for the PI but all publicly funded data should be shared. In my research I use secondary data, particularly from the department of education, and they have proper procedures. I have to show them the tables, not the findings, and they review them and let me know if they are correct. You have to sign a contract with them. So there are exemplars out there." [2]

"We do an awful lot of interviewing of students, faculty, so my first thought is that it is so contextual and contains personal information, I am not sure how useful it will be. My second thought is that it will be awesome; we've terabytes of interview data and it will be great if people can do something with it." [4]

"Data sharing is an interesting idea, since I've done a lot of interviews and the thing with interviews is you don't know where they will go, they might be useful for other people." [6]

"In my area of research there is a lot of talk about sharing of data, and even things like problem sets, but we've run into the issue of data provenance and sometimes we are unable to back-track the analysis or provide the data. I think systems for data management are crucial for data sharing." [6]

"Let me play the devil's advocate and argue that research is such a socially shared enterprise, you can share your data, but I don't have to. It will be interesting to look at the history of folks who actually share data and understand in what ways we can develop shared norms and get buy in. We need to extract value from the shared data – what is the

> value of me sharing it – authorship? Also, there are other examples in the community of creating shared enterprises such as [a hub] that haven't always worked out." [4]

> "Data collection is hard; re-analysis is not that hard. Once you are done with using the data, it is easier to share; if you are still using it, it's not so easy. People I know if they found something in my data, they would tell me and not publish them." [5]

> "One of the reasons I put a paper in the poster session to find out who else got funded to do work I'm doing and I need to meet and talk to these people. I'm not sure why this is [lack of sharing]. Maybe the culture of patents and intellectual property restricts sharing. We need to know what people are already doing; all the money that has been spent collecting data and researchers have very low publication record. We don't get out of the data what we should. I've got a grant expiring this year with no journal papers. There is no substantial long term contribution." [1]

Infrastructure Needs

We next asked participants what kind of support they would need to be able to share data. They expressed a need for an online repository where data could be stored. They preferred the repository to have a permission structure where users could be given different levels of permissions and also permission to some or all of the data. All participants expressed the need for the infrastructure to be easy to use. Furthermore, they expressed different opinions about IRB issues; most saw it as a concern but some suggested that with valid training and credentialing this could be worked out especially since they had added researchers to projects earlier.

> "I haven't looked into tools, websites, etc. So far I've used Scholar site and now it's going away. I believe it was a secure way to share data. [An existing hub-based system] was too painful. Part of me says that the tools have to be simple. Download/upload in multiple ways is too hard. There is no room for projects on [the learning management system used by my institution], which is what my institution will be using. I need a place, a repository, which is easy to use and comes with standardized agreements. There have to be some rules of engagement around sharing the data and if someone else was managing all that, it would make it more inviting. I know that IRB will be a challenge but as long as the site is secure and password protected it should be workable. Maybe a credential based training that all institutions accept. It's got to be easy for me to do. I want somebody else to think about all these hard questions." [1]

"In terms of infrastructure, data will need to be housed and larger datasets will need a data warehouse of some sorts. There has to be some kind of login but beyond that there isn't much more necessary. IRB is tricky; it depends how the first project is pitched; IRB is less strict on secondary data analysis. Maybe change the consent process; it varies so much across institutions anyway." [2]

"I need infrastructure that allows me to post the data and handle the sharing. Right now, if I trust a person I'm happy to email it to them. But if I don't know them, I want it to be iterative and in small steps – share some data and then see how it goes. I also want someone else to do the work. I should be able to just upload it." [4]

"A trust-based exchange is required where permissions can be assigned. Also, I need a way to manage laboratory data and workflows that allow linking of the lab data directly for sharing. Recently, in other fields analytical tools are also available online and that is something else that can be very useful; and also the ability to push data and analysis directly for publishing. I think there are a lot of HCI issues here." [5]

"The repository needs to be searchable. It is important to know things like: who is the interviewer and what is the purpose of the project? A list of sample questions, overall theme, and purpose of conducting the interview are important. Also, clean transcripts (without personal identifier) might need to be shared." [6]

"If someone were to use the data would be good to know, what did they do with it, some form of communication; list of people who have released data in the repository; it is not required to cite me; good to see a list of people who have shared data." [6]

Incentives

When we asked participants about infrastructure and resource needs, another related issue that came up was incentives for sharing the data. One common comment was that without proper incentives it would be hard to motivate researchers to share data and to use shared data. Not surprisingly, faculty commented on the lack of incentives for analyzing shared data if there was no funding related to the effort since funding was required for tenure and promotion. Interview participants suggested that having funding solicitations targeted solely for secondary analysis could incentivize collaboration around shared data. Another idea presented was to use the shared data as way to bring in new researchers and train graduate students. There was also general agreement that community norms have to shift towards valuing data sharing and use of shared data.

> "There has to be within university structure some value for data sharing; if I don't need to collect data versus research dollars. The system drives collect data since you can get money to collect data. One idea for incentives will be to bring in new researchers; people transitioning from other fields might like the fact that data exists. Maybe a mentoring system of sorts with value added for both parties. It's good for the field so people know what's going on." [1]

> "To incentivize data sharing there should be follow-on grants on data analysis and dissemination grant to bring other researchers on board. If NSF changed their model for a year, there is a lot of data out there. I think there has to be some stipulation about who gets authorship when the data is used but I think funding to bring new people on board is essential. There can also be a solicitation focused on secondary analysis. I think one barrier to data sharing is the merit review process within institutions for tenure and promotion; things such as 'how many people accessed your dataset' are not valued." [2]

> "I think there is a personal incentive for me – just to be able to collaborate with somebody else. There is so much data we cannot mine it all and there might be people in the community who want to work on it. If there was a site where everyone was able to look at the data and somebody would push your request and go deeper in the data [different levels of access to data]." [3]

> "I think PIs who want to get funded should need to argue why it's a good test case. Also, who will mentor students who are looking at the data? Publications can be an incentive. Pre-tenure it is hard to do as our institution cares only about money, there has to be some way for the pre-tenure person to get credit." [2]

Advantages

All subjects listed many advantages they thought could emerge with data sharing. Several of them worked with secondary datasets or with shared data and argued that secondary data analysis could be quite creative and was very useful for training new researchers. One participant cited the lack of engineering education data that could be used for quantitative research methods courses. Several subjects also mentioned that creating a dataset that is usable is a very time consuming process and therefore it is prudent to allow as many researchers as possible to use the data once the dataset was available. There would have to be protections for researchers who create that dataset but as a field it would be best to share. Finally, one participant commented that the field would get a lot of credibility if there was more shared research and data were available to other researchers.

> "I analyze secondary data which is not common in this community. I think it's the most creative activity since you can find things other don't. Since you didn't design the research, so that is a caveat, but you find interesting stuff. For sharing of interview data it will depend on how much you need to understand the context. The original PI needs to do member checking for context but quantitative data is easy as most shared datasets come with a pretty detailed code book." [2]

> "I've been writing proposals based on prior data and I've been using existing data to train my students. Later on they can go and collect their own data but it has been hugely valuable to get me up and my lab up and running." [2]

> "I think it is very important for the field to collaborate and to have replication; make sure what somebody produces is actually correct and we are not making big policy changes based on one person's faulty analysis." [3]

> "In a recent project, the data gathering process was going well for a while and then they needed to go talk to their higher-ups and everything slowed down and it took two more months after they got approval from administrators. Once the data was obtained, there was a lot of work just trying to understand what was in it and then to put it into a format for analysis (80% of time goes into making it usable). This was my first large data project it took me a lot of time. I had to learn R; steep learning curve. It was a cumbersome process to say the least. Therefore, if this data can now be shared and others can use it, it will same them a lot of effort. On the other hand, if I had access to similar data, it would save me time and resources." [3]

> "I took a quantitative methods course where you usually brought your own data. The examples in the course were through the "HIGH SCHOOL AND BEYOND" dataset from 1970s or 80s. R has a lot of built in datasets but there are not really any educational datasets I know off. It would be really awesome if there was an educational dataset that was more readily available. If you've a relevant dataset to your field it makes it a lot easier to

> make connections about how to use methods and understand the process of collecting and cleaning data – data determine methods." [3]

> "It might give credibility to our field if we are able to do data sharing." [4]

Concerns of Qualitative Researchers

The qualitative researchers we interviewed all expressed support for data sharing but echoed concerns commonly found in the literature around the complexity of interpreting qualitative data and protecting the privacy of subjects. Overall, though, the participants were in favor of sharing their data as all of them commented that they had collected a lot more data than they could usefully analyze. Interview participants also commented that data sharing could be a pathway towards a more honest discussion of research practices and results in better quality research.

> "As a qualitative researcher I get anxious over where and how to share data. I don't necessarily have setup consent form for data sharing. I tend to be hypersensitive about data. It's not just names but also people's story. I want to monitor and use data respectfully. I don't need to be co-author if I could sign-off. I am not talking about censorship to protect findings but data has to be used respectfully." [1]

> "Qualitative data exposes you a lot. I feel like you will have to defend yourself, 'why did you ask that stupid interview question?' If it the data is anonymous it's different but way less useful, authentic descriptions of settings are important, cold analysis is not what we need." [4]

> "Data sharing is an interesting idea, since I've done a lot of interviews and the thing with interviews is you don't know where they will go; they might be useful for other people. I feel like if people are going to take the trouble to go through your data they are not doing with the intent to critique but to make use of it." [6]

> "It doesn't bother me [quality of data]. Nobody has to see the exact data the same way I do. Given my codes can someone say that make sense; they don't have to like it. I'm not worried about judgment. If they have time to match data to article, they could be doing better things. Someone using my protocol and taking it to the next level; some things are working and some are not; changing and adaptation is good." [1]

> “I made mistake in trying to get the same data at different sites. I collected data all in the same batch. I come back and I want to go back and collect more but I can’t since my funding it over. We need to better understand research projects and how they should be organized and what is good data.” [1]

> “If someone did a workshop at FIE about how they manage data, everyone would attend it. Learning craft and tools of the trade is important and it is easier to share this knowledge then data. Data sharing can lead to this.”

> “You can add people to the IRB; as long as they have done the necessary training so it is not difficult to share data if it is through a trusted repository.” [6]

Secondary Data Analysis and Its Limitations

Finally, participants who had worked on secondary data before were alert to its limitation and the problems that can arise if data are not properly shared. The primary issue is the creation of “rules of engagement” around shared data: Who can use it? Who gets authorship? Participants also felt that although analyzing shared data is great, it is important for the field to continue to collect new data as social structures change and data are embedded in specific social times, which have their own dynamics.

> “My dissertation was on data that other people had collected. It was dream come true for me. I didn’t have to collect 4 years of data but I got to analyze it. What I got out of it made it completely worthwhile. Get what you can out of the data and figure out what you want. Good rules of use in the secondary dataset I analyzed; such as, how it will be used, shared, made available – what are the rules of engagement?” [2]

> “New data is important. A lot of work is based on old data such as XYZ’s work but kids are not the same anymore; kids don’t need to have the same trajectory anymore. We need to pay attention to what is going on in the world right now; race issues this year. When does data expire?” [1]

DISCUSSION AND RECOMMENDATIONS

To better understand the culture of data sharing within the engineering education community, we conducted a small mixed-methods study consisting of a survey and interviews with members of the community. Overall, our findings suggest that few, if any, members of the community currently share research data other than in the context of collaboratively funded projects. Community members expressed a great interest in sharing their data and using shared data with the caveat that they would have to ensure that research participants' privacy is protected. This trend is consistent with those expressed by researchers in other domains (Tenopir et al. 2015). Participants also conveyed a desire for a sharing mechanism that is minimally time invasive and easy to use. Finally, all informants raised the issue of incentives, or the lack thereof, for sharing research data.

Based on our analysis, we recommend the following future actions to improve data sharing within the engineering education community. We believe that a wider discussion within the community is still needed and these points can be used as guidelines for the conversation:

1. <u>Data Management & Repository</u>: Engineering education researchers can better utilize data workflows designed to collect and mark data throughout the research process. The use of data workflows is common across the natural sciences. This will make sharing of data easier by creating common norms across labs and also reduce the time required to clean and prune the data as well as attach metadata to them. The data repository will need to have different access levels so that data can be shared in small amounts initially and with those who have the requisite training. The system should be easy to use where researchers can just capture their entire data workflow and not have to spend time putting it in shape to be shared. The metadata should be easily assignable. Finally, multiple stakeholders, such as graduate students, instructional faculty, seasoned researchers, should be considered in the design of the repository.
2. <u>Incentives for Data Sharing</u>: Although researchers are not averse to data sharing, they feel that there is a lack of incentive to share data. They report that some of this can be mitigated if data sharing becomes required (at least on publicly funded projects). However, this still will not solve the problem of using shared data. Currently, most projects are funded for data collection and analysis rather than secondary analysis of data. This can be changed by solicitations for funding that primarily target analysis of shared data. Researchers are also wary about the lack of credit for creating data infrastructures and for sharing data. Therefore, creating norms for citing shared data and recognizing the activity that in promotion and tenure is necessary.
3. Implementation Issues: Finally, many implementation issues need to be sorted out. For instance, who will create and manage the repository or what mechanisms can be designed for individual researchers to manage the process of data sharing while not having to focus on infrastructural issues? Also, what mechanisms will promote the use of shared data and how can community norms be created that respond to data sharing challenges? Many suggestions were made by community members to address this problem. Initially, it might be prudent to create test-cases

of specific kinds of shared datasets, probably on topics of interest to the community of a population of interest, and these should be opened to a small subgroup of the community. A consistent suggestion was creating a test-bed for graduate students who are currently in the process of research training. Another suggestion was to open it up to instructors who are interested in research and reflection but might not have the resources to collect their own data.

CONCLUSION AND LIMITATIONS

We conducted a mixed-method study to better understand the data sharing practices within the engineering education community. We found that although sharing of research data is not common among engineering education researchers, our respondents were open to the idea of sharing data and using shared data. Lack of incentives to share data inhibit data sharing as both the process of sharing and using shared data is resource intensive. Our dataset is small and therefore generalizations are hard to make but the sample is representative of active researchers in the field.

ACKNOWLEDGEMENTS

This research was partially supported by NSF Award EEC-1408674 and EEC-1306377.

REFERENCES

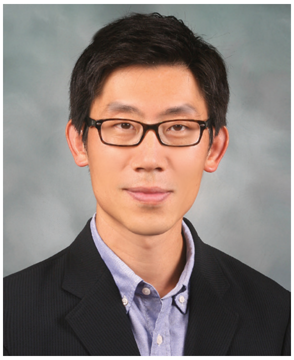
Allameh, E.M., Vorvoreanu, M., Yang, S., Johri, A., & Madhavan, K. "Informing the Sharing and Access of Engineering Education Research Data through Comparative Analysis." Published in the Proceedings of the 122nd American Society of Engineering Education Annual Conference and Exposition, Seattle, WA. (2015).

Tenopir C, Dalton ED, Allard S, Frame M, Pjesivac I, Birch B, et al. (2015) Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. PLoS ONE 10(8): e0134826. doi:10.1371/journal.pone.0134826

Tenopir C, Allard S, Douglass K, Aydinoglu AU, Wu L, et al. (2011) Data Sharing by Scientists: Practices and Perceptions. PLoS ONE 6(6): http://dx.doi.org/10.1371/journal.pone.0021101

AUTHORS

Aditya Johri is an Associate Professor in the Department of Information Sciences & Technology in the Volgenau School of Engineering, George Mason University, Fairfax, VA, USA. He studies the use of information and communication technologies (ICT) for learning and knowledge sharing, with a focus on cognition in informal environments. He also examines the role of ICT in supporting distributed work among globally dispersed workers and in furthering social development in emerging economies. He is a co-editor of the Cambridge Handbook of Engineering Education Research (CHEER), Cambridge University Press (2014). He can be reached at johri@gmu.edu. More information about him is available at: http://mason.gmu.edu/~johri



Seungwon Yang is an assistant professor in the School of Library and Information Science (SLIS), and the Center for Computation and Technology (CCT) at Louisiana State University (LSU). His research is focused around archiving/analyzing/visualizing information, especially applying data mining and natural language processing techniques to social media data from disaster events and online communities. Seungwon earned his Bachelor's, Master's, and Ph.D. degrees all from Virginia Tech (VT). His prior work includes digital library curriculum development, and visual analytics tool development as well. Before arriving to LSU, he was a postdoctoral research fellow in the Department of Information Sciences and Technology at George Mason University. During that time, his main research was to measure informal learning occurring in a large online community of creators applying data mining techniques. In his free time, he enjoys cycling and meditation. Dr. Yang was a postdoctoral researcher at George Mason University when this research study was done.

Mihaela Vorvoreanu is an associate professor in Computer Graphics Technology at Purdue University, where she leads the Human Centered Design and Development concentration. Her research focuses on user experience design, as it specifically applies to creating useful and useable systems. Dr. Vorvoreanu leads the UX research and design on DIA2.com, an NSF-funded project for knowledge mining and visualization. Other research interests include UX education and the impact of new communication technologies such as social media on culture, society, and organizations.

Krishna Madhavan is an associate professor in the School of Engineering Education. In 2008 he was awarded an NSF CAREER award for learner-centric, adaptive cyber-tools and cyber-environments using learning analytics. He leads a major NSF-funded project called Deep Insights Anytime, Anywhere (http://www.dia2.org) to characterize the impact of NSF and other federal investments in the area of STEM education. He also serves as co-PI for the Network for Computational Nanotechnology (nanoHUB.org) that serves hundreds of thousands of researchers and learners worldwide. Dr. Madhavan served as a Visiting Researcher at Microsoft Research (Redmond) focusing on big data analytics using large-scale cloud environments and search engines. His work on big data and learning analytics is also supported by industry partners such as The Boeing Company. He interacts regularly with many startups and large industrial partners on big data and visual analytics problems. He was one of 49 faculty members selected as the nation's top engineering educators and researchers by the U.S. National Academy of Engineering to the Frontiers in Engineering Education symposium.