Advances in Engineering Education

SPRING 2016

# Data Sharing and Reuse within the Academic Pathways Study

GEORGE TOYE

SHERI SHEPPARD

AND

HELEN L. CHEN
Stanford University
Stanford, CA

## ABSTRACT

The Academic Pathway Study (APS) research program within National Science Foundation (NSF) Center for Advancement of Engineering Education (CAEE) ran from 2003 - 2010. It amassed a collection of longitudinal as well as cross-sectional data sets, of varying research method types and formats, from four different primary cohorts that included over 5,300 subjects at over 25 geographically distributed sites. The APS research team was similarly dispersed, at an initial four academic institutions that eventually grew to over eight schools. This diverse research team was co-organized along two major lines of responsibility, affiliated institutions and research methods. Because the team was so geographically distributed and the collection of data so large and heterogeneous, APS leadership recognized early on in the project that effective management of data collection, sharing, and reuse would be essential to the success of their collaborative research efforts. Many observations, lessons learned and reflections have emerged from those experiences with APS cross-institution data sharing and reuse – from features of collaboration tools, to Institutional Review Board (IRB) strategies, to guiding protocols and policies.

**Key words:** Data Sharing, Data Reuse, Database, Institutional Review Board (IRB), Tacit Knowledge

## BACKGROUND

### Academic Pathways Study

In 2003, the Center for the Advancement of Engineering Education (CAEE) was funded by the National Science Foundation (NSF). The largest component of CAEE, the Academic Pathways Study

(APS), was a multi-method study to describe how people navigate their undergraduate education to become engineers. The study was initially designed to be a collaboration of four schools: Colorado School of Mines, Howard University, Stanford University, and University of Washington.

**Multi-Institution Research Team - Organization**

While the overall APS research lead investigator was at Stanford University, there were also representative leads at each of the collaborating schools. Each representative lead (who was also a principal co-investigator) was responsible for coordinating development of one or more of the project's research methods and the consistent implementation of methods across the various campuses. Additionally, each institution was responsible for addressing any concerns related to their respective Institutional Review Boards and the collection of data on their campuses. Although specific campus principal co-investigators were designated to lead different components of the research, the overall research team actively collaborated on all aspects of the project, including subject recruitment, instrument design and implementation, data processing and analysis. The composition of the combined team was interdisciplinary, bringing with them different competencies and backgrounds: Engineering, Education, Communication, Computer Science, Psychology, Evaluation, and Anthropology. One might visualize the APS team as an integrated network of diverse and distributed research collaborators.

Research sub-teams were formed to work on various aspects of data cleaning, processing and analysis for each data set (which are described more in the next section). Almost throughout, the lead principal co-investigator for each respective research method would oversee the cleaning and processing of that method's data. In these teams, a majority of members were geographically clustered at the lead principal co-investigator's institution, but consistently included remote researcher members as well.

At the nexus for data was a specially designated sub-team at Stanford responsible for aggregating, storing, archiving and managing a network server that enabled collection and redistribution of data between the various researchers. In the years between 2003-2010, large volumes of data were shared across the many institutions represented by the research team. This sharing was conducted under selective access permissions, appropriately controlled according to each user's specific participation needs.

**APS Data**

The APS study collected many and diverse data from four primary cohorts:

**Longitudinal Cohort**: Students who expressed interest in majoring in engineering upon admission at four institutions.

- Study group n=160 students;

- 40 students each at 4 universities – including 8 students for additional ethnographic study;
- Data to be collected across 4 years;
- Collected data types: structured interviews (once per year), semi-structured ethnographic interviews (once per year), exit interviews (per occurrence), surveys (twice per year), skills and concept-based tests and interviews (once per year), ethnographic observations of a subset of students (variable), academic transcripts (at least once a year);

**Workplace Cohort**: New professional engineers employed in various settings.

- Actual study group participants n = 111;
- Data collected over a period of approximately two years;
- Collected data types: interviews, ethnographic observations and comparative analyses of Skills and knowledge used in school and work;

**Broader Core Sample**: Engineering undergraduates at the four Longitudinal Cohort institutions who were not in the Longitudinal or Workplace cohorts.

- Study group n > 2000 in original design; actual participants n = 842;
- Data collected at one point in time;
- Collected data type: cross-sectional survey developed from the evolving research results (once);

**Broader National Sample**: Undergraduate students from engineering programs at 21 institutions across the country.

- Study group n > 3000 in original design; actual participants n = 4266;
- Data collected at one point in time;
- Collected data type: cross-sectional survey developed from the evolving research results (once);

The data amassed in the APS came from over 5379 individual subjects and over 100 discrete data gathering events across a period 4 years, at over 25 geographically distributed locations. The official APS data store was vast and diverse. Numerous other unofficial data were accumulated from various pilot studies and trials, as part of the development of the portfolio of APS instruments. In addition, because of the longitudinal research component, key personally identifiable information for subjects involved in that component of study had to be securely maintained as well. Keeping all of these data well managed for researchers to easily retrieve, process, analyze, share, and reuse would be critically important to APS' success.

### Institutional Review Board (IRB)

A certain level of IRB application complexity was anticipated and encountered early, due to the diversity of research instruments, the volume of data collected, the research team being geographically distributed and its multi-component leadership organization. IRB applications had to

be submitted to each of the principal co-investigator's respective institutions to support the work with the Longitudinal and Broader Core Samples; these applications had to be coordinated and synchronized to ensure that 1) procedures for data collection would be performed uniformly across all four institutions, 2) data processing and analysis could be performed and lead by research method leads employed at another institution, and 3) researchers at all four campuses could have equal permissions to access the entire aggregate data set collected from each of the four institution. Jurisdictional approval for how data would be brought together, shared and managed in accordance to oversight by (and at) multiple institutions' IRBs was a concern.

Example of complexity: IRB applications were required to cover the initial four focal institutions for the longitudinal study: Colorado School of Mines, Howard University, Stanford University, and University of Washington. Lacking an active IRB at Colorado School of Mines, Howard University extended their IRB coverage. Independent human subjects applications were submitted to IRBs at Howard University, Stanford University and University of Washington. Each campus' IRB responded differently, requiring custom changes to the specifications for data handling, sharing, and longevity, and in concert, the text in the base subject consent form. These variances would be resolved in time, but caused downstream complications for data captured in that first year of APS.

Another example: an APS researcher transitioned to a faculty position at a new institution that had no previously established IRB. In that situation, a new IRB had to be contracted externally and was eventually established at the institution in order for the researcher to continue the collaboration.

These IRB-related factors shaped the research procedures, defined the extent to how data could be directly shared within the confines of the APS, and prescribed what could be shared and reused in the subsequent years beyond the study.

## DATA'S CONTINUOUS EVOLUTION AND FLOW

### Organizing, Cleaning and Processing

Data are not static. They bear similarities to life – more delicate at birth, expanding and growing in features, hybridizing to spawn new sprouts, and yielding more interesting and productive outcomes over time. In fact, the characteristics of data and consequently how data must be handled is importantly nuanced; this too changes across multiple years on a longer longitudinal research project like APS.

During the time span of APS, researchers were continuously and jointly involved in collating, cleaning, processing and analyzing data. The data set evolved and grew with each new collection event. As defined in the study's original design, there would be diversity in data instruments and

collected data. It was recognized that the non-homogeneous characteristics of APS data compilation would require an organization for this data that is

- Flexible (requirement #1),
- Simple to understand (priority #1), and
- Simple to access (demand #1).

An example of the complexity associated with organizing and managing non-homogeneous units of data: With surveys, data arrived as raw text files with comma-separated values (CSV) – one file per school, per survey deployment. Each CSV file would contain all the survey responses from all subjects at one institution for a given deployment; four original data files resulted from each longitudinal cohort survey deployment. On the other hand, data from an interview deployment would contain raw audio recordings and occasionally supplemented with separate text note annotations for each interviewed subject. In this case, one interview deployment to all 160 longitudinal cohort subjects yielded approximately 180 original data files – a separate data file for each interview's audio, plus additional interview note files with contextually relevant observations that are determined valuable to record and share.

All these data in varied file types and formats had to be organized to be retrievable by individual subject's anonymized ID, by subject's affiliated institution, by cohort, by data collection episode or event time, and by instrument method. The varied types and formats of original data document files included:

- Interview audio - .dss (proprietary format from Olympus digital recorders), .mp3, m4a, .aiff (Apple audio)
- Interview text notes, and text transcriptions - .txt, .doc (Microsoft Word), .rtf (rich text format)
- Survey data - .cvs (comma-separated values), .xls (Microsoft Excel)
- Academic transcript data - .pdf (scans of paper records), .mdb (Microsoft Access), .xls (Microsoft Excel)

As researchers clean and process these original files in preparation for analysis, new files are generated along with possible corresponding new file types. The transcription of a .dss format interview audio recordings begets text in a .rtf (rich text format) or .doc (Microsoft Word) format; the import of a survey's tabular data in .cvs format file into Microsoft Excel begets a .xls format file. These incrementally new files are also added to the overall shareable data set, growing the APS Database's overall size. Moreover in so doing, a distinctive new dimension to the data set is introduced, versioning.

For example, digital audio was recorded for each conducted interview. The process to transcribe audio to text is a particularly arduous one that required creation and maintenance of different file versions along each step of the way. Sometimes, the number of steps were many. There were frequent

additional challenges due to occasional audio recording quality issues, and some subjects' heavy spoken accents. At times, external service providers, such as audio-text transcriptionists, were also engaged. To assure quality of these transcriptions, checks and subsequent double-checks by different people were essential. Then eventually, after all audio text transcripts have been confirmed, finalized and collected, additional multiple review passes are required to anonymize the data, by replacing all the interview subject's personally identifiable information with key codes. Moreover, other name-specified (or descriptively recognizable) individuals or places/locations had to be replaced with codes as well. This would result in multiple incrementally more accurate, complete and more broadly sharable transcripts. Many file versions would be accumulated in this processing – for each interview subject, at each interview event.

Once researchers begin to pursue their analyses using new data analysis tools on cleaned and processed versions of the original data, new application specific data files are created. New corresponding data file types are introduced, such as: .sav (SPSS), .nvp (nVivo), .hpr# (ATLAS.ti). Branching occurs naturally in research explorations. New observations would spark new questions and analyses. In pursuing these branches, data would be reorganized, modified, and reduced in new and different ways. Expectedly, the number of these analysis-supporting data files grow with use.

Not only is versioning an important aspect of data cleaning and processing, it would be integral to research analysis and collaboration. The needs to save, identify, organize and share successive branching generations of analysis data files during the analysis phase, mirror the needs to save, indentify, organize and share incrementally cleaner process iterations of original data files during the previous data preparation phase. In fact, from the APS Database perspective, these analysis-supporting data files are still fundamentally data files, handled identically. These newest additions simply extend the versioning dimension. All are consistent and easily integrated into the overall store of data to be made available for sharing and reuse.

**IRB, Privacy and Data Access Concerns**

Because the data collected from APS research subjects are longitudinal over multiple years and collected via several different instruments, there was heightened sensitivity that by accumulating a combination of data personally identifiable information could be revealed by triangulation. Even though each of the study subjects was assigned an anonymized identification code and respectively collated, data from a subject's survey responses in year one had to be connected to all corresponding subsequent responses to other surveys, interview questions, and academic transcript records. By enabling this kind of longitudinal analysis, the uniqueness of the combined data could reveal the subject's identity; privacy and anonymity would be eroded. As a result, additional attention was required to monitor and manage which researchers might have open access to large number of data sets.

Research instrument-specific methods defined in APS' original design enabled a relatively simple mapping between a researcher's limited usage permissions to data corresponding to a specific analysis method. Notably downstream, mixed methods research was introduced into APS where varied aggregates of data from different instruments would be analyzed together. The data had to be easily retrievable by individual subject, cross-sectionally by institution, across all institutions, by instrument deployment, by year, and across years. Data access requirements for researchers engaged in mixed methods posed a challenge to the existing data organization, user permissions model, and raised some IRB concerns as well.

In the design of APS, it was recognized that secure data sharing would have a central role in its processes and would be required along every step of this research collaboration. However, given the exploratory aspects of research in APS, how the data compilation might be accessed and analyzed could not be well anticipated or accurately predefined. Instead, the APS database system would have to be flexible and evolve along with the research.

**Filenames as Metadata**

Upon reflection, APS' early adoption and adherence to a loose, not too intricate, rationally structured file naming schema turned out to be foundational in helping manage many encountered complexities in file sharing.

Taking stock of the non-homogeneous, file-centric nature of APS data, the only real place to capture valuable metadata pertaining to the data file contents was in the filename. For ease of retrieval and organization, each file's name had to describe with brevity: the data collection event ID, affiliated institution ID, subject's coded ID, item type (audio, text), version number, number within a sequenced set, originating researcher's ID, and optionally in certain cases, a subject's friendly pseudonym. The file naming schema resulted in filenames that in format look like this:

**StudentID-MethodType-EventID-ItemID-ResearcherID-Pseudonym.FExt**

Examples:

```
CSM01F00003-INTS-1-A1_1-GT-Judy.dss
HU01F00025-INTS-2-T1_3-KE.rtf
SU01F00008-ETH-040306-N1_1-TLB.rtf
UW01M00034-INSP-1-S1_1-KO.pdf
```

This file naming schema was flexible enough to accommodate the association between these multiple data files as pertaining to one subject's data for a given research instrument's deployment. For example, data related to the performance tasks instrument included PDF scans of paper documents, as well as interview audio. The schema also could accommodate survey data, where a file

contained data associated with all subjects at an institution. In this case, the student id field in the filename was truncated to just the institution's ID.

By using regular expressions search on filenames, it was equally easy to locate all data associated with a data collection event, with a particular institution, or all data collected by a specific researcher. It would be no more complicated to find all of multiple file parts related to a subject's interview audio recording, regardless of actual format of the audio data (i.e., mp3, .dss, .m4a), or all text transcript revisions for a specified interview.

In having research collaborators from the four longitudinal cohort institutions collecting and uploading multiple data files for 40 subjects for each major data collection event, multiple times a year, across four years, mistakes in file naming would be inevitable. There were indeed mistakes, some more major than others. All were resolvable. There would also be occasional mistakes of not uploading all the data files expected in a collection, and mix-ups of upper and lower case characters in file naming. In striving to maintain consistency in adherence to this file naming schema, many mistakes in human handling that might have otherwise gone unnoticed, were caught and resolved quickly.

**Technology Infrastructure**

Data (information) sharing and reuse has been central in the research domain of distributed collaborative work. Today, there are many excellent internet-based computer supported collaboration work (CSCW) tools available today. These are relatively mature technologies, used widely in deployment of corporate intranets and network database applications. One might reasonably expect many of such commercial offerings (e.g., Dropbox, Box, Google Drive, Google Docs, Microsoft OneDrive) would be able to satisfy researchers' data sharing and reuse needs. However in fact, it was learned in conducting APS engineering education research that extraordinary and complex demands are placed on the use of data sharing tools, in order to adequately address information security and privacy considerations, human subject protocol and IRB requirements.

Mindful of the underlying data sharing complexities in APS' research design in early 2003, TikiWiki, an open source internet web software platform, was selected to be the computer based collaboration tool used to support this geographically distributed APS researcher. At the end of the day, the team needed to communicate and share data. A server with TikiWiki software was installed at Stanford for this purpose and was dubbed the "APS Database" by the research team. This system served as a database, but only in an ad-hoc sort of way. Simple storage structures were created – on demand as needed; data storage organization was informal and was allowed to evolve with researcher requirements. From the start, even with the highly non-homogeneous nature of APS data, data files could be found, retrieved, browsed and collated without significant learning time and effort.

TikiWiki as a feature rich CSCW platform. It offered a full kitchen sink of features that could be modularly enabled and customized. Tool modules and features provided in TikiWiki included (in abbreviated listing):

**Collaboration tool modules**: wiki pages, image galleries, file galleries, blogs, discussion forums, real-time text chat, shoutbox, polls, articles, RSS feeds, personal workspaces (with internal messaging, webmail, web bookmarks, ToDo task lists, calendar, notepad and local files storage), and ...

**Features**: customizable user interface menus and page layout, dynamic access to external databases, content templates, search, and ...

To keep usage simple, only a select handful of collaboration tool modules were enabled. Some were experimentally tried and adopted, while most others were quickly abandoned. The modules found to be most essential to the APS users were: wiki pages, file galleries, articles, and personal workspaces.

File galleries were places where collections of data files could be uploaded and made available for others to access. Independent file galleries were created for each distinct data collection event, such as Howard University's year one structured interview audio data. Wiki pages were created to document research status, and regularly updated with researchers' notes that described the evolving state of a data set or a data collection event. Similar to file galleries, wiki pages could also accept data file uploads and appeared as an attachment at the bottom of a page. But the wiki pages' file attachment feature was commonly used to share data analyses files, rather than raw original data. This is likely because the contextual flow of uploading a current analysis data file to a wiki page, while also updating notes describing the state of that analysis, occurs naturally. TikiWiki's articles module provided space on the homepage to post information intended to be readily visible to all users upon login. This tool was used often to efficiently guide users to various file galleries or wiki pages of high interest and value to users at that point in time. The personal workspace module provided file storage space for users to have backed up important work files, and to privately share files with a limited number of specific users.

Whereas most CSCW software platforms provide only a very rudimentary and global permission system, TikiWiki's permissions architecture and implementation supported fine grained permissions. This was critical to the APS research team. It offered the ability to enable specific individual tool functions to any individually specified research user. Additionally, each new instance of a tool module could have a different set of permissions for a given user. These intricate permission options made

possible the means to meet requirements related to IRB compliance for APS' distributed research organization – having researchers, sub-teams and all different types of data collected from and redistributed – across multiple institutions.

The operating functions and corresponding permissions applicable to the wiki module are distinctly different and kept independent from those for the file gallery.

- An example of permission settings: Two File Galleries are created, File Gallery #1 and File Gallery #2. For File Gallery #1, user X is granted permissions to view an index listing of files in File Gallery #1, upload and download files there. User Y is granted permissions to only view an index listing of files in File Gallery #1, but not upload or download any files there. For File Gallery #2, the permissions for user X and user Y are switched, relative to File Gallery #1.
- Another example: Six wiki pages are created, AA, AB, AC, AD, BB, CB. Researcher #1 is granted view access to only one set of wiki pages: AA, AB, AC, and AD. Researcher #2 is granted view access to only wiki pages: AB, BB, CB. Researcher #3 is granted view, edit and file attachment upload permissions for wiki pages: AA, AB, AC, BB, CB.

This sort of specificity in per user custom permissions for each tool module object instantiation throughout the TikiWiki system is incredibly powerful. The existence and ability to modify such fine-grained permissions in TikiWiki was indispensable to APS. Since each tool module typically has six or more permission-governable operating functions, the combinatorial corresponding to permission options per module instance, per data set, and per users, was enormous. It is easy to envision how, with so many permissions control options in certain tool modules, the task of managing permissions

| Wiki objects: | File Gallery objects: |
|---|---|
| • p_wiki_view_attachments | • p_view_file_gallery |
| • p_wiki_attach_files | • p_upload_files |
| • p_wiki_admin_attachments | • p_download_files |
| • p_view | • p_create_file_galleries |
| • p_upload_picture | • p_batch_upload_files |
| • p_rollback | • p_admin_file_galleries |
| • p_rename | |
| • p_remove | |
| • p_minor | |
| • p_lock | |
| • p_edit_structures | |
| • p_edit_dynvar | |
| • p_edit_copyright | |
| • p_edit | |
| • p_admin_wiki | |

***Table 1. TikiWiki permissions for user operations in two different tool modules, wiki pages and file galleries.***

for new users or a new module instances could become tedious rather quickly. Fortunately, TikiWiki supports creation of multi-dimensional, hierarchical mappings between users and sub-team groups. Administrators could also create a set of permission templates that would apply to users tasked to perform certain roles. These and a few other features made many permissions administration tasks more tolerable.

It also turned out that just figuring out the correct mapping of permissions setting to data and its users was non-trivial. It required a deep understanding of the context of the data/information needing to be shared, how these would be used collaboratively, and between which specific researchers. Moreover, given the vast store of data and the diversity in roles and location of APS users (including VIP guests, outside service contractors, undergraduate and graduate students, summer interns, post-docs, and faculty), it was of paramount importance to secure this data from users attempting to gain uninvited access, and also safeguard this data from inadvertent user errors that could result in accidental loss of data. There were no security breaches and no unintended loss of data, over the service life of the APS server.

Another uncommon feature of TikiWiki was its ability to detect and automatically block repeat uploads of previously uploaded files. This meant there would be no duplicate files, only one copy of any file, in the APS database. If users were allowed to change a file name and upload it, it would appear to others as new and perceived to be a different file, with different contents; there would be ambiguity about which file version is the most "valid" for downloading. By not permitting duplicate files, these confusions are avoided.

This feature also helped reduce clutter and filename maintenance. For example, with good maintenance, files already in the APS database are well named and consistent with the prescribed file naming convention. Often, new researchers would download data files; after studying their contents, rename them, and then attempt to upload them to the original or alternate File Gallery area. Occasionally, a researcher would increment the value in the version number field embedded in the filename, for an entire batch of files even though changes were made to only one or two files. When a researcher finds that a file upload has been blocked due to the existence of an identical file, the error notification served as a gentle reminder to be mindful of the data file naming convention. Adherence to good file naming turned out to be very importance to the group's data sharing effectiveness.

As free, open source software, the price for TikiWiki was right. If custom modifications or additional features were requested, having access to its source code meant that code changes could be made. Overall, this software provided the functionality required in serving the CSCW and file sharing needs of the APS researchers for their work. Some under-the-hood administrator-level manual assistance was required, but only on a few rare occasions.

## DATA OWNERSHIP AND SHARING

### Data Access Protocol

The adoption of TikiWiki and its extraordinary affordances of permissions settings prompted discussions amongst the APS leadership about what procedures would be desired to formally document and implement these permissions. Being mindful of data security concerns around jurisdictional responsibility within each institution's IRB, it was desirable to provide limited, non-equal data access permissions to APS research participants. This discussion was grounded in questions around a number of different user types and different research collaboration activity scenarios.

For example: When a new student researcher under the supervision of a principal co-investigator at a particular campus joins a local research team, what access to the APS Database should that student have? What would this student be working on: data collection, data processing, data analysis? The student's work would be in support of which data instrument and which research method? Who should decide the access permissions in a way that is respectful to key stakeholders?

After a lengthy period of thoughtful consideration, an access granting protocol emerged from the leadership team. It would be based on institutional responsibility and research method ownership. In recognizing this duality of roles by principal co-investigators in this protocol, the cross-functional nature of team organization in APS became more explicit.

In the longitudinal study, principal co-investigators at each institution were responsible for safeguarding those data collected from students at their institution, per the guidelines of their IRB-approved protocols. Each principal co-investigator (as an institution's responsible representative) would be able to request and grant for any researcher from their institution, access permissions to data originating from their institution.

To execute the planned research, principal co-investigators (as research method leads) were also responsible for directing and overseeing their method's research activities. They too could request and grant for any researcher, access permissions to data associated with that research method/instrument. As an additional requirement, requests for access had to be sponsored by one of the project's principal co-investigators and include description of how the requested data were to be used.

In accepting this formal data sharing protocol, the principal co-investigators diffused cumbersome entanglements in the planned APS research. Moreover, the development of this protocol laid the foundation for an important subsequent policy on co-authorship requirements for publications based on research and use of APS data.

**Data Use and Publication Policy**

There was anxiety amongst APS researchers, rooted in experience, that less well informed, misguided use of APS data may make certain analysis suspect and conclusions invalid. And if results from these analyses were published, the credibility of the APS data might be harmed. This concern heighted sensitivity to who and how data were shared, and might be reused. The previously defined data access protocol proved to be inadequate in addressing these risks related to research publications being made by researchers after data were shared with them.

A publication policy linked to data sharing and data reuse was developed. It put into place an agreement by all principal co-investigators that research publications using data sets specific to a research instrument or method shall include as co-authors, the principal co-investigator responsible for that data set. Similarly, if the data sets used are institution-centric, the principal co-investigator responsible for that institution's data shall be included as a co-author. Respective co-authors (co-principal investigators) should be engaged in review of the publication before submission, if they were not already an active collaborator in the research itself. This policy addressed researcher concerns by providing means to assure that analyses remain true to APS instruments' design and data processing methods.

With foresight to include consideration for scenarios where data are shared with external collaborators, and sharing that might occur beyond the time scope of APS, the APS leadership developed data access protocols and publication policies that were jointly consistent and compatible. Together, they provided mechanisms to safeguard the credibility and integrity of APS data sets that continue to be shared/reused with a larger community, now and into the future.

**Data's Underlying Tacit Knowledge**

It is easy to overlook how important tacit knowledge is in APS. Notably, within the core team of APS researchers, members communicated frequently and were well informed. There were weekly large group conference calls, in addition to local meetings and emails at higher frequency between sub-team members. Everyone was integrally involved in all aspects of the data instrument designs, and the data collection efforts. These researchers shared in developing an embedded awareness of each set of data's origin and context. There was implicit tacit knowledge sharing.

As the research team became more established, some members transitioned out while new members joined in. The team began to grow with new junior researchers and new collaborators from outside the group requesting access to the APS data collection. By having new members work closely in person with members of the core research team for a period of time, they too become familiar with the design of the research instruments, the collected data sets, and its organization in the APS database. In this engagement between new and initial core researchers, tacit knowledge

(derived from personal experiences and situational contexts surrounding the data) was shared from original researchers to new researchers.

When data is shared with researchers more remote and external to the APS work, the opportunity to share tacit knowledge is limited. How important is this tacit knowledge?

Consider a hypothetical scenario in which a far away researcher becomes interested in comparing student stress levels at a point in time coincidental to a longitudinal survey's deployment. Fortuitously, the survey instrument included a construct that provided a measure for student stress levels. A request is made to access and reuse this data. However unfortunately, even though the original researchers responsible for this survey's deployment discussed at length tradeoffs related to scheduling, little of those deliberations were explicitly documented. And indeed, practical compromises had to be made in the survey's design and deployment schedule. Concerns related to the difference between semester vs. quarter academic schedules across institutions could not be resolved universally. While students at one school were feeling intense stress because final exams were fast approaching, students at another were relaxed returning from Spring Break and starting a new term. This very simple but easily overlooked nuance of timing and schedule-sensitive context can significantly bias subject responses to questions about their stress levels. Then, if a comparative analysis is performed by this far away researcher on this set of cross institution survey data, it would be easy to erroneously conclude that students at one institution were more stressed than the other.

A quotation frequently cited in knowledge management circles and attributed to Socrates: "one does not know what one does not know." Without involving the original APS researchers for input, it would be impossible for the far away researcher to confidently affirm whether a given set of data processing operations were consistent and proper with the data gathering instrument's deployment. Unknowingly, conclusions from such analysis may be misleading and invalid. This scenario exemplifies data sharing and reuse situations where background context around the shared data has not been also provided. There are inherent associated dangers.

The above scenario is only hypothetical because there are no (near or far away) researchers with those specified interests accessing APS data. However, the described decision factors and undocumented considerations pertaining to survey deployment scheduling in this scenario are based in historical reality. APS researchers recognized and carefully weighed tradeoffs in scheduling each deployment of their longitudinal survey. Many other similar scenarios exist in APS – situations where decisions and judgment calls were made, and affected the outcome of processing and state of saved data.

In an ideal world, one would capture as much of this background as possible, bundle it together with the collected data, and then share the bundle as an integrated data package. However, unrecorded in the APS database, these deliberations instead become tacit knowledge that is lost to

outsiders. Capturing tacit knowledge and making it explicit as viewable records is known to be difficult and a costly challenge. This is a universal problem, and often not considered within the plans and scope of most research projects.

Given the longitudinal nature of this research and longevity of data reuse, and resources to thoroughly capture tacit knowledge would still be overwhelming, the APS researchers felt small organized efforts to capture details of significance in real-time would be justifiable. This lead to creation of a continuously updated internal APS "living document" (Sheppard et al, 2010) that recorded basic accountings and key decisions associated with how data were being collected and processed over time.

Additionally, it was recognized that sharing isolated data sets without more thorough background and historical context information beyond the "living document" has heightened associated dangers. This need to better manage the risk and consequences of lost tacit knowledge influenced development of APS' publication policy. Specifically, it included requirement for original researchers to help review and be involved in future use of APS data. This human-to-human collaborative communication imperative re-opens pathways for tacit knowledge sharing between researchers. Absent such a policy, important tacit knowledge about shared data sets would likely be missed. Going forward, implementation of this publication policy should help mitigate related data sharing risks.

As the Academic Pathways Study research was coming to a conclusion, a final review was made to further assure that the aggregate collected data contained no remaining personally identifying information. The goal was to produce data packages that could be comfortably shared with future other researchers. Now post-APS, data have been successfully shared and reused by additional new researchers within and beyond the original research teams, in ways consistent with the APS data access protocol and publication policy. New research publications have resulted.

## REFLECTIVE INSIGHTS

The Academic Pathways Study was a significant, and intensely collaborative research component of NSF's Center for Advancement of Engineering Education activities. The APS team recognize from the start that effective sharing and reuse of data required diligent attention to handle both technology-related and human-oriented challenges in collaboration. Thus well ahead of any data collection, data collection procedures were planned together with detailed design of data organization, as well as the security and sharing access mechanisms for users of the APS online data sharing platform.

Upon reflection, the APS team's ability to do this reliably, share its rich data set and create an effective data reuse ecosystem, was rooted in the team's portfolio of experience around collaborations:

- Team leaders and senior research staff, having participated in large multi-university engineering education collaborations, such as NSF's ECSEL, and Synthesis Coalitions – provided experienced leadership in planning and coordinating the research activity of this large distributed team.

  It was recognized that strong lines of communication were essential. Per plan, people time and resources were invested toward more travel and in person meetings that first year, which provided solid foundation for the substantive work to be done remotely at a distance in subsequent years. A schedule of regular small and large group conference calls was established to coordinate the team's activities and responsibilities around the growing set of research data. Focused phone calls and email correspondence also occurred frequently between team members.
- Several team members, having research experience in design knowledge capture, team collaboration tools for information sharing and reuse, provided guidance toward design, deployment and effective use of infrastructure collaboration tools provided to the geographically distributed teams.

  These team members utilized learnings from prior research experience around technology deployment to help the APS team avoid many common challenges associated with data sharing and reuse. For example, conscientious effort was spent during the start up first year of APS to jointly discuss and plan details of the data collection, sharing, and analysis – roles, responsibilities and processes. Then before data collection began, an all-hands training session was conducted to familiarize everyone with the logistics and processes around the use of the deployed online platform.

## SUMMARY THOUGHTS AND CONCLUSIONS

In the Academic Pathways Study, we were quick to realize that much more is required to enable engineering education researchers at multiple institutions to share data, than simply uploading and downloading files via the internet. The challenges began with coordinating and synchronizing IRB applications across all the respective institutions. With daily increasing cyber security concerns today, additional scrutiny should be anticipated around the design of how data will be secured and jointly shared beyond the IRB jurisdictional boundaries of each institution.

The more diverse and large the research team's composition, the more flexible and fine-grained a permissions system that will needed to adequately secure the data sharing amongst users. Early, broad adoption of an explicit protocol for requesting and granting data access permissions can help avoid many downstream entanglements.

Collected data come in many different forms and types. Data are seldom static. It will have a life of its own; it will grow and evolve over time. Developing a version management scheme for this data, provides a means to trace its life history, will reduce confusion, errors in handling, and will enable more harmonious data sharing.

Data will also have metadata. Many times, a thoughtful file naming schema can be used advantageously to preserve valuable metadata. There are untold benefits to adopting a file naming schema for collaborators who share many data files, and often.

Data will have surrounding background and context. Researchers who have been heavily involved in developing the instrument, collecting, cleaning and processing the data, will know the data's background and context well. Sharing that data without also sharing its associated tacit knowledge can potentially lead to improper use of the data. And that can be dangerous.

Data has currency. When data that is collected with care, handled with care, and processed with care, it gains credibility currency. When data are carefully analyzed, that too extends the data's credibility. However, in this data ecosystem, data sharing can put the data's credibility at risk. How the next user of that shared data goes about further processing and analyzing it, can alter its currency ever more – upward or downward.

Encouraging new users of shared data to engage with the data's past owners can provide beneficial comfort. Incorporating such encouragement through a publication policy that defines requirements for research based on shared data can help mitigate risk of inadvertent data mishandling.

Much learning about data sharing and reuse was gained through our experiences in the Academic Pathways Study.

## ACKNOWLEDGEMENT

The support of the Center for Advancement of Engineering Education and its Academic Pathways Study through NSF Grant ESI-0227558 is gratefully acknowledged.

## REFERENCES

Chen, H., Donaldson, K., Eriş, O., Chachra, D., Lichtenstein, G., Sheppard, S.D., Toye, G., "From PIE to APPLES: The Evolution of a Survey Instrument to Explore Engineering Student Pathways." In *Proceedings of the American Society for Engineering Education Annual Conference* (2008), Pittsburgh, PA, June 22–25, 2008.

Clark, M., Sheppard, S.D., Atman, C.J., Fleming, L., Miller, R., Stevens, R., Streveler, R., Smith, K., "Academic Pathways Study: Processes and Realities." In *Proceedings of the American Society for Engineering Education Annual Conference* (2008), Pittsburgh, PA, June 22–25, 2008.

Atman, C.J., Sheppard, S.D., Turns, J., Adams, R.S., Fleming, L.N., Stevens, R., Streveler, R.A., Smith, K.A., Miller, R.L., Leifer, L.J., Yasuhara, K., Lund, D., "ENABLING ENGINEERING STUDENT SUCCESS: The Final Report for the Center for the Advancement of Engineering Education", Technical Report (2010) #CAEE-TR-10-02, 2010.

Eriş, Ö., Chachra, D., Chen, H., Rosca, C., Ludlow, L., Sheppard, S.D., Donaldson, K., "A Preliminary Analysis of Correlates of Engineering Persistence: Results from a Longitudinal Study." In *Proceedings of the American Society for Engineering Education Annual Conference* (2007), Honolulu, Hawaii, June 24–27, 2007.

Nonaka, I., von Krogh, G., "Perspective – Tacit Knowledge and Knowledge Conversion: Controversy and Advancement in Organizational Knowledge Creation Theory", *Organizational Science* (2009), **20**(3), May-June 2009, pp. 635–652.

Sheppard, S., Atman, C., Fleming, L., Miller, R., Smith, K., Stevens, R., Streveler, R., Clark, M., Loucks-Jaret, T., Lund, D., "An Overview of the Academic Pathways Study: Research Processes and Procedures", Technical Report (2010) #CAEE-TR-09-03 Summer 2003 – Fall 2008, August 2009 revised May 2010

Toye, G., Cutkosky, M.R., Leifer, L.J., Tenenbaum, J.M., Glicksman, J. 1994. "SHARE: A Methodology and Environment for Collaborative Product Development." *Int. J. of Cooperative Information Systems*, Vol 03, Issue 02, June 1994, pp. 129–153.

Baya, V., Gevins, J., Baudin, C., Mabogunje, A., Toye, G., Leifer, L., "An Experimental Study of Design Information Reuse." In *Proceedings of the 4th International Conference on Design Theory and Methodology* (1992), ASME, Sept. 13–16, Scottsdale, Arizona, pp 141–147, 1992.

Leifer, L., Baya, V., Toye, G., Baudin, C., Gevins Underwood, J., "Engineering Design Knowledge Recycling in Near Real-Time." In *Proceedings of Seventh Annual Workshop on Space Operations Applications and Research (SOAR 1993),* Vol. 1, pp 313–320, 1994.

## AUTHORS

**George Toye**, Ph.D., P.E., is consulting professor in Mechanical Engineering at Stanford University. While engaged in teaching project-based engineering design thinking and innovations at the graduate level, he also contributes to research in engineering education, effective team collaboration in concert with internet technologies. As well, he continues to be active as co-founder in startups and in varied consulting work.

George earned his B.S. and M.S. degrees in Mechanical Engineering from U.C. Berkeley, and Ph.D. in Mechanical Engineering with minor in Electrical Engineering from Stanford University. Since 1983, he has volunteered to organize annual regional and state-level Mathcounts competitions to promote mathematics education amongst middle-school aged students.

**Helen L. Chen** is a research scientist in the Designing Education Lab in the Department of Mechanical Engineering and the Director of ePortfolio Initiatives in the Office of the Registrar at Stanford University. She is also a member of the research team in the National Center for Engineering Pathways to Innovation (Epicenter). Helen earned her undergraduate degree from UCLA and her PhD in Communication with a minor in Psychology from Stanford University in 1998. Her current research interests include: 1) engineering and entrepreneurship education; 2) the pedagogy of ePortfolios and reflective practice in higher education; and 3) reimagining the traditional academic transcript.

**Sheri D. Sheppard**, Ph.D., P.E., is professor of Mechanical Engineering at Stanford University. Besides teaching both undergraduate and graduate design and education related classes at Stanford University, she conducts research on engineering education and work-practices, and applied finite element analysis. From 1999-2008 she served as a Senior Scholar at the Carnegie Foundation for the Advancement of Teaching, leading the Foundationís engineering study (as reported in Educating Engineers: Designing for the Future of the Field). In addition, in 2003 Dr. Sheppard was named co-principal investigator on a National Science Foundation (NSF) grant to form the Center for the Advancement of Engineering Education (CAEE), along with faculty at the University of Washington, Colorado School of Mines, and Howard University. More recently (2011) she was named as co-PI of a national NSF innovation center (Epicenter), and leads an NSF program at Stanford on summer research experiences for high school teachers. Her industry experiences includes engineering positions at Detroit's "Big Three:" Ford Motor Company, General Motors Corporation, and Chrysler Corporation. At Stanford she has served a chair of the faculty senate, and as Associate Vice Provost for Graduate Education. She is currently serving as ME Associate Chair of Undergraduate Programs.