Appendix 1-A
APS File Naming Protocol

Data File Naming

Names of data files within the APS database as proposed here have 6 standard components (plus one optional component) organized in the following order:

1. StudentID: coded per CAEEID
2. MethodType: instrument or method
3. EventID: event_sequence or event_date, or combination
4. ItemID: item_type, item_number and revision_number
5. ResearcherID: name initials
6. (optional) Pseudonym: reference subject's "name"
7. FilenameExtension: document_type

To improve readability and facilitate accurate computer-based parsing, filename components are separated by hyphens. By design, the resultant filename will uniquely identify the context of each data file in the APS database. This filenaming convention would result in filenames that look like

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

This document is a work in progress and may be revised to reflect new needs and functions of the APS database. It is important that this proposed filenaming protocol not only meet the immediate needs related to the current study of the Longitudinal Cohort (Cohort 1), but can also be carried forward to integrate data for future APS cohort studies.

StudentID

In the APS database, we must avoid identifying our study's student participants by real names or other recognizable real-world information, in association with collected data. To abstract a participant's identity, we have developed a coding scheme. This CAEE Student ID code uniquely identifies each student participant and can be broken down into 4 parts. It looks something like: "TPub01F00003". The first part, "TPub", is the school's official acronym (i.e., Technical Public Institution). The second part, "01", is the cohort id. The third part, "F", refers to the gender, female. The fourth and final part, "00003", is a sequentially generated number identifier of the student at the indicated school.

For cohort-1, we have TPub, UPri, SPri, and LPub as the 4 possible school acronyms. When expanding the study in cohort-3, and cohort-4, most U.S. schools have unique acronyms. If we should encounter two or more schools with identical acronyms, we can append a lower case letter (a, b, c...) in sequence to differentiate these schools. Internal to the APS database, there is a table that connects this acronym (known as UnivID) to the full name of the school and its related descriptive data (e.g., university's full name, semester/quarter system, etc.).

With this StudentID code displayed in the filename, a researcher can quickly tell at a glance that the data file is associated with a specific student who attends a given school, is of a given gender, and participates as a member of a specific cohort.

If student data were aggregated into container documents by school, the containers' filenames would only include the school acronym and the cohort id (e.g., "TPub01"). No gender and student identifying sequence number would be included.

MethodType

The MethodType component in the filename identifies the data instrument or method being used to collect the data contained in the file. This portion of the filename is typically 3-4 letters long. For our research as planned, the following MethodTypes would be used:

SURV	survey data
INTS	structured interview data,
INTE	ethnographic interview data
INTX	exit interview data
INSP	problem scoping exercise data within structured/ethno interview session
ETD?	engineering thinking and doing data
ACTX	academic transcript data
ETH?	ethnography data

EventID

The EventID is used to identify the particular instance of the data collection event. The EventID, taken together with MethodType, allow us to refer to a specific data set in a sequence, such as Survey 2, or Structured Interview 1. We will use numeric digits such as {1, 2, 3} to specify the EventID.

Ethnographers will typically contribute new files on a regular basis throughout the year. It may be more appropriate for such research methods to use date in lieu of a sequence number to represent the data gathering event. When applied, the date would be formatted as "YYMMDD", so as to facilitate chronological sorting. This date information should not substitute for the inclusion of more detailed date information inside the file document itself.

ItemID

The ItemID is used to identify one or more data items collected together within the context of a single data collection event. It is composed of three parts, in order:

1. DataType,
2. ItemNumber, and
3. RevisionNumber.

For example, in the course of structured interview #1, we may produce one audio recording file, one interview notes document, and 2 PDF scan files. In this scenario, we would have 4 files with

MethodType = INTS
EventID=1

and the following distinguishing ItemIDs

ItemID=A1_1 (Audio File)
ItemID=N1_1 (Notes File)
ItemID=S1_1 (Scan File #1)
ItemID=S2_1 (Scan File #2)

If after review, the notes file with ItemID=N1_1was revised and resubmitted to the database, the revised file would take on ItemID=N1_2.

At the current time, the following DataType codes are proposed:

A	audio recordings
V	video recordings
N	notes (field notes, interview notes)
T	text transcriptions
R	transcript revision requests
E	MS Excel data
X	tab delimited columnar data
C	comma delimited data
H	Survey rendered from HTML
S	SPSS data analysis
P	digital photos
Z	scanned paper documents

ResearcherID

The ResearcherID identifies the researcher who is primarily responsible for collecting the data in the file. The researcher's initials (in all capital letters) will be used. If we should encounter a situation in which a new researcher has initials identical to an existing researcher, we would append a number to the new researcher’s initials for the ResearcherID. For example, if we have a researcher Gwendelyn Talbot and we add a new researcher Greg Taylor, Gwendelyn Talbot would have ResearcherID “GT” and Greg Taylor would be assigned “GT1”.

Pseudonym

Ethnographers will typically contribute new files on a regular basis throughout the year. In such cases, the CAEE student ID may be hard to write and refer to in discussion. The student pseudonym, appended to the root part of the document name and also recorded in the APS database, would be the identifier that ethnography researchers would use to refer to the student subject.

Student pseudonyms would be created by the research teams at each university, and may be something like "Lego1". Each pseudonym is unique within the subject group at each school; it is not permitted to have another "Lego1" within that school’s subject group. However, it is entirely acceptable to have "Lego1" at another school. This way, researchers have full autonomy and flexibility to choose pseudonyms without fear of conflicts with those chosen by researchers at other universities.

Pseudonym, this filename component, is optional and is not likely used outside of ethnography documents.

FilenameExtension

We use familiar filename extensions to identify their respective applications and data types. The following example filename extensions would be used with associated applications:

.rtf	Microsoft Word, Mac TextEdit
.doc	Microsoft Word
.dot	Microsoft Word (Template)
.xls	Microsoft Excel (spreadsheet)
.csv	Microsoft Excel (comma separated values)
.sav	SPSS data file
.sps	SPSS variable definitions/label file
.dss	Olympus Audio Recorder/Player
.pdf	Adobe Acrobat Reader, Mac Preview
.txt	Microsoft Notepad, Mac TextEdit
.zip	WinZip, Windows Explorer (XP), Stuffit Expander
.jpg	Internet Explorer, Netscape, Mozilla, Safari
.tif	(tagged image format)