

Data Curation Profile – Biomechanics Motion Studies (Kinesiology)

Profile Author	M. Cragin
Profile Author	M. Kogan
Profile Author	A. Collie
Institution Name	Illinois
Contact	M. Cragin (cragin@illinois.edu)
Date of Creation	12/31/2010
Date of Last Update	n/a
Version	1.0
Discipline / Sub-Discipline	Biomechanics
Purpose	<p>Data Curation Profiles are designed to capture requirements for specific data generated by a single scientist or scholar as articulated by the scientist him or herself. They are also intended to enable librarians and others to make informed decisions in working with data of this form, from this research area or sub-discipline.</p> <p>Data Curation Profiles employ a standardized set of fields to enable comparison; however, they are designed to be flexible enough for use in any domain or discipline.</p>
Context	A profile is based on the scientist's reported needs and preferences for these data. They may be derived from several kinds of information, including interview and document data, disciplinary materials, and standards documentation.
Sources of Information used for this profile	<ul style="list-style-type: none"> • An initial interview with the scientist (April, 2008) • A follow-up interview with the same scientist (January, 2009) • The Requirements Worksheet questionnaire was completed during the follow-up interview (January, 2009)
Scope Note	The scope of individual profiles will vary, based on the author's and participating researcher's background, experiences, and knowledge, as well as the materials available for analysis.
Editorial Note	Any modifications of this document will be subject to version control, and annotations require a minimum of creator name, data, and identification of related source documents.
Author's Note	This Motion Studies Data Curation Profile is based on analysis of interview and document data, collected from a researcher working in this research area or sub-discipline. Some sub-sections of the profile were left blank; this occurs when there was no relevant data in the interview or available documents used to construct this profile.
URL	http://www.datacurationprofiles.org

Brief summary of data curation needs

This scientist needs infrastructure to accommodate large amounts of data as these projects generate gigabytes of raw and “filtered” data, and megabytes of reduced and abstracted data that are held in several forms. Beyond the raw, proprietary quantitative data, the processed data types include Matlab files, MS Excel files, codebook texts, and graphical files. The data most suitable for deposit into a shared repository is a collection of spreadsheets, along with the code book (which contains details of individual trials in the data set). The codebook file(s) will require additional preparation before submission, to remove personal identifiers.

With respect to value over time, these data ought to be maintained for a minimum of five (5) years, but less than ten (10) years. Data generated during movement studies with people from special populations (e.g. post-surgery; have a diagnosed movement disorder) are seen to have high value for re-use because of the very high cost of replacing the data. That is, developing a new, specialized data collection would require both great human and financial resources.) The scientist does not see the need for an embargo period for these data. Access restrictions for the data were suggested based on ethical concerns, such as the possibility of correlating movement characteristics with personal characteristics. The value of these data for use by the general public is seen to be very low or none.

Overview of the research

Research area focus

This scientist works across several different research fields, including kinesiology, biomechanics, developmental psychology and community health. He conducts research on movement and balance; the project reported here concentrates on movement analysis of various groups of human subjects that span age (lifespan) and special populations with injuries or disabilities.

Intended audiences

- Scientist’s interdisciplinary group of campus collaborators
- General population of kinesiology researchers
- Researchers in other disciplines pertaining to biomechanics, which is a very interdisciplinary field

Funding sources

The NSF and NIH are the primary funders for this scientist’s research.

Data kinds and stages

Data narrative

These data are produced with a motion capture system, which includes hardware and proprietary software. Motion capture markers are attached to various parts of the body, usually the joints. While the study subject performs target motions, the marker coordinates are recorded by the motion capture system. The precise placement of markers is very important for the quality of the data and its reliability. The scientist normally uses about 40 markers on each subject. These motion capture systems are produced by several companies; currently there are not any standard for this sub-discipline. This scientist uses a specialized motion capture system.

The 3-D marker coordinates (x, y, z) are captured multiple times (t_x) in a session; this raw output is proprietary, and some bit of automatic filtering happens within the motion capture system. The data are then moved to Excel for automated and human “filtering,” to removing errors and noise, which occur due to the system being sensitive to light (e.g. reflections) and marker occlusion; this filtering process is tedious and time-consuming work.

In order to more directly deal with the raw data, researcher transfers it into an Excel spreadsheet. After that more automatic threshold-based filtering is carried out, along with visual review of the data and manual cleaning. This process takes place in either Excel or Matlab. The resulting raw “filtered” data is usually kept in either Matlab or Excel files.

The filtered raw data set is best shared as a collection of spreadsheets. Each spreadsheet represents the recorded trials of a single study subject. For these research projects, these data are moved into MatLab, and converted to represent several variables (e.g. angle data, displacement velocity, or acceleration of joint segments). Data are then aggregated across subjects. This aggregate data is also stored in an Excel spreadsheet. In the following stage it is further reduced, either to statistical values (means and standard deviations), or to wave patterns, depending on question of interest. This data is also maintained in an Excel spreadsheet form.

One analytical process results in wave patterns, which are a representation of the study subjects’ movement with respect to some parameter. The wave patterns are stored as coordinates that can be easily plotted into a visual representation; these are often published in peer reviewed papers.

The categories in the “data stages” column listed in the table below were developed by the authors of this data curation profile. The data specifically designated by the scientist to make publicly available are indicated by the rows shaded in gray.

Data Stage	Output	Typical File Size	Format	Other / Notes
Movement Data				
“Raw”	Coordinate data over time	200 files X 20-100Mb each	Proprietary format, then MS Excel	40 markers, about 2 min/ trial, 240-1000 Hz sampling rate
“Raw filtered”	Filtered coordinate data over time	200 files X 20-100Mb each	MS Excel or Matlab	40 markers, about 2 min/ trial, 240-1000 Hz sampling rate
Processed, aggregate data	1. Angles coordinates 2. Velocity profiles 3. Acceleration profiles	Smaller than previous stages (10s-100 Mb?)	MS Excel or Matlab	
“Reduced”	numerical data	1. Few Mbs 2. Kbs	MS Excel	These are means and standard deviations for the angles coordinates, velocity profiles and acceleration profiles.
Analytical product	Wave patterns: quantitative measures differentiating elliptical shapes	Small, about 18 rows	MS Excel, and graphics (format unknown)	
Augmentative Data				
Codebook				May indicate personally identifiable information

Note: The data specifically designated by the scientist to make publicly available are indicated by the rows shaded in gray. Empty cells represent cases in which information was not collected or the scientist could not provide a response.

Target data for sharing

The filtered raw data a set of approximately 200 spreadsheets is perceived to have the most informational value for re-use as all the other analytical measures can be derived from it. (It was noted, however, that some researchers in this sub-discipline want the original raw data in order to apply their own filtering processes.) The codebook file enhances greatly the usability of the raw filtered data as it provides background information on the pool of study subjects, including demographic and physiological characteristics. As noted above, personally identifying content would need to be removed prior to deposit.

Use/Re-use value of the data

Two different conditions adhere for the maintenance of this type of data for re-use: The value for data sets derived from studies of typically developed and developing subjects is tied to technology cycles. The motion-capture research field is technologically intensive, and new methods for generating, processing and analyzing data are in constant development, with new data collection and analysis systems developed approximately every five (5) years. At that point it becomes easier to collect new data, but more difficult to use the existing data. Data sets that are collected from study subjects who are typically developing/developed (and without injury) have valuable only as long as the technology used for its collection remains viable and accessible.

However, data sets that are seen to have high value for preservation and re-use are those generated during movement studies of special populations (e.g. subjects who have particular medical conditions like post-injury/post-surgery, or a diagnosed movement disorder). First, the cost of replacing these data are very high - it is difficult to recruit study subjects from special populations, and developing these data sets requires extensive human and financial resources. As a result, these data also have high value for local re-use, and use by others interested in analyzing them. The scientist noted that these data sets ought to be kept for a much longer period of time than the data collected on typically developing or functioning subject populations, though he did not state a time frame.

Contextual narrative

This research is data intensive, but also highly dependent on the expertise of the research staff, who are quite involved in the set-up of the trials. Motion capture systems generate large amounts of data; each participant may perform 10-40 trials some task (like walking on a treadmill) for approximately two (2) minutes each at a rate of ~1000 Hz (cycles/second). The data that are generated includes coordinate data (x, y, z) are collected over the trial (time) for about 40 different points using markers attached to the person. The sampling rates and number of data points captured per second are important. It was also stated that, "where the markers are placed on the body is incredibly important; and knowing that the individuals who place the markers are trained is kind of important."

For general data management, a separate spreadsheet is produced for each participant. This data set is considered to be static; while individual subject data can be added to the aggregate set, the data generated from each individual study subject's motion study is finite. Data may be processed in different ways depending questions being addressed in the study, resulting in (joint) angle data, velocity profiles, or acceleration profiles.

Data integration or re-use is a problem in this field, as gait labs and clinics do not use the same methods for collecting data, thus, "the data from various gait clinics doesn't really map on well to the data collected to in other gait clinics." This scientist's concern is not with the data per se, but the lack of standards for data collection procedures, and he sees this as having a higher priority for the field than standards for the "maintenance and keeping of the data."

Intellectual property context and information

Data owner(s)

For the interdisciplinary project that was the main focus of these interviews, the scientist states that he and his collaborators own this data collectively. Equal access is given to each of the project's group members, and the group holds regular meetings to talk about research questions and data analysis. The intellectual property seems to be a non-issue for the collaborators driving this project.

He noted, however, that for another project where each collaborator collected data in their individual labs, sharing is not automatic and use has to be negotiated for each instance. There is no standard arrangement or written agreement to guide such negotiations.

Stakeholders

The primary stakeholders are the project collaborators, and the funding agencies - NSF and NIH.

Terms of use (conditions for access and re-use)

The scientist has some general concern about access to his data by the broad public, and additional concerns about potential inappropriate use of this type of data (for example to create gait profiles are tied to the identity of specific individuals). Thus, he would like to maintain some restrictions on access. There is more on this in the Access Control section below.

Attribution

The scientist did not mention whether receiving attribution from others who use his data was important for him; however, he was very willing to provide attribution and even (offer) co-authorship to one set of colleagues whose data he used.

Organization and description of data for ingest (incl. metadata)

Overview of data organization and description

Most of the data collected by the scientist is stored in the Excel format. Time sequence data (including the raw, filtered, and angle data) is listed by rows that indicate the incremental data points during data collection. The columns represent marker coordinates or variables derived from those coordinates. The file collection is organized by the file names and the folder structure, while also cross-referenced through a "codebook" file.

Formal standards used

There are currently no formal standards (including metadata, vocabularies, or ontologies) in this field.

Locally developed standards

None.

Crosswalks

None.

Documentation of data organization/description

The filtered raw data is organized through a file folder system, where each trial is documented in a single spreadsheet, and all the files from particular study are stored in the same folder structure. Another level of organization is added by the codebook file, which connects and cross-references data files and study subjects through the use of the subject ID. The subject ID follows certain naming conventions, indicating both the study name and participant number.

Ingest

The scientist indicated that the ability to personally submit the data into a repository was of high importance, as was having the submission process be automated to some extent.

Access

Willingness / Motivations to share

The scientist is willing to share raw, cleaned, and processed data with his immediate and close collaborators on that project; he definitely gives preference to people he knows and trusts. Otherwise, he stated, "I really wouldn't be willing to share the data set until I've run an initial analysis on it." While is willing to share his data with other known colleagues before publication on a case-by-case basis, and this depends on the reasons why other researchers need his data, what projects they hope to undertake based on the data. He would be open to sharing the analyzed data with the people at his research center (department).

Immediately before and after publication, the scientist would share within the research institution, and his professional societies; he indicated reluctance to share beyond these professional boundaries, and, again, said it would really depend on the request.

The scientist has some reservations about publically sharing the data, due to potential for unethical use. As noted above, such possible misuse includes constructing individual gait profiles, indentifying individuals based on those, and possibly correlating the gait characteristics with mental and moral traits. He would prefer, therefore, that data access to be mainly limited to the professionals in his specific field. However, he also views scholarly publication of the results derived from the data as sharing with the public. And, despite his reservations on making data widely available, the scientist believes that after the findings have been published, the public is entitled to access the data if desired.

Embargo

The scientist does not see a need for embargo on this kind of data, since the processing period is so extensive, he noted that this acts as a sort of embargo.

Access control

The ability to restrict access was indicated to be a high priority for the scientist. If the data were to be deposited before publication, the scientist would like to have restrictions on access. While he used the term 'data warehouse, much of the discussion pertained to the possibilities of a domain-based repository that would exist for this type of data. He suggested that there should be an advisory board, strong rules and standards regulating both access and deposit, and data access should be restricted to "practitioners in the field." Given these sentiments, the scientist also acknowledged that, "most journals and most professional organizations require that we share the data with anybody that asks to see it."

Secondary (Mirror) site

The ability to access the data set at a secondary (mirror) site if the repository goes off-line is a low priority.

Discovery

The scientist indicated that it is very important for researchers in this field to be able to find this data. Enabling researchers from outside of her field, and enabling discovery through internet search engines were both low-priority concerns.

Tools

The researcher indicated that the ability to connect the data to visualization and analytical tools is a high priority.

Interoperability

The scientist indicated that support for the use of web services APIs is a low priority. There was no mention of a need to make data sets interoperable, or to a need to be able to aggregate data sets.

Measuring impact

The scientist did not specifically discuss a need to measure the impact of making his data available to others.

Usage Statistics

Usage statistics were a low priority for this scientist.

Gathering information about users

The scientist did not specifically discuss gathering information about users.

Data management

Security/Back-ups

Currently, these data are stored on a grant-funded server in the lab of one of the project collaborators (in a different department). The costs of back-up, upkeep and data migration to new media are covered from the grant as well. The data is backed up and maintained by the Engineering lab. Various parts of the data are also stored on the local machines of the collaborators.

On the local machines in the participating scientist's lab, the data management is performed by the graduate students who spend about ten hours a week on those tasks. The data is backed up daily.

Secondary storage sites

A secondary storage site is a low priority for the scientist, though he did recognize the risk of having the back-up copies in the same location as the 'original.'

Preservation

Duration of preservation

In the interview the scientist indicated that he tries to keep all his data indefinitely. This preservation attitude is due to the high cost of producing and filtering raw data. However, the scientist mentioned that these data get less useful as technology used to collect it becomes obsolete.

As noted above, the filtered raw data are the most informationally useful data for sharing, and ought to be kept for at least five (5) years, but less than ten (10).

The data on the special populations are much harder to recreate, carry more value and should be preserved longer, despite the technological change.

Data provenance

Documentation of any and all changes made to these data sets is a high priority for the scientist.

Data audits

The ability to audit the data is a high priority for the scientist.

Version control

Version control is not applicable to this dataset.

Format migration

The ability to migrate datasets into new formats over time is a high priority for the scientist, and he noted that this is done now for current data sets in the lab that handles the collection and management.

Personnel – This section is to be used to document roles and responsibilities of the people involved in the stewardship of this data. For this particular profile, information was gathered as a part of a study directed by human subject guidelines and therefore we are not able to populate the fields in this section.

Primary data contact (data author or designate) - The Scientist

Data Steward (ex. Library / Archive personnel)

Campus IT contact

Other Contacts

Notes on Personnel