Me'phaa Language Documentation Project Meta-Data guide

Hugh Paterson III
hugh.paterson@sil.org

"...certain things being stated, something other than what is stated follows of necessity from their being so." - Aristotle [Organon]

The purpose of this paper is to describe the meta-data schema used by the Me'phaa language documentation project.This would include how the meta-data is collected, where it is stored, the collection processes, conventions, and the challenges faced in the acquisition of meta-data as it relates to this project. This paper is not a presentation of the corpus or of the methodologies used to create the corpus, it presupposes upon a readers knowledge of both of these. However, a distinction is made between workflow and methodologies. Workflow is the various stages in collection of language data and metadata, whereas methodologies are the manner of things done at each stage. It is necessary to understand several issues related to workflow to understand some of the needed flexibility in such things as the *File Naming Convention.* Where possible and applicable this paper presents what is understood to be best practices for meta-data collection for Language Documentation and presents how we have interpreted best practice, our attempt to implement those practices, or our rejection of those practices and our rational for doing so.

1. Introduction to the Project Data Structures:

A. Project:

The entire project corpus consists of whatever is done.

> There is no principled research plan explaining or proposing what will be attempted from either an anthropological framework or from a linguistic perspective. There is also no developed plan to guide documenters in the development of a multi-dimensional corpus (Himmelmann 1998, particularly section 4.2). The project's efforts are motivated by what is interesting to the primary investigator. However, whatever is archived is what will be declared to have been done. The Data which is developed in the process of pursuing various interests will live in the Project Data Folder. This Data must needs be described by Meta-data.

B. Categories:

Everything which is done in this project can likely be related to something, or multiple somethings. i.e. There might be top level events and then also *Grammar Exploration*. Under Grammar Exploration there might also be *Verb Conjugations*.

Name	Date Modified	Date Created	Size
Metadata_Index-Event-v8.xls	11:36 AM	4/5/11	14.3 MB
E026 – Emilia plant Texts	11:21 AM	9:52 AM	14.1 MB
E020 – Interview in Nanzintla	9:07 AM	2/16/11	5.05 GB
Unprocessed Videos	8:34 AM	12/27/10	29.49 GB
L021 – The 183 Word Wordlist	4/13/11	3/21/11	2 GB
E010 – Filming of Coffee Text in the Coffee field and Animals that eat Coffee	4/13/11	1/3/11	5.63 GB
E009 – Mixed Questions	4/13/11	12/29/10	748 MB
E013 – Tongue twisters	4/13/11	11/29/10	3.9 MB
E008 – Bird Texts	4/13/11	----	834.9 MB
E011 – Filming of the Making of Tortillas	4/13/11	1/3/11	1.89 GB
L022 – The 200 Word Wordlist	4/13/11	3/21/11	44.48 GB
Grammar Exploration	4/13/11	1/7/11	1.26 GB
L002 – Vocatives	4/13/11	12/28/10	257.9 MB
L010 – Locatives	3/28/11	12/30/10	348.5 MB
!L034 – Estar Solo PAST	3/25/11	3/25/11	83.1 MB
!L033 – Estar Cansado PAST	3/25/11	3/25/11	78 MB
Verb Conjugations	3/25/11	1/12/11	868 KB
!L018 – Relative Locations	3/14/11	1/22/11	494.7 MB
Adjectives	3/14/11	1/22/11	53 KB
!L016 – Yes No Questions	1/22/11	1/22/11	66 KB
!L015 – General Questions	1/22/11	1/22/11	66 KB
Unprocessed Audio Files	4/13/11	12/31/10	4.62 GB
L043 – Town Names in Mephaa	4/13/11	4/10/11	559.3 MB
E024 – Filadelfia Texts	4/13/11	4/10/11	5.46 GB
E025 – Elias Funny Story	4/13/11	4/13/11	447.5 MB

Categories have a one-to-many relationship with events, and potentially a many-to-one relationship with events but the event folders must live in only one location. This means that location of an event folder is in a Category folder. This is so that our event files which are numerically ordered can be grouped together topically.

C. File structure:

Project Data Folder >> Category folders >> Event Folders >> Files

For pragmatic reasons Files are located in Event folders. Event Folders are located in Category folders. And the Category folders are located in the Project Data Folder.

(The researchers did not want to view many event folders at one time in one folder, and it makes some organizational sense to organize files with non-sequential EventIDs to each other by placing them in the same category.)

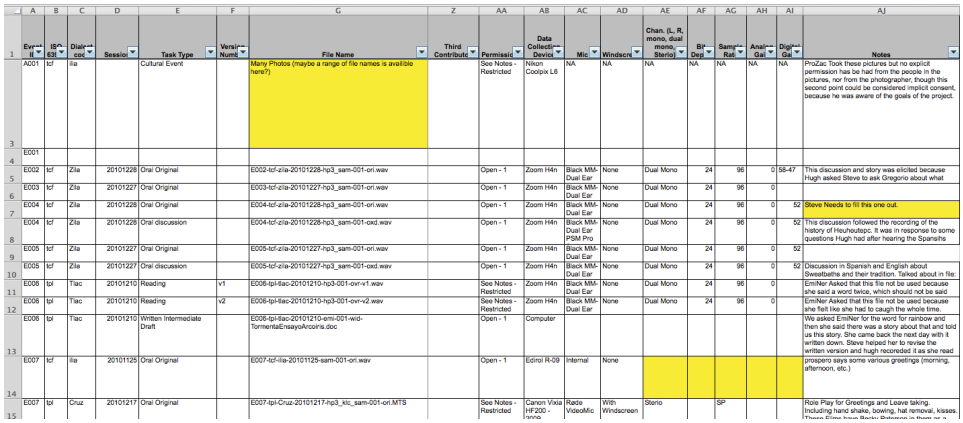
Paterson III Hugh Joseph Jan 15, '11, 11:20 PM
This needs some work.

2. Meta-data tracking:

The goal of this section is to describe the meta-data storage, visualization and organization solution.

Each metadata value is organized in an Meta-data worksheet (Excel spreadsheet). This choice of technology to implement data storage, visualization, meta-data organization and meta-data

collection has its benefits and costs.



	A	B	C	D	E	F	G	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ
1	Event ID	ISO-639	Dialect code	Session	Task Type	Version Number	File Name	Third Contributor	Permissions	Data Collection Device	Mic	Windscreen	Chan. (L, R, mono, dual mono, Sterio)	Bit Depth	Sample Rate	Analog Gain	Digital Gain	Notes
3	A001	tcf	Ilia		Cultural Event		Many Photos (maybe a range of file names is availible here?)		See Notes - Restricted	Nikon Coolpix L6	NA	NA	NA	NA	NA	NA	NA	ProZac Took these pictures but no explicit permission has be had from the people in the pictures, nor from the photographer, though this second point could be considered implicit consent, because he was aware of the goals of the project.
4	E001																	
5	E002	tcf	Zila	20101228	Oral Original		E002-tcf-zila-20101228-hp3_sam-001-ori.wav		Open - 1	Zoom H4n	Black MM-Dual Ear	None	Dual Mono	24	96	0	58-47	This discussion and story was elicited because Hugh asked Steve to ask Gregorio about what
6	E003	tcf	Zila	20101227	Oral Original		E003-tcf-zila-20101227-hp3_sam-001-ori.wav		Open - 1	Zoom H4n	Black MM-Dual Ear	None	Dual Mono	24	96	0		
7	E004	tcf	Zila	20101228	Oral Original		E004-tcf-zila-20101228-hp3_sam-001-ori.wav		Open - 1	Zoom H4n	Black MM-Dual Ear	None	Dual Mono	24	96	0	52	Steve Needs to fill this one out.
8	E004	tcf	Zila	20101228	Oral discussion		E004-tcf-zila-20101228-hp3_sam-001-oxd.wav		Open - 1	Zoom H4n	Black MM-Dual Ear PSM Pro	None	Dual Mono	24	96	0	52	This discussion followed the recording of the history of Heuhueteqc. It was in response to some questions Hugh had after hearing the Sparsihs
9	E005	tcf	Zila	20101227	Oral Original		E005-tcf-zila-20101227-hp3_sam-001-ori.wav		Open - 1	Zoom H4n	Black MM-Dual Ear	None	Dual Mono	24	96	0	52	
10	E005	tcf	Zila	20101227	Oral discussion		E005-tcf-zila-20101227-hp3_sam-001-oxd.wav		Open - 1	Zoom H4n	Black MM-Dual Ear	None	Dual Mono	24	96	0	52	Discussion in Spanish and English about Sweatbaths and their tradition. Talked about in file:
11	E006	tpl	Tlac	20101210	Reading	v1	E006-tpl-tlac-20101210-hp3-001-ovr-v1.wav		See Notes - Restricted	Zoom H4n	Black MM-Dual Ear	None	Dual Mono	24	96	0		EmiNer Asked that this file not be used because she said a word twice, which should not be said
12	E006	tpl	Tlac	20101210	Reading	v2	E006-tpl-tlac-20101210-hp3-001-ovr-v2.wav		See Notes - Restricted	Zoom H4n	Black MM-Dual Ear	None	Dual Mono	24	96	0		EmiNer Asked that this file not be used because she felt like she had to caugh the whole time.
13	E006	tpl	Tlac	20101210	Written Intermediate Draft		E006-tpl-tlac-20101210-emi-001-wid-TormentaEnsayoArcoiris.doc		Open - 1	Computer								We asked EmiNer for the word for rainbow and then she said there was a story about that and told us this story. She came back the next day with it written down. Steve helped her to revise the written version and hugh recorded it as she read
14	E007	tcf	Ilia	20101125	Oral Original		E007-tcf-Ilia-20101125-sam-001-ori.wav		Open - 1	Edirol R-09	Internal	None						prospero says some various greetings (morning, afternoon, etc.)
15	E007	tpl	Cruz	20101217	Oral Original		E007-tpl-Cruz-20101217-hp3_kic_sam-001-ori.MTS		See Notes - Restricted	Canon Vixia HF200 - NTSC	Rode VideoMic	With Windscreen	Sterio		SP			Role Play for Greetings and Leave taking. Including hand shake, bowing, hat removal, kisses.

The benefits:

I. By using Excel as a team we had a document type which was accessible by everyone on the team.

II. The file was a flat file, as apposed to a multi-demential database structure. This helped people understand where data belonged in relationship to the file.

III. The data can be exported to other programs via CSV format.

IV. Depending on the version of Excel used to access the meta-data file there can be powerful filters to find the data entries desired.

V. Some drop down lists could be created for the selection of meta-data in some fields.

The cost:

I. Up to 108 bits of meta-data per file created. This is a lot of time to input meta-data.

II. Meta-data Fields which can be assumed based on the structure of the meta-data still had to be inputed individually.

III. Only one person could write to the Excel file at a time.

IV. Permissions on the file were getting corrupted, affecting access to the file by team members. This resulted in having to version the meta-data file.

V. Some fields were not relevant to all files. This resulted in some confusion because there were sometimes “holes” in the chart.

VI. Visualization of relationships expressed in the meta-data did not make themselves available for input in to the analysis of the primary data.

What should a meta-data tracking solution look like?

During the course of the project it became clear that tracking meta-data was an extra burden (overhead) rather than something which actually added new insight to the analysis, or even energy to the project. This is partially because the data (even the files containing the data) was not always organized in a helpful way [1]. Another reason was that files could not be sorted and arranged based on meta-data values associates with the files. If the meta-data could be presented back to the inputing user in a manner which adds new insight to the analysis, the the user is probably going to make the decision that the pay-off in time and effort to input the meta-data is worth the time it takes to collect their data.

With a workflow influenced by a language documentation view of a project, it follows that files need to be organized so that they can be submitted to the archive. However, this forces the organizers of the files into a paradigm of viewing their files (this paradigm might be provided by the archive or it might be system of organization created just to survive the onslaught of data during the project). However, the data and meta-data are not without relationships; relationships to each other, to the participants, to the local geography, to other languages, to project plans and objectives, to linguistic theory, to community interests (just to name a few). A preformed paradigm of organization will not respect all of these relationships. Each of these relationships show the research team something useful, which during analysis can affect the documentation project. The interface to the data needs to be able to help the project participants view and connect with the data in a relevant, intuitive, informative, collaborative, progressive, time efficient and relational ways. In a way this is a data visualization issue.[2] Displaying the complexities of the relationships in the meta-data and the data is something that the file system of modern Operating Systems (OS X, Windows, Linux) can not deliver. Something else that the file system can not deliver is an organization by “custom” meta-data attributes. Yet one more limitation of the files system is that it does not allow for the re-use of metadata (implied meta-data based on associations of files). I can not drop a file in a folder and then know that the speaker on that file was “xyz” because the other two files in the folder both had the speaker “xyz”.

Must Haves – Features

* The ability to implement a solution across Windows, OS X and Linux platforms.

---

[1] “Helpful” in this context should be explored a bit further than just acknowledged on the surface level. There were several schools of thoughts among the participants, and there were different kinds of objectives for each of the participants. This was reflected in the way we worked and what we worked on. Our needs (so things can be “helpful”) also depended on if our approaches were event oriented or product oriented. (I am doing a paper on place names so I need to collect place names… as an example of product oriented. This is contrasted with “I am recording a story and a place name comes up”… as an example of event oriented.) Organizing all the recordings containing place names would be one way to organize the files, another way to organize the files would be to mark each event which contains a place name but to organize all the files related to one recording event together. i.e. The audio file, the video file, the transcription, the translation, the geo coordinates relevant to the recording, etc. The next logical question, if the organization by event option is selected, is to choose how to organize the events.

[2] There are many kinds of data visualization techniques and tools. Data visualizations are really important to the way we think and understand things in a time-space continuum (Item “a” was in location “x” at time “y”, while item “b” was in location “w” at time “y”). Visualizations are also really good for communicating things which are removed from our local cognitive environment (like the massive interrelatedness of the internet) and abstract concepts like workflows and conceptual categories (put the data in this bucket or that bucket). They are also good at demonstrating the relationship between two (or more) concepts (These are sometimes called mash-ups.)

* The ability to use the same solution on a personal computer and on a server on a local network, or a Wide Area Network (web hosted and a single storage solution for data files).

* The ability to have a scalable solution (in terms of users accessing the solution and in terms of the amount of Gigabytes of data it can hold).

* The solution, if browser based had to be IE, FF, Safari, Chrome compliant (CSS2 minimally, even CSS3).

* The ability to leverage good UI design on the final product to enable end users to quickly adapt to concepts presented by the work and to intuitively know how to use the tool.

* The install process of the final product needed to be simple enough to allow for wide spread adaption with a low impact on needed support for customizations.

* The ability to share information among project users (researchers and contributors, language consultants) in a collaborative environment.

* The ability to share information with the greater linguistic community via established norms.

In addition to these requirements there are certain technologies which are prevalent to the field of Language Documentation. These range in scope from file types, API's, and data sets, to research frameworks. Also to be taken into account are other technological tools used in a language documentation project. i.e. FLEx, ELAN, Toolbox, Tools from the Max Plank Institute. Methodologies from both Language Documentation and Language Description (Descriptive Linguistics) need to be accounted for, as projects, depending on their purpose, staffing and desired impacts are likely to use methodologies from both sub-fields of Linguistics. These research requirements also needed to be accounted for in a systematic way. The method used to account for these was to walk through a Language Documentation project to experience first hand the work-place data handling requirements.

In a way there are three levels of requirements:

1. There are organizational and processing requirements like being able to extract from and write meta-data to file types, view files on a map related to the location they represent or the type of document they are.
2. There are functional requirements .i.e. UX requirements like drag-and-drop file additions.
3. There are design requirement i.e. UI requirements like clear lines between different sections of the CMS so that a user can know where they are and navigate to where they need to go, or when there is a difference in being logged in as a user/researcher and as a site god or admin.

These requirements relate loosely to three areas of the process as they are experienced by the end user:

1. What the user wants to get done.
2. How the user gets it done.
3. How it is organized for the user to get it done.

Ultimately what we are looking for is a platform to facilitate the collaboration around a language documentation (research) project. A large part of this collaboration revolves around appropriate organization, perhaps the rest of it revolves around access to actionable information; having the right information collected together to and processable so that the right decisions can be made.

3. Workflow organization and Parts of the Package

A. Workflow

Work usually started like this: The PI considers a problem[3] and begins to draft a paper in Spanish addressing the problem. The PI asks the language consultants questions. The PI produces an elicitation tool and applies the tool. The received data is input to the computer and the knowledge is applied to the in-progress paper. A printout of the paper is created for the consultant to review and for the researcher running the recording equipment and the audio recording session. The recording session occurs. Appropriate remarks are made in notes and filed along with the recordings for future archival. Final papers are published and data is turned over to the archive. In general, it is safe to assume that this project does not set out to create a single corpus with a certain breath of coverage, but rather to create an open set of "final products" of a linguistic nature. These final products can usually be conceived of as papers for publication in some fashion.

B. The Complete Event

In practice, it could be conceived that there are different kinds of events, that is events with different kinds of end goals, but formally this distinction is never made nor carried into the data structure. This is evidenced by some events not being annotated while other being annotated. Some events will have audio and video, while others will only have audio. Some events have sessions which are recorded in a homemade recording booth, while other recordings are made in homes or work places of the consultants. This adds some dimensions to the background noise heard in the recordings. Because of the different end for which the events were conceived there may also be some difference in the "completeness" of the event when looking at the types of files contained in the event. Therefore the total amount of data types and data coverage possible may not, or is likely to not be present in every event. However, the *Complete Event* could have all or any of the following things in it. The following tree is an attempt to graphically represent the options which might be found within the *Complete Event*.

1. The Event
- 1.1.The Blank Elicitation Tool
- 1.2.Scan of used Elicitation tool
- 1.3.Typed Version of Elicitation tool (corrections)

2. The Audio Form
- 2.1.Audio Recording
- 2.2.Annotation
  - 2.2.1.Oral annotation (slow speech)
  - 2.2.2.Written annotation
    - 2.2.2.1.Transcription
      - 2.2.2.1.1.Phonetic
        - 2.2.2.1.1.1.Narrow

---

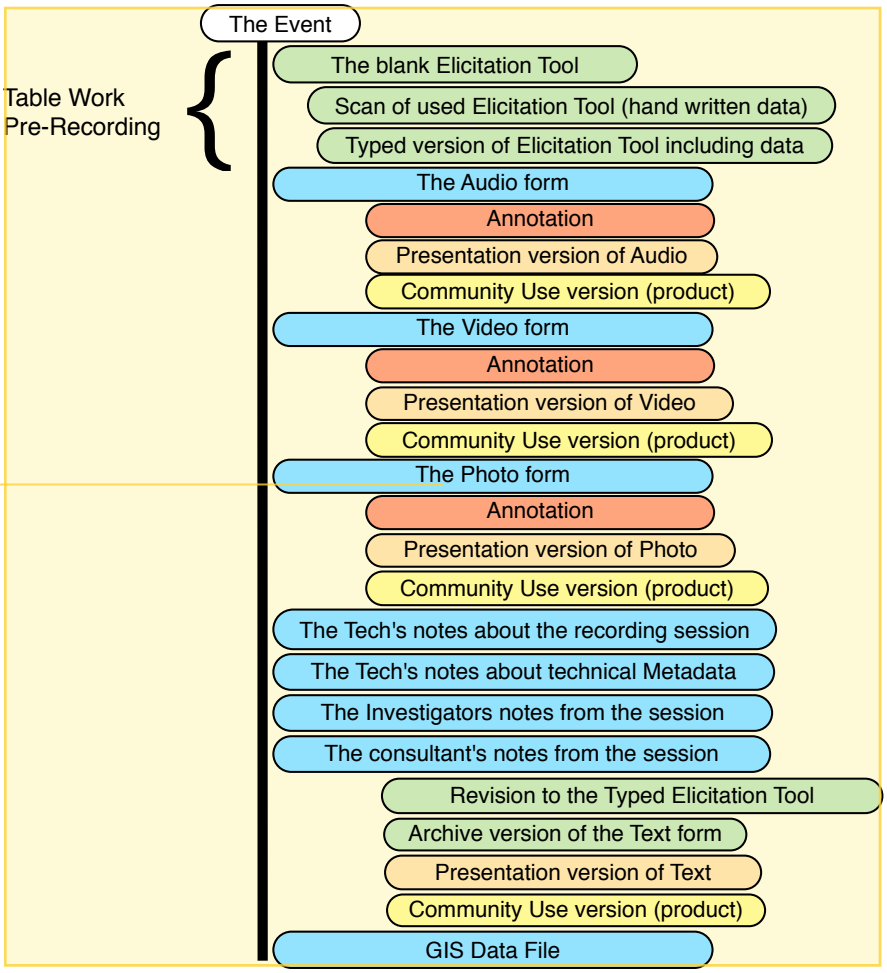
[3] Three kinds of relevance when the PI considers a problem:
1. What is interesting, that is: "What has theoretical importance?"
2. What are the outstanding questions that the SIL BT project still has?
3. What are some simple, basic things that can be presented. i.e. Some short stories, something about the numbering system, etc.

Paterson III Hugh Joseph Apr 15, '11, 8:30 AM
Additionally this could be strict IPA or not. It could be strict some other system or not.

Paterson III Hugh Joseph Apr 15, '11, 3:20 PM
Per language
So there could be a Spanish target language gloss. There could be a English other language Gloss.

2.2.2.1.1.2.Broad
2.2.2.1.2.Phonemic
2.2.2.1.3.Orthographic
2.2.2.1.4.Proper (done according to the grammar of the language)
2.2.2.1.5.Musical
2.2.2.2.Gloss
2.2.2.2.1.Word Gloss
2.2.2.2.2.Morpheme Gloss (Leipzig rules would be strict, but it could be possible to add to Leipzig rules)
2.3.Translation
2.3.1.Oral Translation
2.3.1.1.Free Translation
2.3.1.2.Literal Translation
2.3.2.Written Translation
2.3.2.1.Free Translation
2.3.2.2.Literal Translation
2.4.Presentation version
2.4.1.Audio Presentation version
2.4.2.Text presentation version
2.4.2.1.Interlinear text version
2.5.Community use version (product)
2.6.Notes about these session (during the recording)
2.6.1.Technician's notes about the recording
2.6.2.Technician's notes about the technical metadata for the recording
2.6.3.Investigator's notes about the session (during the recording)
2.6.4.The consultant's notes about the session (during the recording)

3. **The Video form**
3.1.Oral annotation
3.2.Visual annotation
3.3.Written annotation
4. **The Photo Form**
5. The Text form
6. Geo-spacial Data



Paterson III Hugh Joseph Apr 15, '11, 8:09 AM

the Annotation level does not address the differences in glossing and interlinearization.

4. Naming Conventions:

A. Folder Naming Conventions:

Project Data Folder is named Project Data. This folder is a given and a constant.

Category Folders are contrived without principle with the exception that their title should be short and in some manner relate to the content. However, For each category folder created a category needs to be added to the meta-data worksheet. Every *Event* will at least fit into one *Category*. A single *Event* might fit into to multiple categories, but it is impossible with the technical limitations of the file structure on Windows and OS X to have a folder in multiple containing folders at equal levels in the file system. That is, the folder structure on Windows and

OSX dictates a one (folder) to many (sub-folders) relationship, where a many to many relationship would be ideal, therefore we must deploy a one to many structure.

Event folders contain all the files relevant to a single event. They are named with the following pattern: “EventID - Event Title”. Events may contain several sessions but sessions do not have folders of their own.

B. File Naming Convention:

EventID-ISO639_3-TownID-SessionID-CreatorID-FileNumberID-TaskID-(VersionID)-Keyword.Extension

Example 1. E001-tcf-Zila-20110127-hp3-001-ori.wav
Example 2. E001-tcf-Zila-20110127-hp3_sam-001-oxd-v1-LiveFreeDiscussion.wav
Example 3. E001-tcf-Zila-20110127-hp3-001-dpf.wav
Example 4. E001-tcf-Zila-20110127-hp3-002-ori.wav

Paterson III Hugh Joseph Mar 23, '11, 9:12 AM

Is it version 1 of the oxd or is it version 1 of the 001?
Is this clearly marked? – no it is not.

In this project there are many kinds of digital files, which can be broadly categorized into audio, video, text, remarks and notes, annotations, and data types. More generally there are: .pdf, .doc, .wav, .mts, .mp3, .jpg, .gpx, avi, mov, .txt, .ris, .enw, etc. The goal of the file naming convention is to embed the relationships of files to each other, to Events, to the speaker’s language variety and to the creators of the file, as well as the creator’s purpose. It is realized that these relationships need to be embedded in the file’s name so that when the file is removed from it’s containing folders that it will still sort alphabetically by file name and group with it other associated files. There are ten elements to the file name. These elements are separated by hyphens.

C. The order of the File Name Elements:

The first element is the EventID. This is brought to the front so that files will sort by Event.

EventID is the first part of the file name because it is assumed that the event is the basic building block of the research agenda.

ISO 639-3 code (LanguageID) comes second so that in multi-lingual events that files will sort according the each language in the event.

TownID (DialectID) comes third because its relevance depends on the language code.

SessionID this is assumed to be the second level building block in the research agenda. It could be positioned before the LanguageID, but then separate sessions would sort together and it is assumed that various languages could be encountered in the same session. Therefore, to get the files of the same dialect to sort together then for the files within the same session to sort together, the SessionID comes after the dialect ID.

CreatorID I CAN NOT REMEMBER WHY THIS IS HERE IN THIS POSITION rather than after the VersionID and before the Keywords.

FileNumberID: This is assumed to be the third building block of the research agenda. It comes comes before the TaskID because want files of the same task to sort together regardless of which version or of what kind of task the are.

TaskID goes close to the end so as to not affect the sort order.

VersionID: The VersionID is pushed to the last because all other things being equal if there are two files differing by versions then they should sort together.

Keywords are optional and therefore are added at the end so as to limit their effects on sorting.

D. The Elements in the File Naming Convention:

Event Name Convention (EventID): This is a single Alpha character followed by three digits: A001, E002, etc. An event is a loosely defined notion. It may incorporate one or more Sessions. It may relate directly to a research goal or it may not directly relate to a research goal.

There is a challenge with the event idea not being versatile enough. There are Recording events and Research events. Research events may span several recording events and recording events may involve several research events. What is used in the name is the Research Event ID not the Recording Event ID. There is no Recording Event ID. The concept of a Recording Event ID is divided between SessionID and an EventID.

Certain Letters are reserved and are used in reference to the following categories:

Use "E" for Communicative Events	Story, text, monologue,
Use "T" for Discussion Topic	People discussing something. Dialogue. beyond Stories.
Use "L" for Lists	Word Lists,
Use "A" for Anthropological events and data. i.e. Photos do not have a "language" value but they do have an anthropological value. Photos happen in an event therefore this needs to be an event. Photos can also be in other groups (filed under "E001" or under "L001") "A001" is for non-communicative events with Anthropological value. It is also possible to have no language data, but to have cultural, visual, and geo-spacial data at the end of an event.	

There is an attempt to match these categories to the relevant discussion by Himmelmann (1998).

ISO 639-3 Code: This is the ISO 639-3 code (a three letter code, alpha characters only) of the language the the ISO says the speaker in the recording belongs to. This is not the ISO 639-3 code of the language being used in the recording. Or the language that the Speaker says he is speaking. It is an ISO recognized language for the area that person comes from. (So, in the case of the Me'phaa language documentation project, there are 4 ISO 639-3 codes for what native speakers and Mexican law consider 9 languages. The speaker may call his language one thing and the ISO may call it something else. In this case we are going with the ISO.) When speakers from separate languages are speaking in one recording both ISO codes are not used. There is no guidelines on which ISO code to use. Additionally only the codes in use by the ISO 639-3 standard at the beginning of the project are in use throughout the project. There is a certain instability in the code points of the ISO 639-3 set of points. As credible researchers investigate languages and make their findings available to the ISO these code points may divide or merge. This instability in the code is resolved, for the purposes of this project, by only using the codes available at the onset of the project.

Issues with the ISO codes: There is no Date (or version number) on the ISO 639-3 Codes. ISO 639-3 Codes change. They may even change mid stream of a project. What does one do when two languages or more are spoken in the same sound file? What is one to do when the only language spoke in the sound file is the L2 of the researcher and the consultant? how does one mark the file? Additionally, What is one to do when the ISO 639-3 misidentifies a lect based on geography? These are some of the issues we have had to deal with in choosing to use the ISO 639-3 code in our naming convention.

TownID: This is a four alpha character code representing the town the speaker says he or she is from (however that is defined). More info on how a consultant identifies with this town and more info about this town can be found in the Contributors and Locations meta-data respectively. The idea is that the TownID will most closely identify this speech variety with a geographical location. This TownID is matched to the Spanish, Me'phaa, INGEI, and Ethnologue names for the town where possible. It is also connected to GPS coordinates of, Latitude, Longitude, and Altitude; and to INGEI coordinates of, Latitude, Longitude, and Altitude. Additionally, the INGEI coordinates need to have a notes filed.

Session ID Name Convention: The Session ID is determined by the Date that the session was started and can continue to multiple other dates without changing the SessionID. Multiple sessions can occur in the same event. The session ID is a date in the format of Year Month Day with a 4 number year a 2 number month an a 2 number day all run together. Example for 15 January 2011 :: 20110115.

CreatorID: this is the initials of the creator of the file. Creator is a rather simplified way of looking at role. Current best practice dictates that the roles of participants as they relate to an item being archived is declared (Simons, Bird & Spanne 2008).

Author	Photographer
Illustrator	Facilitator
Editor	Compiler
Translator	Advisor
Instructor	Transcriber
Singer	Interviewer
Recorder	Developer
Singer	

However, we have chosen to not use a role based schema to mark participants in our metadata schema. However, in our workflow it is easy to conceptualize the difference between a session planner, a session videographer, a session recorder, a language consultant who shares their language, and an author. However, for the purposes of the naming convention, A creator is a researcher rather than a consultant and it is the researchers who have had the greatest responsibility in creating the file. One example of how this would work is that Researcher1 would develop a list to be recorded, Researcher2 would record that list with the consultant. Researcher1's initials would go on the list file, Researcher2's initials would go on recording and the recording notes files. The initials of researchers can be found in the table for the researchers Bio data.

Multiple Creators can be expressed as joined with an underscore hp3_klc, otherwise there needs only be one creator expressed as the initials of the creator.

File Number ID: These numbers are to track the number of files created in a session by one Creator. So, if in a session, the creator creates several files these will be numbered: 001, 002,

Paterson III Hugh Joseph Mar 11, '11, 8:05 AM
There are 2 problems with this ordering:
1. if there are multiple people working on a session with different creator IDs the there could be some files with both, some with name "A" some with name "B".
2. The sort by name prevents sorting the files by numeric order of file creation. so file 001 of a session created by person "B" may not get sorted before file 003 created by person "A".

003, etc. (with two leading zeros). However, when a file is versioned or used as an elicitation tool then it retains the same FileID number as the original file. At this point it is the TaskID of the new file which changes. That said, there is one logical scenario which may push the limits of this naming convention: if eliciting a wordlist, the paper form of the wordlist will have a number, say, 001 because the form was created first. Now lets assume that while recording the list the recording gets stopped and a new audio file is created when recording is resumed. This introduces a one-to-many relationship between the paper file and the audio recordings. The second audio recording will use the 002 FileID rather than the 001 FileID as would be expected in this situation. The one to many relationship will be recored in the written meta-data.

TaskID: The TaskID is the short (abbreviation) version of the *Task Types.* The Task Types is an open category of defining the kind of event that created the file. It is open to the addition of more task types, but the addition of more Task Types will require the review of previous files to determine if they belong in the new Task Type or if they belong the stated Task Type.

Paterson III Hugh Joseph Apr 15, '11, 9:23 PM
this is different than a video original. Video original does need to be different because we have audio rolling simultaneously for the same event on two items. if they are both ori then they must be marked 001 and 002. But if they are marked ori and vid then they can both be 001

Task Type	Abbreviation	Explanation
Oral Original	ori	This is an original orally elicited text or wordlist
Oral Transcription	otc	They repeat something slowly while listening to and *Oral Original* being played.
Oral Translation	otl	They repeat something in another language while listening to either an Oral Original or an Oral Transcription.
Oral Discussion	oxd	This is a multi person discussion in the target language. Where an Oral original is Monologue, this is di-, tri-alogue (and more)
Written Transcription of Oral	wtc	This is the written form of an Oral Original, Oral Translation, Oral Discussion. It is Spoken first then transcribed.
Written Translation of Oral	wto	This is the written translation of the an Oral text. There was no Oral translation, the process went directly from target language to a written translation. This signifies both a transition in modality of communication and in language of communication.
Written Discussion of Oral	wxo	This is a written discussion about what was discussed orally. (it could be the transcript of an online chat room where the content of the chat is a discussion of the things said or the topic of the oral discussion). This is not a transcription of an oral discussion.
Reading	ovr	Oral version of something written. This is used for the reading of a text. Sometimes a story will originate in written form. The individual or the researcher may want to get an oral version of this text. The distinction between Oral Original and Reading is very important. Readings are not Oral Originals because the prosody and speech patterns of a text coming from memory is different than the prosody and speech patterns of a reading.
Written Translation of Written	wlw	This is where a written text originates in the target language and then gets translated to some other language in written form.
Written Original Draft	wod	Usually a hand written version (The thought here is that a written text might go through several edits or version before it is considered complete by the author.)

Written Intermediate Draft	wid	Some in between version usually typed. (The thought here is that a hand written draft will be the first draft and a second draft will be a computerized version. Although this is not strictly true.)
Written Original Final Version	wof	Usually a final type written version. (A final version is one where the author considers the work acceptable, and presentable. In language documentation, documenting the written genres of the language is as valid as documenting the oral genres of the language. Documenting the drafts are useful for comparison of texts with the language structures of the language of wider communication and with other genres within the language.)
Photo Prompt Tool	ppt	This is a Photo which was used to confirm or prompt an idea during elicitation. The Prompt Tool category is really larger than just photos, it could include Video, texts, audio, or objects. These other parts are not added to this meta-data introduction because this project has not yet used these items in a prompt session.
Written Original Prompt Tool	wop	This is the prompt tool created by the researchers. This is especially important for readings of lists, so that a person reviewing a corpus later will be able to "see" what the language consultant "saw" during an elicitation. i.e. A consultant might be given a list of words and asked to read each word three times. these lists are also important because a consultant might note written changes to the files on the copy of what they are reading.
GPS Data File	gps	This is a .gpx file that was collected. .gpx files are XML files which are used to store data from GPS devices. Data from these files can be extracted and then embedded in various other file types as embedded meta-data.
Recording session copy	rsc	This is what the consultant used during the recording session. This is important because as a consultant reads the prompt, they may indicate that changes need to be made to the prompt, even if the prompt is a printing of their own work. This revision to their work needs to be captured and expressed.
Audio Recording Notes	arn	These are the Recording technician's notes. These are important because they show what the recording technician was doing (i.e. volume settings or where items were repeated) during a recording. (although these are titled ARN they equally apply to Video Notes or to Logbook Pages)
Cultural Event	cev	Some Sort of Cultural Event. Although everything in a sense is a cultural event, this is reserved for something larger than a single text or a text collecting event. This usually occurs outside the studio. Cultural events will most likely all fall under the event prefix "T". So event "T001" may contain a file with -cev- but is not required to. It is hard to imagine an "L001" which would also generate a file which would have a -cev- as part of the file name. The -cev- Type Task is conceivably mostly used with Audio and Video Recordings. i.e. the recording of a wedding, or a celebration.
Derived Product File	dpf	This tag is used for a product that has been post-processed. (The post processed file gets this tag rather than the "ori" tag.) i.e. like an original video file is created and then down sampled and effects are added and the video is made into a consumer product. The new file becomes a derived product file.

Paterson III Hugh Joseph Apr 15, '11, 8:10 AM

This is different than a glossed version.

A Glossed version is also different than an interlinearized version.

Paterson III Hugh Joseph Mar 11, '11, 2:42 PM
What about video recording notes?

Paterson III Hugh Joseph Mar 11, '11, 7:52 AM
I need to add a file type for photos.

VersionID: In our workflow it is conceivable that a particular file might be versioned. Especially working with digital text based texts. So if a consultant does create a written text by hand; that copy would be scanned to a PDF. That PDF would have -wod- in the file name. When that text is typed it would then have -wid- in the file name in stead of -wod-. Now if, that -wid- was then edited prior to becoming a -wof- (written final version), it would need a VersionID to distinguish it

from prior iterations. All versions which are not marked for version are assumed to be version 1 (It is in this sense that the tag is optional). So the second, and later versions, are marked with -v2-, or -v3-, etc. tag. The VersionID has the syntax of the letter “v” for version and a numeric of the version. No leading zeros are added before a numeric.

Keywords: Because it is helpful to the researchers to have one term at the end of a file name which is Mnemonically associative with the content of the file a single Keyword is permitted immediately prior to the file extension. The Keyword is permitted to be several real words long but no spaces are allowed and the first letter of the real word must be capitalized. ThisProcessIsKnownAsCamelCase. There is no maximum length on for a Keyword but it is suggested to keep it as short as possible. Keywords are also transferred to derivative files. So if an -ori- file does have a keyword then the -otc- file describing that -ori- would also bear the keyword.

The File Extension: The computer uses this to determine what kind of file this is. The researcher really does not have control over this part of the file.

4. What Meta-data:

In general the meta-data collected reflects a desire to explain the data collected by explaining the context of the speech event being recored. It does this by explaining the “Who?”, “What?”, “Where?”. “When?”, “Why?” and “How?”. This entire paper builds upon work and ideas set forth by Gary Simons and SIL International’s Strategic Research Unit. Together they have established a methodology of both meta-data collection (Simons 2009) and language documentation (Reiman 2010). However there are some limitations on methodologies employed by Reiman. This current project pushes the organizational structure presented by Simons to the limits of its original intent by adding new elements to the language documentation workflow as presented by Reiman. These challenges have been discussed in the above section on workflow. In general there are four major categories meta-data which affect a language documentation project:

Paterson III Hugh Joseph Jan 19, '11, 2:42 PM

get NISO citation.

- *Descriptive meta-data*: supports discovery, attribution and identification of resources created.
- *Administrative meta-data*: supports management, preservation, and appropriate usage of resources created
  - Technical[4]
  - Distributability, Use Limitations, Access, Copyright, and Intellectual Property Rights
  - Procedural or processing
- *Structural meta-data*: maintains relationships between the parts of complex, multi-part resources (Spanne 2008)
- *Situational (Bergqvist 2007)*

We are collection meta-data about:

- People
- Equipment
- Equipment settings during recording
- Locations
- Recording Environments

[4] It is important to note that the line between *descriptive* and *administrative* can be blurry especially when users might browse an archive's holdings by license or by frequency sampling.

    - Times
    - Situations
- Linguistic Dynamics

A. People:

There are two major types of people: us and them, or in a more Politically Correct phraseology: Researchers and Language Consultants. More formally this is the difference between those who have signed an informed consent form to contribute data and those who are mentioned in the IRB proposal (though theoretically people in a public place may also also be recored without an informed consent form).

Researchers: Names, Contact info, initials, IRB status, Gender.

Language Consultants each have these fields: Short Name, Paternal Surname, Maternal Surname, Given name(s), Name preferred in oral contexts, Preference for name in print, Birthdate, Birthplace (registered as a Location in the Locations section), Major place of residence when growing up (registered as a Location in the Locations section), Current major place of residence (registered as a Location in the Locations section), Gender, Education, Marital status now[5], Family information when growing up, Place of L1 learning, L1 ISO 639-3, L1 [lect], L2 ISO 639-3, L2 [lect], Mother's language ISO 639-3, Mother's language [lect], Father's language ISO 639-3, Father's language [lect], Time away from home, Other Notes, Informed Consent file, L3 ISO 639-3, Contact Info.

B. Equipment:

There needs to be an equipment list from which the members of the project can draw details of the equipment from. This needs to minimally include recording devices and microphones. Other information like serial number, value, insurance policy, and owner should also be tracked. (Indeed there is such a list. It is called “The Project Inventory”.) There is also a third equipment list that contains the possible options of equipment, recording patterns of mics, links to the equipment manuals, etc.

Which device created the file: Pick from Equipment List

Mic used: Pick from Equipment list.

C. Equipment settings during recording:

Recording settings: Bit Depth, Sample rate, Analog gain, Digital Gain, Modifiers in the line (amplifiers, battery packs to boots signals, etc.). Recording volume on start of the recording (changes to the recording volume go in the notes field).

Channels: Right Mono, Left Mono, Dual Mono, Stereo

Windscreen: With Windscreen, Without Windscreen.

Mic Settings: Low pass filter or high pass filter (yes or no.)

Wires used: (gold plated connections or mixed, or crappy cables, etc.)

D. Locations:

---

[5] This needs some thought. Marital status v.s. the person is married to: What if the person is a widow or divorced?

Paterson III Hugh Joseph Jan 17, '11, 4:02 PM
Change this value. to reflect the word “location”.

Each location has a: Project Name (which becomes the TownID used in the file naming convention), Spanish name, Indigenous name, Ethnologue name, Our GPS coordinates (Long), Our GPS coordinates (Lat), Our GPS coordinates (Alt), The Datum used in our measurements, INEGI Name, INEGI coordinates (Long), INEGI coordinates (Lat), INEGI coordinates (Alt), The Datum used in INEGI measurements, Notes on INEGI Data, GoogleEarth coordinates (Long), GoogleEarth coordinates (Lat), GoogleEarth coordinates (Alt), The Datum used in GoogleEarth measurements, Notes on GoogleEarth Data, Notes on Locations which includes notes on data sources, INALI designation.

E. Recording Environments:

Location: this is the Location of the recording event. It gets all the possible attributes of all the locations. (It is just added to the project locations and its short name is added in the meta-data about the file.)

Distal Settings: Distance from the Speaker to the mic, arrangement of the mics relative to the speakers. This is usually put into the notes field.

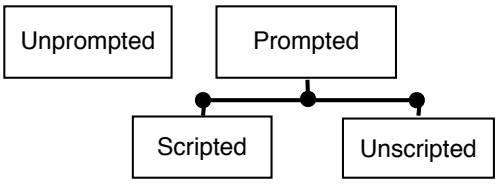
Setting: give the location from a more cultural standpoint, such as “In X’s home” or “Outside the church” or “Beside the river” These kinds of descriptions are setting the viewer, of the archived work, and expectation and a layout awareness to under what conditions the recording event happened. i.e. was in a homemade recording studio?

Situation: This meta-data element flows from Bergqvist (2007), where the conditions (socio-linguistic and participant relationships are explained. This might look like: “We were recoding at so-and-so’s neighbor’s house, they were expecting us, they were talking with so-and-so’s dad and there were “these” people in the audience”, but better examples can be found in Bergqvist.) Give any other details that seem relevant about the context for the event or the course of events that caused the event to take place.

Times: Time of start, time of finish, date, This is different than duration, because the duration is the duration of the recording, this is measuring the start and stop times of the Elicitation sessions.

F. Linguistic Dynamics:

Elicitation dynamics:
There are several dynamics which we have esteemed to be useful to track in regards to elicitation techniques used in this project. These follow two sets of two variables each with one set being subordinate to the value of one of the parts of the first set.



Unscripted v.s. *Scripted* - Was the elicitation scripted?

Even if an elicitation is unscripted it does not mean that it was without prompting. i.e. give me a sentence for each of these words, could be an example of sentences need which were

unscripted but prompted. A reading is always scripted, a translated list is always scripted. The generation of a list of birds without prompt tool would be unscripted.

Prompted v.s *Unprompted* - Was the elicitation prompted?

An *Unprompted* text is a texts which the Language Consultant Volunteers without a prompt. i.e. The Language Consultant walks in one day and says: “Look a this story I wrote this weekend.” Whereas a *Prompted* text is one where a general response requested. i.e. Do you have any stories about birds?

Written Texts

Drafted v.s. *Final*

Texts which originate in a written form may go through several revisions. Texts which start in an oral form may also reduced to words and then the language consultant may be encouraged to “revise” their text. So we find that written texts are different from transcripts and as such might be in a “drafted” state for several revisions. It is anticipated that at any point in a revision additional material of sentence length or paragraph length might be added. Equally plausible is the option that some material might be removed, however, it is more likely, based on experience of the P.I. that material will be added rather than removed. We are noting that there is a difference between a Draft and the Final version of a text.

Version number

Every text should have a version number.

The first version would be assumed to be the hand written version.
The second version would be the keyboarding it into FLEx with corrections. These corrections need to be noted in FLEx so that we do not introduce the mistakes from version 1 into version 2 and then a PDF of version 2 needs to be printed to the file structure on the server and entered into the Metadata worksheet.

Transcript of audio

Elicited via translation: give me a translation of this. Whatever “this” is.

Discourse

We are tracking one discourse dynamic, that is is the segment we are recording representative of connected speech or disconnected speech.

Disconnected v.s *Connected*

Is it connected speech? That is, is it grater than the level of the sentence?
If our text is greater than the sentence level we are classifying it as connected speech, if it is less than the sentence level then we are classifying it as disconnected.

Connected speech would ideally also have a genre associated with it. Additionally there are other ways to classify discourse including classifying the discourse by its agent orientation and contingent temporal succession.The following four discourse types (as shown in the diagram)

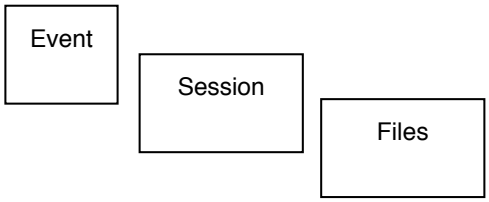
		Agent orientation	
		+	–
Contingent temporal succession	+	Narrative	Procedural
	–	Behavioral	Expository

are derivable according to Longacre (1996), as is brought out in Dooley and Levinsohn (2001). Himmelmann (1998, section 4.2) discusses in detail several continuums for eliciting various kinds of oral events as would be stereo-typical of a language documentation project. However this project's goals do not require this sort of classification. Not that this project won't collect texts which would be in different genres according to Dooley & Levinsohn or Himmelmann, it is just not in our purview to classify the texts according to such.

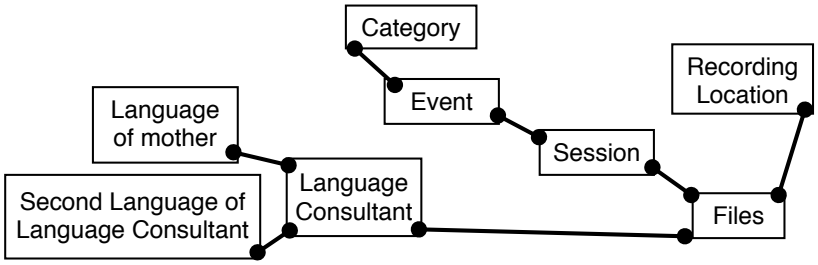
Paterson III Hugh Joseph Jan 17, '11, 3:45 PM
Write up all the fields.

5. The Relationship of Meta-Data to the recording event:

Some meta-data elements can be associated with the event and therefore be applied to all the files in that event. Some meta-data elements are specified to a particular session and can only be applied to the files from that particular session. Still other meta-data elements are only applicable to a particular file within a session.



Some meta-data can be attached to other nodes and then these nodes can be connected to the Event, Session and File nodes.



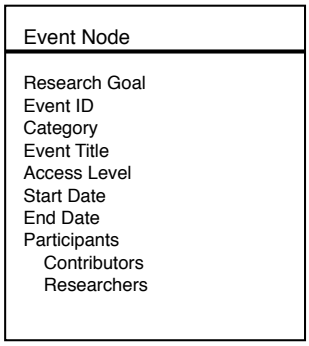
For instance the language of a consultant's mother does not need to connect to the event node but rather to the Language consultant. However, the recording location needs to relate to the file node rather than the Session.

Just because a meta-data element describes a file does not mean that it can not be inferred upon a Session or upon an Event. The following is an attempt to clearly Identify and define all

elements we are tracking in our meta-data schema. They are listed by their primary Node of attachment. Some of the Meta-data elements are described in a prior section of this paper as those elements relate to the file naming convention.

A. Event Node:

Event ID, category, Event Title, Access level, Start Date, End Date, Participants, Research Goal



Research Goal:
This is an open field which is a brief description on why we are involved in a particular event. It is also a statement of the research question we hope to answer by being involved in this event.

Event ID:
This is a unique ID given to the event based on properties of the event.

Category:
This is a term used to give semantic meaning to an event. It also allows for the cross referencing of events together based on that semantic idea.

Event Title:
This is a Unique title given to each event. The purposed is to be able to quickly describe what was happening in the event.

Access level:
This is a derived value based on the sum of the access levels of all the files with the EventID.

Start Date:
This is the earliest start date on any of the files in the Event. It can be derived based on information gathered at the File Node.

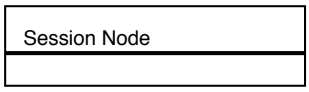
End Date:
This is the latest ending date on any of the files in the Event. It can be derived based on information gathered at the File Node.

Participants:
There are two classes of participants: *Researchers* and *Contributors*. Researchers are those doing the investigation. Contributors are those who are being interviewed for language data.

B. Session Node:

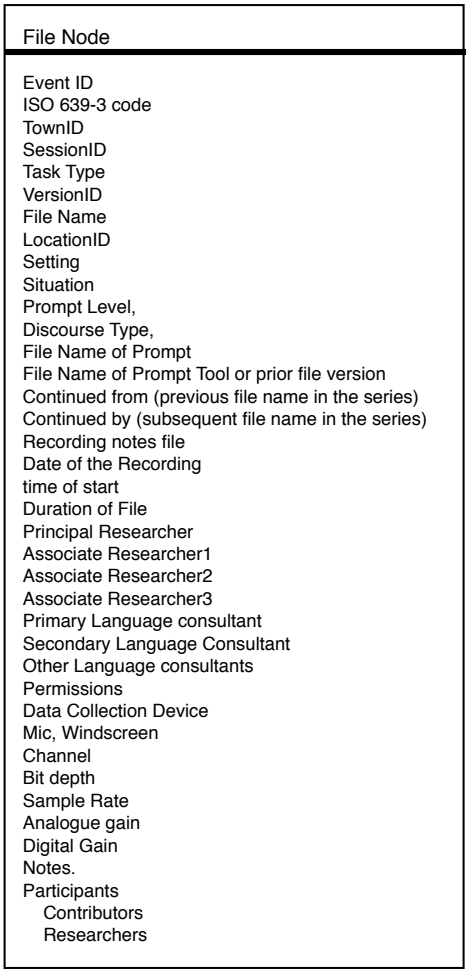
Nothing - all relevant data to the session has been esteemed to be tracked on the file node. (Although in theory it is possible that metadata could be more efficiently attached to the session

node rather than being over specified at the file node.) One particular field that would be helpful would be to have a session notes field. In this field one might generally describe why this session is different than the other sessions in the event, what was the goal of the session, etc.To this end some information is derivable about a session without actually linking the meta-data to the session. With a data-base driven solution, a Session report can be derived to include the ranges of data in the files of a given session. In our practice, a session is generally separates an event's activities based on languages being recorded.



C. File Node:

EventID, ISO 639-3 code, TownID, SessionID, Task Type, VersionID, File Name, LocationID (LocationID's are the same values as TownID's but represent the place of recording rather than the place where the consultant identifies with.), Setting, Situation, Prompt Level, Discourse Type, File Name of Prompt Tool or prior file version, Continued from (previous file name in the series), Continued by (subsequent file name in the series), Recording notes file, Date of the Recording, time of start, Duration of File, Principal Researcher, Associate Researcher1, Associate Researcher2, Associate Researcher3, Primary Language consultant, Secondary Language Consultant, Other Language consultants, Permissions, Data Collection Device, Mic, Windscreen, Channel, Bit depth, Sample Rate, Analogue gain, Digital Gain, Notes.



Event ID:
This is a unique ID given to the event based on properties of the event.

ISO 639-3 code
TownID
SessionID
Task Type
VersionID
File Name
LocationID
(LocationID's are the same values as TownID's but represent the place of recording rather than the place where the consultant identifies with.)

Setting
Situation
Prompt Level
Discourse Type
File Name of Prompt Tool or prior file version
Continued from (previous file name in the series)
Continued by (subsequent file name in the series)
Recording notes file
Date of the Recording
time of start
Duration of File
Principal Researcher
Associate Researcher1
Associate Researcher2
Associate Researcher3
Primary Language consultant
Secondary Language Consultant
Other Language consultants
Permissions
Data Collection Device
Mic
Windscreen
Channel
Bit depth
Sample Rate
Analogue gain
Digital Gain
Notes

D. Location Node

Paterson III Hugh Joseph Jan 27, '11, 11:23 AM
Find out from Kevin what these values really are.

Paterson III Hugh Joseph Jan 19, '11, 2:51 PM
This needs updating to sync with the other item mentioned above.

Location Node

Project Name (which becomes the TownID used in the file naming convention)
Spanish name
Indigenous name
Ethnologue name
Is this town mentioned in the Ethnologue?
Is this town mentioned by INALI?
Is this town shown by INEGI?
Our GPS coordinates (Long)
Our GPS coordinates (Lat)
Our GPS coordinates (Alt)
The Datum used in our measurements
INEGI Name
INEGI coordinates (Long)
INEGI coordinates (Lat)
INEGI coordinates (Alt)
The Datum used in INEGI measurements
INEGI Epoch.
Notes on INEGI Data
GoogleEarth coordinates (Long)
GoogleEarth coordinates (Lat)
GoogleEarth coordinates (Alt)
The Datum used in GoogleEarth measurements
GoogleEarth Epoch
Notes on GoogleEarth Data
Notes on Locations which includes notes on data sources
INALI designation.

E. Consultant Node

Consultant Node

Short Name
Paternal Surname
Maternal Surname
Given name(s)
Name preferred in oral contexts
Preference for name in print
Birthdate
Birthplace (registered as a Location in the Locations section)
Major place of residence when growing up (registered as a Location in the Locations section)
Current major place of residence (registered as a Location in the Locations section)
Gender, Education
Marital status now
Family information when growing up
Place of L1 learning
L1 ISO 639-3
L1 [lect]
L2 ISO 639-3
L2 [lect]
Mother's language ISO 639-3
Mother's language [lect]
Father's language ISO 639-3
Father's language [lect]
Time away from home
Other Notes
Informed Consent file
L3 ISO 639-3
Contact Info.

Paterson III Hugh Joseph Jan 17, '11, 3:46 PM

This needs some thought. Marital status v.s. the person is married to: What if the person is a widow or divorced?

F. Researcher Node

Researcher Node

Names
Contact info
initials
IRB status
Gender.

Bibliography

Bergqvist, Henrik. 2007. The role of metadata for translation and pragmatics in language documentation. In Peter K. Austin (ed.), *Language Documentation and Description*, vol. 4, 163-73. London: SOAS.

Dooley, Robert A. & Stephen H. Levinsohn. 2001. *Analyzing discourse: A manual of basic concepts: Corrected version 2007*. Dallas: SIL International.

Himmelmann, Nikolaus P. 1998. Documentary and Descriptive Linguistics (full version). *Linguistics* 36:161-95. <Accessed: 2 January 2011>. http://www.uni-muenster.de/imperia/md/content/allgemeine_sprachwissenschaft/dozenten-unterlagen/himmelmann/linguistics98.pdf

Longacre, Robert E. 1996. *The grammar of discourse*, 2nd edn (Topics in Language and Linguistics). New York: Plenum Press.

Reiman, D. Will. 2010. Basic Oral Language Documentation. *Language Documentation & Conservation* 4:254-68. http://hdl.handle.net/10125/4479

Simons, Gary. 2009. A Plan for Managing the Data from a Language Documentation Project.

Simons, Gary, Steven Bird & Joan Spanne. 2008. OLAC Metadata Usage Guidelines. <Accessed: 10 January 2011>. http://www.language-archives.org/NOTE/usage-20080711.html

Spanne, Joan. 2008. Metadata: Why, What and How (the “Who” is You). Presentation for Audio and Video Techniques. Dallas: GIAL. 29 July 2008.