Archiving Lexical Datasets and Lexical Resources

Table of contents

Background info
This post is a collection of comments which were started as a response to the unclear communication surrounding the archiving of FLEx databases, Toolbox Databases, Bit-text Corpora.

Some relevant links from which this data has been collected (or originally published):

- http://hugh.thejourneyler.org/share-the-questionnaire/
- http://hugh.thejourneyler.org/share-the-questionnaire/engaging-archives/
- http://hugh.thejourneyler.org/share-the-questionnaire/dichotomy-of-lexical-resources/
- http://hugh.thejourneyler.org/share-the-questionnaire/understanding-archive-resource-publicity/
- http://flex.hughandbecky.org/

It should be noted that the server to which these links points has been undergoing a SPAM attack and is in the process of moving. Therefore some of the pages maybe unavailable. The content has been moved here and this is the most recent copy as of 10. January 2014.

Red sections of this page are sections which I realize need more work. I either (1) don't have time today, or (2) need input from someone else to complete this section.

You might be interested in...
Most of the content here relates to communications with end users about archiving. However, the following is likely to be of interest to these specific parties:

Lexicography Service Group:

- === Clear communication about services ===
- Specific Objections and Confusion surrounding Archiving, SIL's Cloud based services, and FLEx
- 5.2 Value of the results to SIL Organizational Units
- What does a Lexical Data Set entry look like for archiving?

LSDev:

- Specific Objections and Confusion surrounding Archiving, SIL's Cloud based services, and FLEx
- 5.2 Value of the results to SIL Organizational Units

1. All Lexical Resources

Linguists, archivists, and data wranglers all think in different terms (worldviews). It is an active act of translation to communicate the semantics, concepts and implications from one community to another community. For instance communicating to linguists what it means to archivists to archive a lexical resource, and communicating to data wranglers what resources are in the archive and how they are constructed. It is important to have this conversation so that as archivists and as linguists (submitters) we know where in the spectrum of lexical resources FLEx and Toolbox data sets fit.

Why do we need a typology of resource types?

1. Marketing of products (on the behalf of linguists and economic partners, including the matching of resources to SIL's services).
2. Communication between submitters and archivists.
3. Effective matching of metadata to other metadata systems.
4. The longevity (Data maintenance strategy) and upgrading of data formats. (Because these strategies are dependent on the data formats and the file types and the data resource types.)

The archive's (and LSDev too) audience (linguists) often think in terms of "the dictionary" or "the wordlist". This is an end product orientation. It has been well argued (by Steve Echerd, Gary Simons and others) that SIL would rather see their staff oriented towards "the lexical database" because that is the source. From such a source multiple end products can be produced. While it would be ideal to flip a switch and change these "linguist's" orientation, such a switch does not readily present itself. Therefore, the SIL services which deal with Lexical data must be clear, persuasive and educational in their communications.

With respect to the distinction between a lexical data set and an out-put end product like a "dictionary", the distinction must also be clear in the archiving records. That is, on the back end of the service to archive resources, the archive record's architecture and organization needs to reflect the derivative product relationships. Since there can be multiple derivative products per lexical data set, in a DSpace architecture, it seems that both objects should be items with a relationship "is_derivative_of" So, a dictionary item is_derivative_of the lexical data set item. This is discussed in section 1.3 below.

Linguists (especially American linguists) are trained deconstructionists. This means that one of their first questions is, what do you mean by "lexical database" or "lexical data set"? Clarifying for them what we mean is important as we strive to provide clear services to this class of consumers.

1.1 Resource Types

A consistent typology of lexical resources is challenging for several reasons. One of those reasons is that lexical resources are usually at the apices of several intersecting continuums. Some of these continuums are presented below.

Spectrum edge 1	Spectrum edge 2
Wordlists	Encyclopedic entries
Monolingual	Poly-lingual
Print	Non-Print (Oral)
Single mode (i.e. textual only)	Multi-mode (i.e. text + audio, images, video)
Physical	Digital
Edited	Non-Edited
Single Author	Collaborative Production
Single IP	Multiple IP
Corpus Based	Non-Corpus Based

Beyond these continuums there is also purpose both of data collection and of the out-put product. It is in this purpose that interactive ideal is established (linguists, like many other classes of individuals often leave this idea un stated). What do I mean by purpose? If we take the dictionary as an example, then there is the "Learner's dictionary", the "Bi-lingual dictionary", the "Picture dictionary", the "Domain specialist dictionary", etc.

1.2 Databases Types

Beyond the description of the thing-ness of lexical databases using the continuums above, there is the technical description of the database. We can talk about character sets (UTF-8, UTF-16, etc.), and we can also talk about the description of "the thing" by the application which we used to create "the thing". So it might be a ToolBox database or a FLEx database, etc. But even within these descriptions there issues like database schemas, or customizations which need to be documented if we are going to think about passing our data on to other users.

1.2.1 "The Things"

So, what is this "thing" we (linguists) need to actually submit to the archive? or the "thing" we (archivists) need to expect from linguists?

In a complete toolbox project file one should expect to find the following.

```
Some-zipped-toolbox-project.zip
├──── .typ - File defining the database structure
├──── .lng - File defining theLanguage encoding
├──── .prj - Project file
└──── Datafile - with one of the following file types
  ├──── .db
  ├──── .dic
  ├──── null - meaning no file ending
  ├──── .txt
  └──── .xml
```

In a complete FLEx 6 and previous project file one should expect to find the following.

-- Some tree of files and what those files represent or include and why

In a complete FLEx 7 and Newer project file one should expect to find the following.

-- Some tree of files and what those files represent or include and why

In a complete FLEx 8 and Newer project file one should expect to find the following.

-- Some tree of files and what those files represent or include and why

1.2.2 Are the same "Things" Equivalent?

Inter-version non-equivalence.

It follows then that as we look at various databases (For instance a FLEx database) as produced by various version of software (for instance FLEx 6 vs. FLEx 8) that the thing-ness of the digital object changes. This means from a reusability standpoint that the things are different. Notice that I am not talking about user changing the data in their databases over time, but rather I am talking about the technical composition of the object. This variation would suggest that the archive should have some method of grouping like "things" together. So, one should be able to get a report on all the "FLEx 6" databases or all the "FLEx 8" Databases.

Same version non-equivalence.

A second level of non-equivalence exists and may not be obvious to non-application users (especially archivists). To this point in the discussion we have been talking about FLEx and Toolbox databases and datasets as if they are only databases, or grids of words and their relationship to grammar and meaning. However, both applications can be used in multiple ways (and in deed are by various linguists). Let me take FLEx for instance, because it is more familiar to me (but in our communications with linguists we should provide examples from ToolBox and FLEx). A FLEx 8 database used by anthropologists may include texts, but rather than word level annotations about meaning and grammar, there are a plethora of annotations for notes on culture and anthropology (with very little marked in the database for grammar). This kind of FLEx database stands in contrast to the dictionary resource which is mostly focused on grammar and meaning.

However, the example of the anthropologist using FLEx with texts points to a larger challenge when considering and categorizing the output of tools like FLEx and ToolBox. That is, these tools are not just grids of words and meanings they also have texts in them. I refer to these texts as bit-text because they are in the written mode rather than in the oral or video mode. This pluralistic function of these resources is an important element to highlight and make available to discovery for linguists. In archiving terms it is as if the FLEx item contains other items which may not be archived independently. A FLEx database may have over 100 bit-texts which are parsed and glossed embedded inside of the "FLEx database". Therefore the kind of database which is based off of rapid word collection strategy is very different in terms of content from the database based off of bit-texts. When communicating the nature of the archived database with linguists this is an important element to communicate about. This is also an important element to realize for data transfer and an Archive's Data Preservation Strategy. In the transition from FLEx 7.2.7 to FLEx 8 I have seen no less than two discussions on the FLEx users group where data migration was botched because the texts were lost. The ability of FLEx to handle texts is also a point of critique by well established Toolbox users. That is, some ToolBox users either don't understand the current power of FLEx to process (bit-)texts, or they don't understand how to move (bit-)texts processed in ToolBox to FLEx, or ToolBox really is more flexible in processing (bit-)texts than FLEx. But both applications have bit-text elements, as well as grid-like elements.

1.3 The Archive Record

As previously discussed above, the archive record needs to consider the dictionary as an item but also the data used to create that dictionary. As we see in 1.2.2 bit-texts may be a part of that foundation. I think the crucial question to ask is: Is a dictionary a lexical database? are a lexical database and a dictionary the same thing? - If they are not then should they be put in the same record (Item) or should they be independent items with a relationship connecting them?

```
Archive Institution
└──── DSpace
   ├──── Community 1
   │   ├──── Collection 1
   │   │   ├──── Item 1
   │   │   │   ├──── Bitstream 1
   │   │   │   └──── Bitstream 2
   │   │   └──── Item 2

   │   │   │   ├──── Bitstream 1

   │   │   │   └──── Bitstream 2
   │   └──── Collection 2
   └──── Community 2
```

Once we have an answer to the Is a dictionary a lexical database? are a lexical database and a dictionary the same thing? question then we can move on to asking what does each record need to contain. In many respects this is like existing package development going on in ILPT for training resources and with respect to type-setters and the products and outputs they have.

1.3.1 What does an archive's catalogue entry for a dictionary need to look like?

Best practices for file archiving of Dictionaries in SIL's Archive. (or What should the dictionary package include?)

- All dictionaries should have a lexical database associated with them.
- All dictionaries should have a PDF with them.
- All dictionaries should have the cover or jacket PDF (if one was created, if not then a comment to that effect should be in the description).
- All fonts and scripts used to format the lexical data into the PDF should be included.
- All dictionaries should have a write up of which materials in the Lexical database were included in the dictionary and how this was decided.
- All dictionaries with more than lexical content should include source files for those pages (portions) of the dictionary.
- All dictionaries with images should include the original source images in this archive package.

1.3.2 What does an archive's catalogue entry for a Lexical Data Set need to look like?

Best practices for file archiving of lexical databases in SIL's Archive. (or What should the lexical database package include?)

All lexical data sets should have a write up explaining which custom fields are used and for what they are used. ****

All lexical data sets should have in their description the texts which are included in their texts portion. (These texts should also get their own item description.)

Not all lexical data sets have a dictionary output. All lexical datasets should have a .lift output. (even thought .lift is not everything in a FLEx dataset. - ie. LIFT it does not include bit-texts)

- All ShoeBox files should have_______ file ending
  - A remark about SFM v.s MDF (the Schema used)
- All ToolBox Files should have_______ file ending
  - In all ToolBox files should be ________ components.
  - A remark about SFM v.s MDF (the Schema used)

- All FLEx databases should have_______ file ending
  - All FLEx databases should have a remark about the FLEx version.
  - What is included in a FLEx archived package?
  - What is included in a FLEx back-up package?
  - What is transferred to Language Depot? Is this the same as what is included in a FLEx Backup file?
    - How long is data on Language Depot kept?
    - Who owns the data on Language Depot?
    - What is the license of the Data on Language Depot?
    - Who has access to the files on Language Depot?
  - Is Language Depot Use considered Archiving?

**** The guidance currently provided by the archive is really confusing because, as a surveyor, I could choose to put all my words collected in the FLEx database and because of my task goal it would be "complete" however, an encyclopedic lexicographer would not consider this complete. There are really two factors which I feel are trying to be answered by the single piece of guidance curently provided by the archive. 1st) Is the answer of coverage. It should be a statical feature of the application to be able to determine how many headwords are in the lexical database. Then the application should be able to look at those head words and determine how many fields are used for each lexical item. If the database has 1500 items, and on average each item has 5 other fields with data in it but across the database a total of 30 fields are uses with many of he 25 odd fields being used under 10 times, then the total database report should be able to quantify which files are used what percent of the time, and the complete list of named fields used. For instance 1500 head words, 1495 definitions, 1374 pronunciation fields, 1500 english glosses, 300 French glosses, 500 example sentences, etc. This is an example of coverage. However, Coverage is only one metric of "completeness" review and accuracy is also a metric. If we have only 300 items of those 1500 which have been reviewed by a second speaker, or a lexicography consultant then that is a separate part of this report, and it needs to be treated separately in instructions to those archiving lexical databases.

A second thing to think about is data licensing --- I talk here about the onion model. (in the comments on that page, do a search for 'onion') I would like to clean up and clarify those ideas as I bring them into this discussion.

As a cursory remark each contributor to the data set contained in a Flex database may not also be a contributor to a dictionary produced from those data sets. This means that contributors, their roles and their contribution need to be tracked within the data base and a means for attribution at each node within the data model needs to be available.

What does a lexical dataset description need to look like?

- Number of entries
- Level of editing
- Date of active range collection
- Name of language input editor
- Name of contributors and their general contributions "tom added head words"; "sue added example sentences", etc.
- If content is derived from text then citations to those texts
- SFM or data structure used (MXB style SFM data structure is different from Philippines data structure.)

=== Data Maintenance Strategy ===

All Shoebox, ToolBox and FLEx databases should be archived once a year, at project's end and prior to conversion to another format (or version of)- like a FLEx database.

All Data conversion should be first attempted by the active project. All data from inactive projects should be updated annually with the release cycles of newer versions of FLEx. - This might could be scripted and conducted in the collaboration between the SIL Archive and the SIL Lexicography Data Conversion Service.

Lexical content Browser

What does Versions mean in a REAP context as apposed to a language depot context?

=== Clear communication about services ===

Explain the relationship between Webonary and Archiving

Explain the relationship between Language Depot and Archiving

Explain the relationship between Language Depot and Webonary

Explain how one becomes a client of the Language and Culture Archive

How does one becomes a client of the Language and Culture Archive?

- Anyone can request content from the archive via sil.org mechanisms.
- Only SIL Members may submit content to be archived.

Contracted Archiving Services

- If collegial organizations or partner organizations (and their staff) want to archive their own content at SIL's repository (REAP) then an MOU with the Language and Culture Archive must be signed, then special collections and arrangements are created for these organizations and their staff (see here for more details).

Who is Eligible to use the service?

- Anyone can request content from the archive and browse archive listings from sil.org
- Archive listings to persons with insite access may browse listings directly in REAP (the advantage over using SIL.org is that additional listings may be available which are not suitable for public access via sil.org)
- Only SIL Members and staff of organizations with active archiving services contracts may submit content to be archived

Cost?

- Cost is currently free for SIL Members
- Fee structures for staff of organizations with contracted archiving services are available through the client organization.

What data is needed?

How is content submitted?

- Via RAMP: http://ramp.insitehome.org

Who owns the data?

- The author, compiler or editor continues to own the data.
- If the work was done as work-for-hire, then the hiring organization also has a claim on the work (Currently, the SIL model is that if work was completed as part of an SIL assignment then SIL owns the data, because that person was doing work-for-hire for SIL.)

What is accessible to whom?

- Does a submitter have continued access to submissions even if they leave SIL or change roles within SIL?
- SIL Reserves the Right to control access to data regardless of who is the data owner. SIL may move content from open access to restricted access at any time pursuant to its business interests.

What happens to the data when service users submit their data to the service providers? and what happens to the data when the service request is filled?

What is the expectation for Data Maintenance or preservation? (transmission to new data formats, to keep data current)

Explain how one becomes a client of the data conversion service offered by the Lexicography Service Group.

- Who is Eligible to use the service?
- Cost?
- What data is needed?
- What is accessible to whom?
- What happens to the data when service users submit their data to the service providers? and what happens to the data when the service request is filled?

Explain how one becomes a client or user of Language Depot.

- Who is Eligible to use the service?
- Cost?
- What is uploaded?
- What is accessible to whom?
- How is this different than a FLEx Back-up?

Explain how one becomes a client of the Webonary service.

- Who is Eligible to use the service?
- Cost?
- What is uploaded?
- What is accessible to whom?
- How is this different than a FLEx Back-up?

---

2. Explaining Archived Resources

2.1 Un-archived Resources

May be:

- known (common knowledge, grant funded, etc.)
- unknown (e.g. individual research projects, that only select few know to exist)
- discoverable (posted on a personal or departmental website)
- privately kept (without public discovery)

However, no instance (and therefore also record) of these resources exists in the curated catalogues of professional libraries or institutional archives dedicated to the care and stewardship of language resources. Furthermore, in the above scenarios there is no long term preservation plan for these resources, even if a redundancy fallback copy of the data exists.

2.2 Archived but Private Resources

Private resources are severely restricted. Most people (including specialists in the language family, some archive staff and even some community members) do not know about them.

- Meta-data is hidden (not shared publicly).
- Archived objects have restricted access.

While archives may not be able to directly report on these objects, they can indirectly report what percentage of the archive's total content these items comprise.

2.2.1 Reporting on Private Resources

Example of an indirect report: 10% of XYZ archive's total contents are severely restricted. Most corpora in the archive contain less than 0.1% of severely restricted content.
Such reporting is healthy for:

- Funders - to help understand the nature of how language data is viewed by various communities. It also communicates that the archiving institution is being as transparent as possible with the data it does have - a mark of faithful stewardship.
- Archive administrators - to monitor basic trends across individual corpora (language projects or submission sets), across their entire archive's submissions, and across the larger language archiving community.
- Language and linguistic specialists - to realize that these options do exist and if these options need to be exercised, that these options for archiving are used within industry "norms". To this end, linguists also need some example use cases.
- Communities - to realize that archives have not forgotten that they have a connection with communities which are not listed in more public places.

Some restrictions are necessary. They help to build trust in archiving institutions and appropriate expectations for various stakeholders.
Note: The reasons for these restrictions should be documented so that when archive staff change, the rational for the restrictions is not lost. Additionally, the archive staff and the depositors should be in contact at a pre-determined interval to establish the continued necessity for this level of resource suppression. Frequency of communication can vary (but 3-5 years is a long time in today's world). Additionally communication about the contents of any collection or restricted collection should not be dependent upon the depositor.

2.3 Archived but Restricted Resources

Meta-data is publicly advertised via a clean navigable website, is discoverable to industry leading search engines, and through specific archiving and linguistic industry standard venues like OLAC. In contrast the the high visibility of the meta-data, the archived objects have restricted (permissions based) access.

- Meta-data is open and discoverable.
- Items have restricted access.

To maintain trust in this context, items should have: a stable endpoint (URI address) for citation purposes and a contact method for requesting access to the item (not necessarily the whole corpus). Resource items also need to be able to display their relationship to (1) the corpus as a whole and (2) other items in the corpus (especially those which are needed to function together). Additionally, archives should have in place a stewardship protocol granting them authority to administer the deposits in such cases that the original depositor is disinclined to remain alive or in contact with the archive.

2.4 Archived and Open Resources

Meta-data is publicly advertised, and the resource is openly available.

- Meta-data is open, discoverable and shareable under open licenses.
- Items are open to public access either through direct click and download or through an automated human verification (like login or recaptcha).

2.5 License and Access are not the same.

Creative Commons or CC0 data may be housed by the archive but not made accessible. An archive may choose to not share data even if the data was freely received and the archive has permission to share the data freely. SIL only shares items in manners which meet strategic goals. If it is strategic to not share an item then that item will not be made accessible regardless of the license applied to the item. This may mean that there can be a growing class of items which are "open" (by license) but not "available" (by access).

---

3. Clarifying and differentiating Archiving from

other activities

3.1 What is the difference between archiving and back-up?

Short answer: Back-up protects us from data-loss, whereas archiving inducts the data into practices of: data preservation, formal description, systematic access, and data protection.

Archiving and Back-up are different.
First lets explain what we mean by, "Copy" and "Back-up" and then we will contrast this with what we mean by "Archive". In the IT industry there are generally three kinds of Back-up (explained below). Additionally, we often see the word "archive" associated with IT products. e.g. Google Mail (Gmail) has a feature called "archive", as do many IMAP email systems. Amazon Cloud Storage promotes their some of their data storage products as being "good for 'archiving'". We take issue with how the term is used in these contexts and clarify what we mean by archiving below.

'Copy' and 'Back-up'

1. Same Drive - Onsite :: we call this a Copy. - If your computer is stolen or the drive goes bad both "copies" are lost. This is not a back-up.
2. Separate Drive - Onsite :: This is where a copy of the data lives on a second drive, but the drive is in the same location as you computer where the file exists. If one of the two drives dies then the data is recoverable from the second drive. However, because both drives are in the same location, it is highly probable that if there was a catastrophic event: Fire, Explosion (war), or theft, that both devices would be rendered unusable. An example of this kind of back-up solution for OS X is Time-Machine.
3. Separate Drive - Offsite :: This kind of back-up solution may come in two varieties: (1) Same Region or (2) Different Region.
    - Same Region works like this: The person backing things up takes the drive and makes a back-up copy and passes the drive off to a custodian who stores the drive (and data) entrusted to them in a secondary location in some other part of the city or small country. Often for this to work well, it must be done at regular intervals. In one language documentation project Hugh Paterson was involved in he used Time-Machine and replaced data on the offsite disk weekly. - Note: many linguistic field projects, including SIL entities around the world have solutions at this level.
    - Different Region works like this: The person backing things up (usually) uses a dynamic service which stores a copy of their hard drive in one or more data centers. There are several commercial services which work like this but all function slightly different. For example: CrashPlanPROe, Carbonite, or RebuSync (though RebuSync as of 2010 did not keep versions of files, and was difficult to work with the large file size of primary data in a language documentation project). At this level of back-up solution, if a datacenter in a hurricane zone like Florida was destroyed, then the data would be expected to be stored somewhere else like Los Angles, or Tokyo. The data could then be restored or accessed from this second location.

    Note: What about Google Drive, Sugar Sync, or DropBox aren't these back-up solutions? No. We would not classify these as back-up solutions. We classify them as collaboration and file sharing solutions. Here is why: when these kinds of solutions are used they, (typically) sync materials from your computer to the user's cloud account. If a user accidentally deletes their content from their computer then this deletion is also replicated to these remote file stores. Thereby also deleting the file in the offsite location. (We recognize that some services do offer versioning which does give users limited capability to recover deleted files, but these features are not automatic, and usually come bundled with premium version of these products/services.)

So then what does 'Archived' mean and how is it different from 'Back-up'? - Whereas back-up is primarily concerned with data loss prevention, Archiving is concerned with preservation of usability (of the data), discoverability, provenance (history) and identification (of the data), and then also access to data.

3.2 If my Data is in the cloud does that mean it is archived?

This is also a great question, because there are a lot of issues involved with cloud data. First cloud data is often social, and implies variation though versions (updates or changes to the same dataset) and forks (dataset splits where each set is then modified independently). Second cloud data is not on the local machine and therefore can feel to some like an "offsite back-up" solution.

With regards to lexical datasets SIL offers two independent cloud services:

- languagedepot.org - A web service which enables the send and receive functions of lexical dataset building teams to share their data with each other through FLEx's built in Send/Receive function.

- webonary.org - A website where FLEx data can be hosted and viewed.
- Additionally, some linguists and lexical data user use third party services like github.com or bitbucket to host their data so they can communicate with with collaborators.

http://en.wikipedia.org/wiki/Gnolia

http://www.wired.com/business/2009/01/magnolia-suffer/

First lets address the cloud - What is the cloud is the data secure? (and secure from what)
social data and the iterative nature of lexical resources. There are lots of tools like DropBox, SugarSync, Google Drive, etc. These are not necessarily even successful back-up strategies.

4. Specific Objections and Confusion surrounding Archiving, SIL's Cloud based services, and FLEx

and some answers to those objections

SIL Member says:

> (1) Did you want me to archive my data? I have never been asked to do so, or given any instructions on how.
> (2) I tried transferring my data from toolbox to FLEx, but my branch is actually a small group with no technical department (and I was not in the country at the time), and my sending organisation was too short of expertise to help me. I got by, but I wonder if other members are hampered by lack of technical assistance, or if it was just me.

SIL Member says:

> Note re why the dataset was not uploaded to REAP: I intended to upload them to REAP, but when I started to do that a couple of years ago, I did not know whether just the downloadable files should be in REAP or the entire website. At that time, there were no instructions in REAP about how to archive web sites. I then wrote to Laurie Nelson the REAP administrator and the matter of how to archive web sites with lexical data was taken on my others. Only just recently have the PDF files of the print publication been archived in REAP.

SIL Member says:

> Since it was submitted to SIL for e-publication and is online with Webonary.org, I assumed that it was de facto archived. If not, what should I do?

SIL Member says:

> I have posted a version of the dictionary on the web. Reason the lexical database is not archived is because it is never finished - still in process.

SIL Member says:

> What counts as an "SIL" project?

SIL Member says:

> Just tried to use FLEx couple of times, but it's really working too slow for real work (20.000+ entry dictionary).

SIL Member says:

> The project is on Language Depot for purposes of collaboration.
> I have an appointment tomorrow to begin archiving my language materials on REAP. I made the appointment before receiving your email.

Non-SIL Member says:

> I tried migrating from Toolbox to Flex but the lack of sufficient fields in the text function of Flex discouraged me from continuing.

Non-SIL Member Says:

> I am also using ELAN for annotating video recordings. I would like to export files for analysis from ELAN to FLEx, but don't have enough experience. I also want to export earlier Toolbox files to FLEx but have problem

> in doing that.

Non-SIL Member says:

> I like FLEx, but wish there were more developers on the coding side of the the tool. There are a lot of great things that just need small tweaks but are not completed because of the queue of features being added.

Non-SIL Member says:

> While we have not yet archived the FLEx database, we do intend to once work has progressed a little further - it is still in the early stages. The database will be archived with PARADISEC. Print publications are also planned, at least for community circulation, but again these will not happen until we have more data.

Non-SIL Member says:

> I am currently using Toolbox and am happy with it. But I would not mind giving FLEX a go, however, I am unable to locate a handbook or manual for the software to get started. Do you have one? I am also currently using three parallel linked dictionaries in Toolbox (multilingual Northern Australian environment) and I am not sure if this is also possible in FLEX?

5. Understanding the Results of Hugh's Lexical Database Archiving survey

5.1 Context of the survey

The Impetus:

Hugh was looking at several issues and realized that clearer communication with end users about SIL's services and products and their interrelationships would benefit end users and likely have positive effects on SIL service offerings. Hugh was particularly interested in how digitally delivered training and helps could be delivered online for various digital tools which SIL produces. The survey developed was focused at tool users (both SIL and non-SIL users). Hugh created the questionnaire and then chose Google forms to collect data. After seeing the results it became clear that several of SIL's service groups might benefit from the results.

How it was presented to Participants:

The following paragraph (or something close to it) was at the header of the invite to participate in the questionnaire.

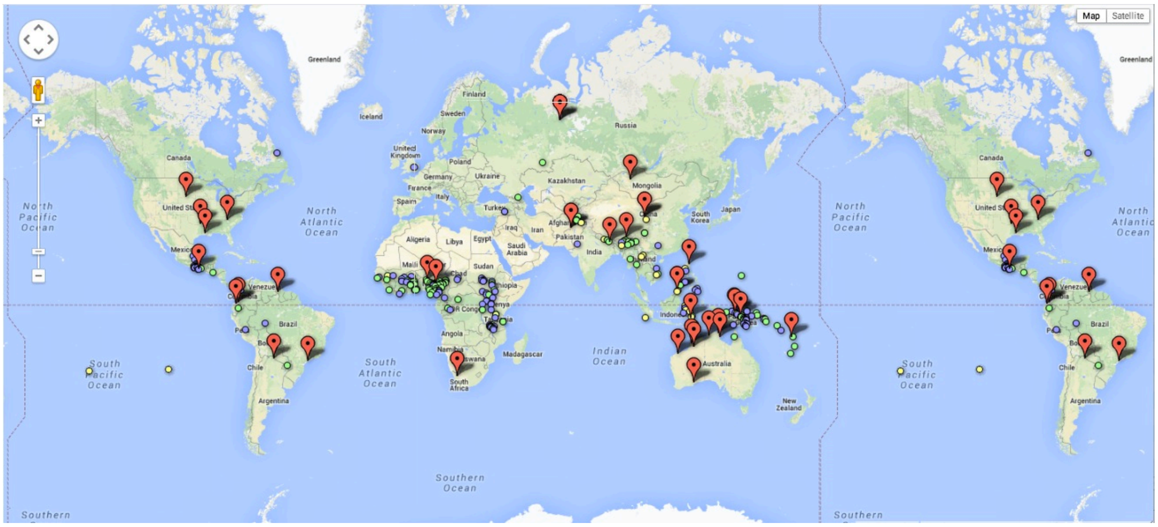
> Hugh Paterson III, in cooperation with Jeremy Nordmoe of the SIL Language and Culture Archive, is investigating the trend among lexical database users to archive their work. In their poster presented at ICLDC3 , it was claimed that less than 1% of SIL projects archive lexical datasets. Hugh and Jeremy want to know if this is common among all lexical database users or just SIL users of FLEx & ToolBox.

5.2 Value of the results to SIL Organizational Units

- LSDev - The open ended and un filtered contents show a strong desire of respondents to be able to migrate their data from ToolBox to FLEx. The Pathways to do this are not very clear to the user group.
- Lexicography Services Group - This list of respondents should be of interest for two reasons (1) these are the people who need (and sometimes even want) help via the data migration service, and (2) These are the people who are candidates for the webonary service.
- Communications - How are we communicating about SIL policies, SIL services and SIL products. The great many respondents with questions and lack of clarity should be an indicator to our success or failure in communication.
- ILPT - How are people connected with the training elements surrounded with our various products and services. The level of confusion displayed with these comments should be an indicator in our success or failure in accessibility and consumption of training resources.
- Language & Culture Archive - Are people really archiving? No. they are not. But they have some interesting mis-understandings about (1) the reason for archiving, (2) what it is, (3) how to do it, (4) what to archive.

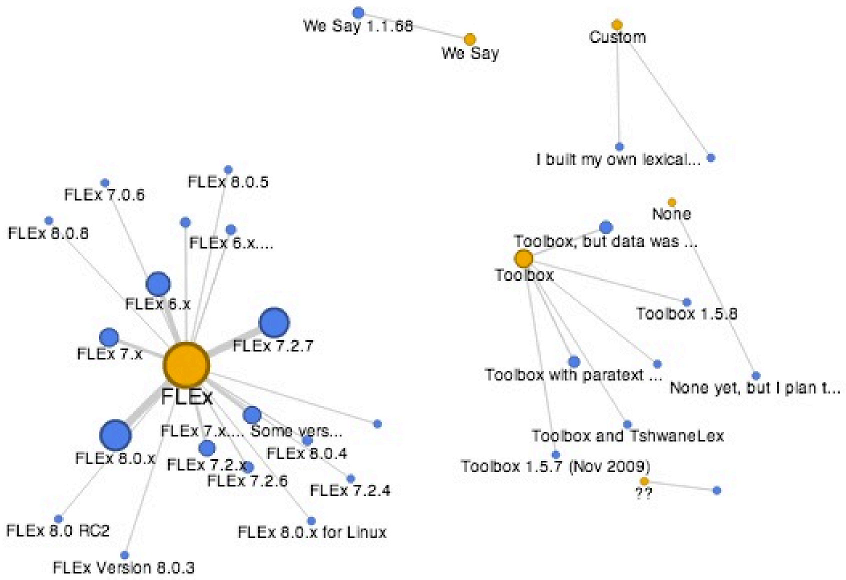
5.3 Data from the survey

Those who responded said that they work on languages which SIL.org says are from the following places. Dots of any kind represent the languages which are being described in the lexical databases. Blue and Yellow dots are respondent identified SIL projects. Green and Red dots are respondent identified non-SIL projects, or projects where this information was not provided. Red and Yellow dots are "endangered languages" according to SIL.org, while Blue and Green dots are "robust languages" according to SIL.org



Software and version can make a lot of difference in opinions about the usefulness of completing tasks.

This is a graph of the kinds of responses about software versions received through the survey. The launch of this questionnaire unintentionally coincided with the launch of FLEx 8. So, FLEx users may be a moving target. However, note that some go as far back as FLEx 6 and often users do not like to "upgrade" during a "project" - much like how students aviod installing a new OS while doing a thesis.

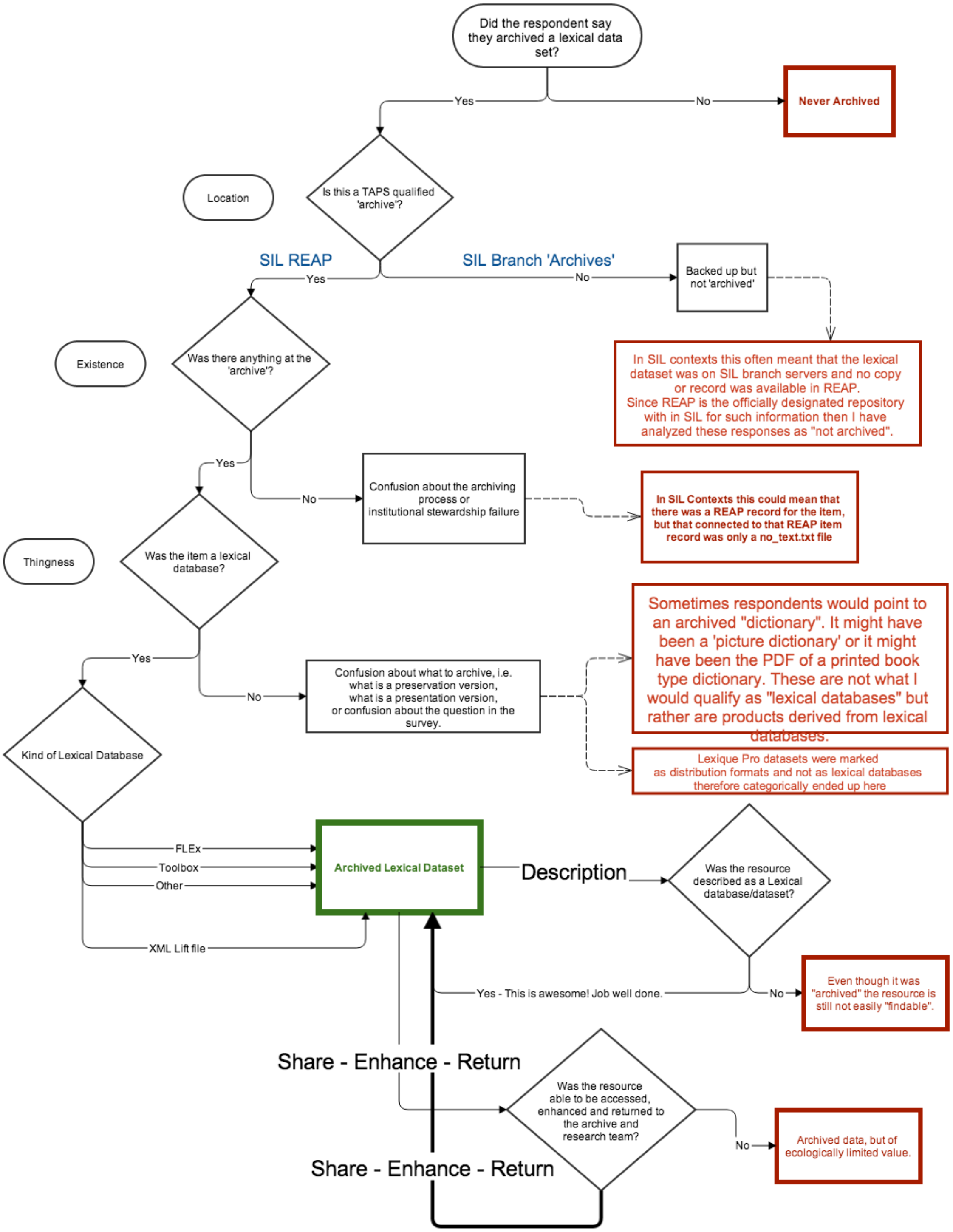


SIL Entity participation

For some reason MSEAG was the only SIL entity which actively requested that I not connect with their teams, and that they would not support the propagation of the survey questionnaire. The most supportive of the questionnaire were SIL PNG, SIL

Nigeria Group and SIL Mexico Branch who connected me with appropriate staff directly or propagated the questionnaire on my behalf and returned the results to me. Other entities were were either passive or non-interactive - Individuals from a variety of unnamed entities participated on an individual basis. The survey is still open so things could still change.

Data interpretation process



Did the respondent say they archived a lexical data set?
Yes
No
Never Archived
Location
Is this a TAPS qualified 'archive'?
SIL REAP
SIL Branch 'Archives'
Yes
No
Backed up but not 'archived'
In SIL contexts this often meant that the lexical dataset was on SIL branch servers and no copy or record was available in REAP. Since REAP is the officially designated repository with in SIL for such information then I have analyzed these responses as "not archived".
Existence
Was there anything at the 'archive'?
Yes
No
Confusion about the archiving process or institutional stewardship failure
In SIL Contexts this could mean that there was a REAP record for the item, but that connected to that REAP item record was only a no_text.txt file
Thingness
Was the item a lexical database?
Yes
No
Confusion about what to archive, i.e. what is a preservation version, what is a presentation version, or confusion about the question in the survey.
Sometimes respondents would point to an archived "dictionary". It might have been a 'picture dictionary' or it might have been the PDF of a printed book type dictionary. These are not what I would qualify as "lexical databases" but rather are products derived from lexical databases.
Lexique Pro datasets were marked as distribution formats and not as lexical databases therefore categorically ended up here
Kind of Lexical Database
FLEx
Toolbox
Other
XML Lift file
Archived Lexical Dataset
Description
Was the resource described as a Lexical database/dataset?
Yes - This is awesome! Job well done.
No
Even though it was "archived" the resource is still not easily "findable".
Share - Enhance - Return
Was the resource able to be accessed, enhanced and returned to the archive and research team?
No
Archived data, but of ecologically limited value.
Share - Enhance - Return

How many of these projects claim to be archived?

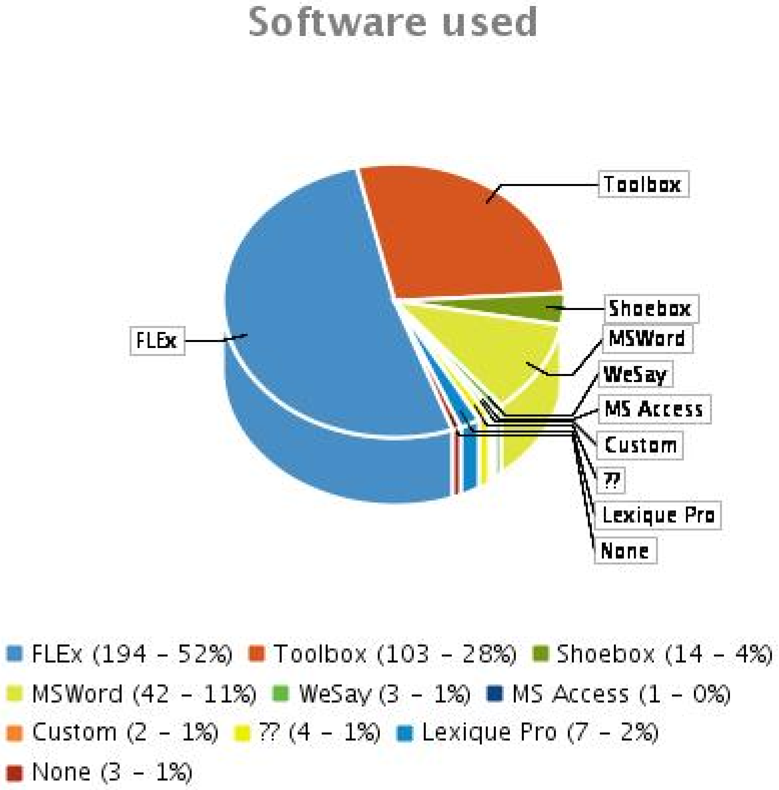
Items on the left are claimed to be archived, whereas the circles on the right are claimed by the respondents to not have ever been archived.

There were some cases where the respondents did not know if their lexical databases were archived.

Link to Raw Data :: link to processed data.

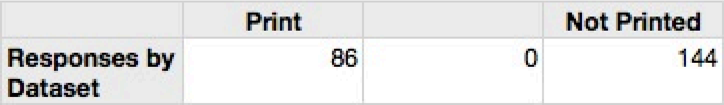
5.4 Basic Results

As of 23. May 2014 the results in the following table are accurate and represents data from questionnaire respondents. (Red figures may be based on January stats.)

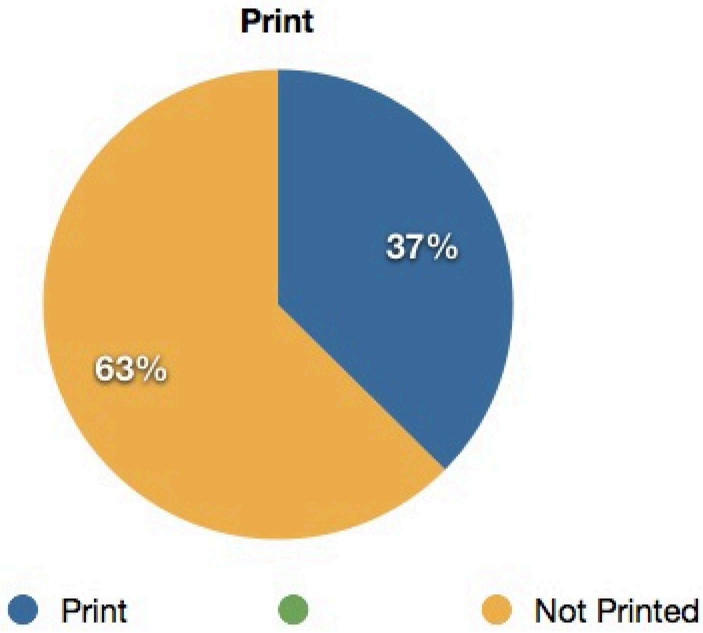


Total Databases	373	Archive Status of Responses by lexical data set.		Print Publication was produced from Lexical data set		Thinks they have archived but have not	45
FLEx	194	No Never	234	Yes	86+	Cloud Data	2
Toolbox	103	Archived with ELAR	11	No	144+	Offsite Backup	28
Shoebox	14	SIL-REAP	19			Derivative Product is archived	4+

MSWord	42	Archived with PARADISEC	14			Review (messy data)	11
WeSay	3	Kaipuleohone	2				
MS Access	1	MPI's TLA (DoBeS project)	3				
Custom	2	LMU ITG	1				
??	4	Lakota Language Consortium	1				
Lexique Pro	7	The don’t know	3+				
None	3	SIL Branch Data Store	37				
		AIATSIS	14				
		Derivative Product not Full Dataset	4+				



	Print		Not Printed
Responses by Dataset	86	0	144



-up

5.5 The Questions asked in the survey

The Questionnaire

1. What Lexical Database Solution do you use: *

FLEx, ToolBox, Lexus, TshwaneLex, etc

- FLEx Version 8.0.x
- FLEx Version 7.2.7 (Latest Stable Release)
- FLEx Version 6.... Some version in the FLEx 6 Series
- Hey I use ToolBox ‼
- Lexus
- I built my own lexical database solution.
- Other:______________________________

2. ISO 639-3 language code of the language you are analyzing/studying in your Lexical Database: *

Three letter code from "ethnologue.com". Only use lowercase a-z. Use "und" if you don't know, but be sure to put your email address and the language name under the optional answers section.

- ISO 639-3 code:_______________________

3. I have archived a version of my current Lexical Database at an Institutional Archive *

An archive like SIL's L & CA, or SOAS's ELAR, or MPI's TLA. - It doesn't have to be one of these three.

- No. - Never Archived it.
- Archived with SIL
- Archived with ELAR
- Archived with TLA
- Archived with PARADISEC
- Other:_______________________

Optional Questions

The following information would help us in our research but are completely optional.

Email address: ____

Name: ____

SIL Project

Is the FLEx (or other) database used, part of an SIL project?

- Yes
- No

If you are using FLEx or ToolBox have you produced a Print publication?

Perhaps a dictionary for local use or a more formal publication. Can you tell us about it? Got a citation or a link?

____

____

Anything we should know?

Keep it short and important - We intend to read these. ;-)

____

____

5.6 Mailing lists invited to participate

COMMUNITY INVITED TO PARTICIPATE	DATE SENT
ANU Austronesian Mailing List	16. November 2013

Yahoo! Lexicography List	17. November 2013
RNLD list	18. November 2013
SIL-LDL	15. November 2013
SIL-Survey	15. November 2013
ALGONQUIANA on Linguist List	18. November 2013
ENDANGERED-LANGUAGES-L on Linguist List	18. November 2013
FLEx Users Group	25. November 2013
Various University of Oregon Linguistic Department lists	27. November 2013
SIL-UND FaceBook Page	28. November 2013
ToolBox Users Group	28. November 2013
lingtransoft	09. December 2013
SEALANG-L on Linguist List	09. December 2013
TIBETO-BURMAN-LINGUISTICS on Linguist List	09. December 2013
SALON	11. December 2013
Wycliffe Nigeria	12. December 2013
SIL Lexicography Service Group List	13. December 2013
FaceBook Group on Lexicography	22. May 2014