Challenges of implementing a tool to extract metadata from linguists: the use case of RAMP

By Hugh Paterson & Jeremy Nordmoe

Cite as: Paterson, Hugh J, III and Jeremy Nordmoe. 2013. Challenges of implementing a tool to extract metadata from linguists: the use case of RAMP. Poster presented at 3rd International Conference on Language Documentation and Conservation, at the University of Hawai‘i Mānoa, Honolulu, HI. February 28 – March 3rd.
Behavior in the Archive Submission Process:
the role of User Experience in the success of acquisitions

Comments invited via WordPress or YouTube:
Hugh.Paterson@sil.org & Jeremy_Nordmoe@sil.org
• SIL has nearly 80 years of history working with minority language communities.

• About 1 million relevant non-digital objects are estimated to exist in SIL networks.

• About 50 million relevant digital objects are estimated to exist in SIL networks.
A challenge not without a plan

An institutional repository is needed

DSpace 1.6.2 is chosen (now up dated to 1.8.x): the submission process User Experience is determined to be wanting, from three perspectives, linguists, archivists, and users (browsers).

RAMP is developed to fill the need for a simplified form based approach to submissions.

The expectation is that: Submitters will avoid the standard DSpace entry method, and want to use RAMP.

RAMP is simple compared to DSpace. - True

But do people really want to use RAMP?
RAMP and the SIL archive submission process
Previous models of submissions in SIL

1947-2010

Local bibliographer determines meta-data needed about the item and this is usually done from a librarian, rather than an archivist’s perspective.

Local "archives" are functionally operated as private libraries.

A majority of content described and handled in these models is non-digital.

Notice is sent to the corporate bibliographer (but possibly also a copy of the item).

Linguists would also sometimes directly submit works to the corporate bibliographer.
The RAMP model of submissions

**2011 - Present**

Linguists and Staff submit content via RAMP to specific DSpace collections

Content is then piped to appropriate access points to serve various communities

pending approval by curators, objects get included into DSpace data stores

Curators of specific collections check the integrity of metadata and content

2013 marks the end of the SIL Bibliography and the beginning of the online presence of the SIL archive.
The “RAMP” Effect

Prior accession rates over the last 10 years have averaged between 1,500 and 2,000 items per year.

SIL Archive Accessions over the last decade

RAMP submits in 2012 alone produced 4,223 new accessions (bit streams).

81% of new bit streams contain more than one digital object.

81% of new bit streams contain more than one digital object.

81% of new bit streams contain more than one digital object.

Direct to DSpace v.s via RAMP ratio

RAMP is a desktop application which was created to overcome complexities in DSpace UI so that field workers could directly submit to the archive.

RAMP contributed an increased capacity (200%) to accession materials, but was it equally successful in reaching its intended audience?
So who is making submissions

If RAMP truly meets a recognized need in the organization as perceived by linguists, we would expect to see lateral peer-to-peer spread in the user base of the software. This is not the case. Most users, by quantity of submissions are archivists.

But why are there so many single use submissions? What can we learn from this?

Based on the population of SIL staff working in language projects, it is not unreasonable to expect the user base of RAMP to exceed 2000 unique users per year.

**Penetration among the targeted user group is around 2%.**
78% of all submissions to the archive in 2012 were made by SIL staff with a specific role in archiving.

In 2012:
- 2.3% of all SIL staff globally made RAMP submissions
- 3.5% of language development staff with roles in active projects made submissions via RAMP

Repeat DSpace submitters, who have never used RAMP tend to be in publishing roles.

Of the 141 people who have made submissions to the archive in 2012, only 122 of them used RAMP. Each job type (except archive-temp) and all six major administrative units of SIL are represented by those 19 users who did not use RAMP at all and made submissions to the archive; 12 of those 19 only made 1 DSpace submission.

8 of the top 10 RAMP users have a role in archiving.

Of the 122 RAMP submitters in 2012, 36 of them did not use DSpace and also only submitted one item.
- 30% of RAMP users chose not to use the software again (having never compared it with DSpace).

Of the 36 users:
- 12 were submitting objects on which they were not contributors, e.g. not author, not composer
- 12 continued to submit materials to the archive, but chose to do so though another person, or via a non-digital means.
- 6 individuals who had previously submitted items to the archive through another person, or non-digital means, chose to attempt to use RAMP, but had no desire to continue to use RAMP (or possibly further opportunity to use RAMP, the archive does not know).

8 of the top 10 RAMP users have a role in archiving.

Of the non-single use submitters, 38% (32 users), had an archiving role. If persons with a publishing role are added then it goes to 45% (39 users).
Why?

Why is the largest user group, by number of contributions, archivists, rather than field workers (linguists)?

*RAMP is simple compared to DSpace.* - True

*But do people really want to use RAMP?*

*Are less than 6% of SIL staff concerned with archiving?*
The “DSpace” Effect

Of the 37 DSpace submitters in 2012, 13 of them have an archiving role.
- **60% of DSpace submitters are non-archivists.**

This stands in contrast to all 141 submitters, of whom 38 have an archiving role and 13 of these used DSpace.
- **60% of SIL archivists don’t use DSpace for submissions.**
  Of the top 12 repeat DSpace submitters, 10 of them have roles in archiving or publishing. Others were, 1 each: training, and fieldworker. - *Those SIL archivists who use DSpace use it a lot (or for batches).*

For 12 DSpace submitters (none of whom have an archiving role), the DSpace experience was the only digital interaction with the archive that they had in 2012 (they were non-RAMP users). All of these users only contributed once. They were all contributing their own content.
  - 4 of these 12 would go on to contribute materials to the archive via another method (non-digitally, or through another individual submitting the content). But again not as a user of any digital system for archiving with SIL.
  - 5 of these 12 had already submitted something to the archive either through another individual or through non-digital means when they tried DSpace for the first time. They have a belief that archiving is important, or their works are in corporate publishing workflows which result in archiving.
Designing Experiences

"Designing for experiences is fundamentally about people, their activities, and the context of those activities..."

User adoption of RAMP can not be solely attributed to its User Interface and feature set. Social attitudes about archiving in an SIL context, language program management strategies which do or do not require archiving, and the task perception by the RAMP user must also come into account.

The really great experiences are deeply rooted with insights into motivations, desires, emotions, cultural & social patterns, beliefs and other deeper considerations.
Social Attitudes of Linguists

Nordmoe (2011) claims that archiving meta-schemas remain too complex for linguists... We find this objection un-grounded coming from linguists who devise meta-schemas for for describing language... (Though we make no claim that any schema is innate).

Linguists use a variety of complex metadata schemas during their working day - though some linguists may be unaware of them. The user experience challenge for archivists is: can archivists access these data at the point of first use?

So is archiving truly complex or is the perception that the questions are irrelevant and therefore the process is perceived as unnecessarily complex?

Other attitudes encountered:
- The archiving institution:
  - loses content and materials
  - can/will not restrict content access appropriately
  - can not publish content to open access points in a timely manner
  - does not value certain types of content or will charge for access
  - does not maintain accurate records because all the information provided by the linguists does not fit into the institution's metadata schema.
- Confusion about the organizational structure of the archiving institution.
- Opinions that archiving should only take place in the country where the language is indigenously spoken.
- I have heard a linguist say "I hate Metadata". For a linguist to value the archive more, more than discovery metadata must be exposed about the data in the archive.
- The I don't care attitude: "I'll just turn it over to the archive to do whatever they do."

Monday, March 11, 2013
Do Attitudes affect when linguists chose to archive?

604 items archived in 2012 are known to have been created or published in 2012.

One of SIL archivist’s perceived challenges is to acquire data and resources in a reasonable amount of time relative to that object’s creation. As time passes linguists are more likely to not be able to provide metadata details to the archive for the benefit of the archivist or for the benefit of future users of those resources.

Are items being accessioned in appropriate amounts of time? Or, do linguists retain the attitude: archiving is my last task before death?
Task perception becomes a major issue in user interface design. The user interface also has a major role in setting the mood for the entire interaction. An archivist wants to know what the object is that the submitter has.

Linguist:
How does RAMP relate to my other data in my workflow?
How does RAMP enable me to keep the promises I made for funding?

Archivist:
How does RAMP tell me what the item is so that I know which “shelf” to put it on?

The perception of where RAMP is situated in the entire eco-system is foundational to widespread user adoption.

Archivist:
How does RAMP relate to my other data in my workflow?
How does RAMP enable me to keep the promises I made for funding?

Linguist:
The perception of where RAMP is situated in the entire eco-system is foundational to widespread user adoption.

Which screen should come first?

The linguist is trying to give the archivist something. The linguist is also the initiator of the conversation between submitter and the archivist.
Reaching Meaningful

“Emotion and cognition conjointly and equally contribute to the control of thought and behavior.” (Gray 2002) Often the design of linguistics based software is focused on specific tasks, not creating meaningful experiences.

Does RAMP cross the chasm and become meaningful to its users? If it did, would we expect to see lateral spread (peer to peer) in the user group, rather than organizational tree based spread?

What is the emotional impact on the RAMP user’s attitudes towards archiving? Is it the meta-schema which is too difficult or is it the relationship through the software? - To the RAMP user, is the experience worth repeating and telling their friends about?
Working with existing metadata

Most media files have metadata of some kind embedded in them. Working with this metadata (and allowing the user to verify it) rather than completely ignoring it would save the user effort, and in some cases time. The perceived gain would make the user appreciate the software more. They feel like they have to answer fewer useless or obviously answered questions. But these questions are and can be pertinent in accessions of non-digital objects.

Matching keywords to existing embedded metadata, linguistic ontologies, and archive ontologies, (and perhaps text searches) could have a significant return on investment for the archive as it endeavors to leverage its materials to users.
More efficient input

Across the 2490 items which had a creation date, there were 29 different ways that date was expressed. An additional 4 ways of expressing the date were found in the issue date field.

Airlines find ways for customers to often select two dates per ticket purchase. What could app designers learn from other industries’ designs?
Preparing through embedding

RAMP does not prepare files for the archive by taking the metadata provided by the linguist and then embedding it into the file types native metadata options.

Some archiving institutions would rather do this after receiving the files, other institutions would rather the linguist to do this prior to submission.

RAMP also does not visually let the linguist know what metadata is embedded in the files they are uploading. This embedded metadata, if it later becomes available with the file as the archive provides it, could have unintended consequences. Part of the submission process for archives should be (to the best of their ability at a given point in time) for the archive to discover what the intended consequences are of distribution.
What kinds of digital objects are being submitted?

One of the big questions in archiving is: *are digital objects clumped or divided appropriately?*

To assess, this an archive might look at how many .zip files and archive type files (.iso, .tar, .gzip, .etc.) it might have accumulated. This year the SIL archive added 1,030 new .zip files. Zip files may be a reasonable transmission or storage format, but if the reason for the submission in a .zip format is because the submitter didn’t want to take the time to archive each digital object independently, when the it is more appropriate for them to be added to separate bitstreams, then something is wrong with the *user experience* in the submission process. These pressure points become the new wave of bottlenecks in distributed archive submissions.

Another way to assess clumping and dividing is through relationships like: *X has part Y or Y is a part of X.*

1,821 of the items added in 2012 have relationships to other items in the archive.

In 2012 there were **475** active participants in the FLEX Google Group yet there were only **4** instances of a FLEX data set Archive. - these were submitted by **2** contributors and one instance was a version of a previous instance. (Not all SIL FLEX users are in the Google Group, nor are all 475 members are SIL staff.)

---

**Types of items submitted in 2012**

<table>
<thead>
<tr>
<th>Type of item</th>
<th>Total number of items across all submissions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Textual based objects (presentations, papers, PDFs)</td>
<td>5,598</td>
</tr>
<tr>
<td>Image based objects (.psd, .jpg, .raw, .tiff)</td>
<td>2,104</td>
</tr>
<tr>
<td>Unknown (obscure object formats, fonts, ISOs, .zip)</td>
<td>1,484</td>
</tr>
<tr>
<td>Audio based objects (.mp3, .aiff, .wav)</td>
<td>1,003</td>
</tr>
<tr>
<td>Text-Data based objects (toolbox files, FLEX, .xls)</td>
<td>67</td>
</tr>
<tr>
<td>Web formatted files (html, css)</td>
<td>45</td>
</tr>
<tr>
<td>Video based objects (.mov, .vob, .mp4)</td>
<td>32</td>
</tr>
<tr>
<td><strong>Total Digital Objects</strong></td>
<td><strong>10,333</strong></td>
</tr>
</tbody>
</table>

60% of images may be part of text based scanning of old documents to archival .tiff formats.
Project Management

There are three issues which affect project management, two of these have a drastic impact on how and where the output of a project is archived:

- **Community Involvement** - is a major concern for access of materials but less so for how a project conducts its archiving
- **Funders of the project** - often put requirements on projects to be archived
- **Institutional relationships** of the researchers often dictate where a researcher can or will not archive.
Major Areas of Administration

International and Global Collections
Submission methods used

Each administrative area of SIL has different strategies for archiving content. These management strategies affect which tools are presented to various sets of linguists and therefore also who does the work related to submission to the archive. Roughly speaking, higher rates of RAMP usage mean that the area staff is more self-sufficient in terms of submission to the archive. Compare Yellow/Green to Red/Blue.

<table>
<thead>
<tr>
<th>Administrative Area</th>
<th>DSpace Submits</th>
<th>RAMP Submits by Area Non-field Staff</th>
<th>RAMP Submits by Area Field Workers</th>
<th>RAMP Submits by Non-area Staff to Area Related Collections</th>
<th>All Submits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td>253</td>
<td>244</td>
<td>133</td>
<td>481</td>
<td>1,113</td>
</tr>
<tr>
<td>Americas</td>
<td>345</td>
<td>317</td>
<td>4</td>
<td>941</td>
<td>1,607</td>
</tr>
<tr>
<td>Asia</td>
<td>40</td>
<td>455</td>
<td>372</td>
<td>422</td>
<td>1,289</td>
</tr>
<tr>
<td>Eurasia</td>
<td>3</td>
<td>163</td>
<td>3</td>
<td>34</td>
<td>203</td>
</tr>
<tr>
<td>Int’l</td>
<td>3</td>
<td>223</td>
<td>19</td>
<td>82</td>
<td>910</td>
</tr>
<tr>
<td>Pacific</td>
<td>3</td>
<td>129</td>
<td>50</td>
<td>151</td>
<td>333</td>
</tr>
</tbody>
</table>

**DSpace Submits**
- Africa: 253
- Americas: 345
- Asia: 40
- Eurasia: 3
- Int’l: 3
- Pacific: 3

**RAMP Submits by Area Non-field Staff**
- Africa: 244
- Americas: 317
- Asia: 455
- Eurasia: 163
- Int’l: 223
- Pacific: 129

**RAMP Submits by Area Field Workers**
- Africa: 133
- Americas: 4
- Asia: 372
- Eurasia: 3
- Int’l: 19
- Pacific: 50

**RAMP Submits by Non-area Staff to Area Related Collections**
- Africa: 481
- Americas: 941
- Asia: 422
- Eurasia: 34
- Int’l: 82
- Pacific: 151

**All Submits**
- Africa: 1,113
- Americas: 1,607
- Asia: 1,289
- Eurasia: 203
- Int’l: 910
- Pacific: 333

International is indicated in the table above but not on the map because it does not have field workers, but rather consultants which act and publish with experience from a variety of contexts. Green numbers in this column reflect this difference.

**Lower Table:**
Green is ratio of RAMP submits to DSpace submits.

Yellow is administrative unit personnel submitting to collections specifically designated for that administrative unit.

Red is the percentage, measured off the whole, which were submitted by field workers.

---

**Percentage of total submits which were submitted via RAMP.**

<table>
<thead>
<tr>
<th>Administrative Area</th>
<th>DSpace by Admin</th>
<th>RAMP by Field Workers</th>
<th>RAMP by Non-area Staff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td>59%</td>
<td>33%</td>
<td>12%</td>
</tr>
<tr>
<td>Americas</td>
<td>59%</td>
<td>33%</td>
<td>12%</td>
</tr>
<tr>
<td>Asia</td>
<td>59%</td>
<td>33%</td>
<td>12%</td>
</tr>
<tr>
<td>Eurasia</td>
<td>59%</td>
<td>33%</td>
<td>12%</td>
</tr>
<tr>
<td>Int’l</td>
<td>59%</td>
<td>33%</td>
<td>12%</td>
</tr>
<tr>
<td>Pacific</td>
<td>59%</td>
<td>33%</td>
<td>12%</td>
</tr>
</tbody>
</table>

**Estimated percentage of RAMP submits by area personnel as opposed to non-area personnel contributing via RAMP.**

<table>
<thead>
<tr>
<th>Administrative Area</th>
<th>Non-area Personnel</th>
<th>Area Personnel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td>43.94%</td>
<td>25.44%</td>
</tr>
<tr>
<td>Americas</td>
<td>43.94%</td>
<td>25.44%</td>
</tr>
<tr>
<td>Asia</td>
<td>43.94%</td>
<td>25.44%</td>
</tr>
<tr>
<td>Eurasia</td>
<td>43.94%</td>
<td>25.44%</td>
</tr>
<tr>
<td>Int’l</td>
<td>43.94%</td>
<td>25.44%</td>
</tr>
<tr>
<td>Pacific</td>
<td>43.94%</td>
<td>25.44%</td>
</tr>
</tbody>
</table>

**Estimated percentage of RAMP submits by area personnel (who have submitted materials) classified as field workers.**

<table>
<thead>
<tr>
<th>Administrative Area</th>
<th>Field Workers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td>15.50%</td>
</tr>
<tr>
<td>Americas</td>
<td>0.32%</td>
</tr>
<tr>
<td>Asia</td>
<td>29.78%</td>
</tr>
<tr>
<td>Eurasia</td>
<td>1.50%</td>
</tr>
<tr>
<td>Int’l</td>
<td>5.86%</td>
</tr>
<tr>
<td>Pacific</td>
<td>15.15%</td>
</tr>
</tbody>
</table>

Though Americas Area leads SIL in total contributions to the archive, it also leads the areas in not encouraging its field workers to submit content directly to the archive via RAMP.
The tool or the fit

What is the experience that archives are designing for their constituents?

The tool is simply a component of the eco-system designed to involve users in a particular experience.

Fit is the way that a tool interacts with the entire eco-system. Fit is not just the relationship of the tool to any other part of the eco-system, but also it is how the entire system breathes together to create needs and solutions for users. Pressures or benefits in one part of the system can drive users to use the tool less or more.

Inappropriate feel in the User Interface or insufficient detail to features can lead to bad report for a tool. Insufficient detail to the overall economy of the larger eco-system can lead to abandonment of the tool, even if it is well designed.

In user experience analysis we must be careful to not attribute faults of the tool to the fit, and vice versa. Because of the organizational economics of archiving, there is relatively little return on investment for linguists to archive.
Archives, a dispensable service

The core business of any archive is the marketing of its relevance, often via content promotion and curation services, to both submitters and content users. The more it can convince each group of its value, the more valued it becomes in the eco-system.

Unlike most two sided markets (Parker & Van Alstyne 2000), the interaction over content (almost exclusively) happens asynchronously.

- Do language projects and programs, or future programs in the same language community benefit from archived resources?
- Are there other ways for SIL projects to disseminate developed resources? Are these methods direct or indirect competitors to motivating factors for linguists to archive?
- Does the linguist even need the archive? (Assuming that the linguist’s only need to archive was to share files with colleagues, and Dropbox works, meeting present needs more efficiently.)
- Does the archive and the content it houses, serve the linguist, the archiving institution, or the various political interests of the communities? Who is the direct customer and who is the beneficiary of archiving services?
- Who is the one who manages the relationship between the archiving institution and the language community?
Hinderances to the successful two-sided market lie in the inter-constituent relationships in the social network

The archive's ongoing challenge is to assert its value to its constituency.

What is the Administrator's responsibility with respect to data as they manage the project in which the linguist works?

What is the data management responsibility of the administrator as they manage the relationship between the community and the organization?

---

Monday, March 11, 2013

---

Is the consultants' relationship with the minority language speaker/community a 2, 3 or 4 party relationship, with one of the parties being the organization?

Do parties in the relationship understand or see it this way? What is that effect on data management, or archiving?

---

Getting stuff can be great! But content usually travels along the lines of the social network.

---

I can't contribute back even if I originally got the data from the archive, because I am not SIL

---

I might work with an SIL colleague.

---

I could contribute to DSpace, but why? I get nothing out of it...

---

Getting stuff can be great! But why am I interested?

---

I might hire a student, but then I am going to handle all the data exchange.

---

I might work with a student at a school, but then is the data excluded because we are working at the school?

---

I can't contribute to DSpace, but why? I get nothing out of it...

---

Getting stuff can be great!

---

Why should I look at the SIL.org archive site, because my role as a consultant is to bring in solutions which are not already known or available. To do this I use a wide variety of resources which are external to the organization.
Conclusions

- The challenge is not creating a tool, but rather a tool which fits the frame of reference of linguists and monopolizes on metadata created by linguists at the time of object use or object creation.
- Archives have relatively little that persuades linguists to archive; this power resides with the project funders. (But even then, there is often no way to revoke funding if the project is not archived.)
- Archives have the power to entice linguists to submit data, but the power of this enticement resides with user interaction design.
- The current user group of RAMP is not the primary intended user group. But the SIL archive is happy to see an increase in accessions.
- RAMP has seen uptake in use by archivists but not by field linguists. Archivists are happy to answer the questions asked by the application. RAMP has not seen wide adoption by field linguists. This is either because field linguists don’t see the value in archiving, or the time required to use the application is not justified for furthering the linguist’s ends.
- There is no significant return on investment for time spent to archive materials in the current fit between language program execution and the activity of submitting materials to the archive.
- Archiving via RAMP is still perceived as an end of project task.
- There is a high degree of probability that the DSpace UI dissuades non-publishing staff and non-archivists from using it for submissions.
- There is also evidence that the RAMP interface may also be having a dissuading effect among field linguists, but the statistical evidence is inconclusive.
- In SIL, implementations of archiving policy very greatly.