

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

Hugh Paterson III & Jeremy Nordmoe



RAMP

1 The Archiving Experience

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

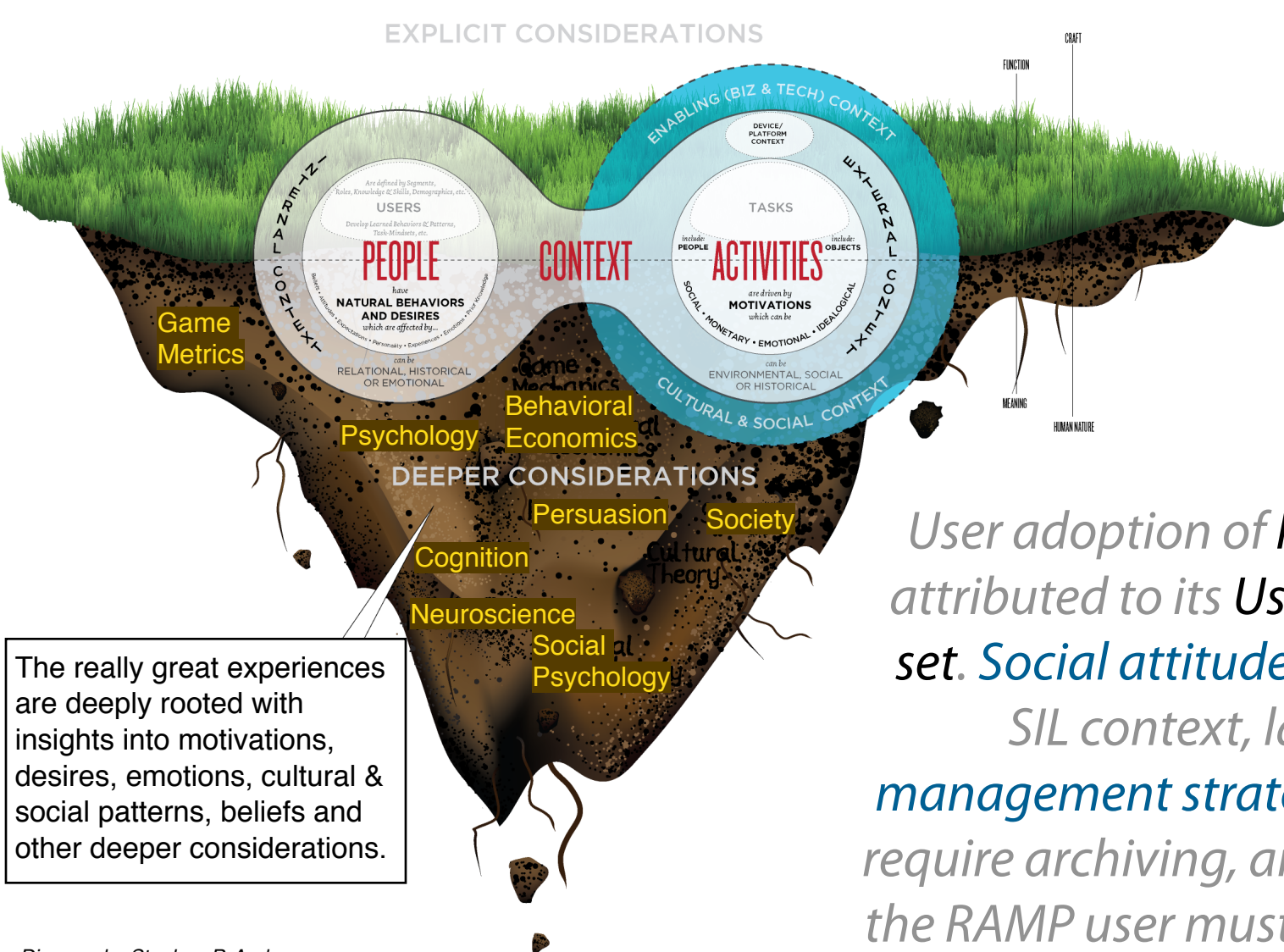
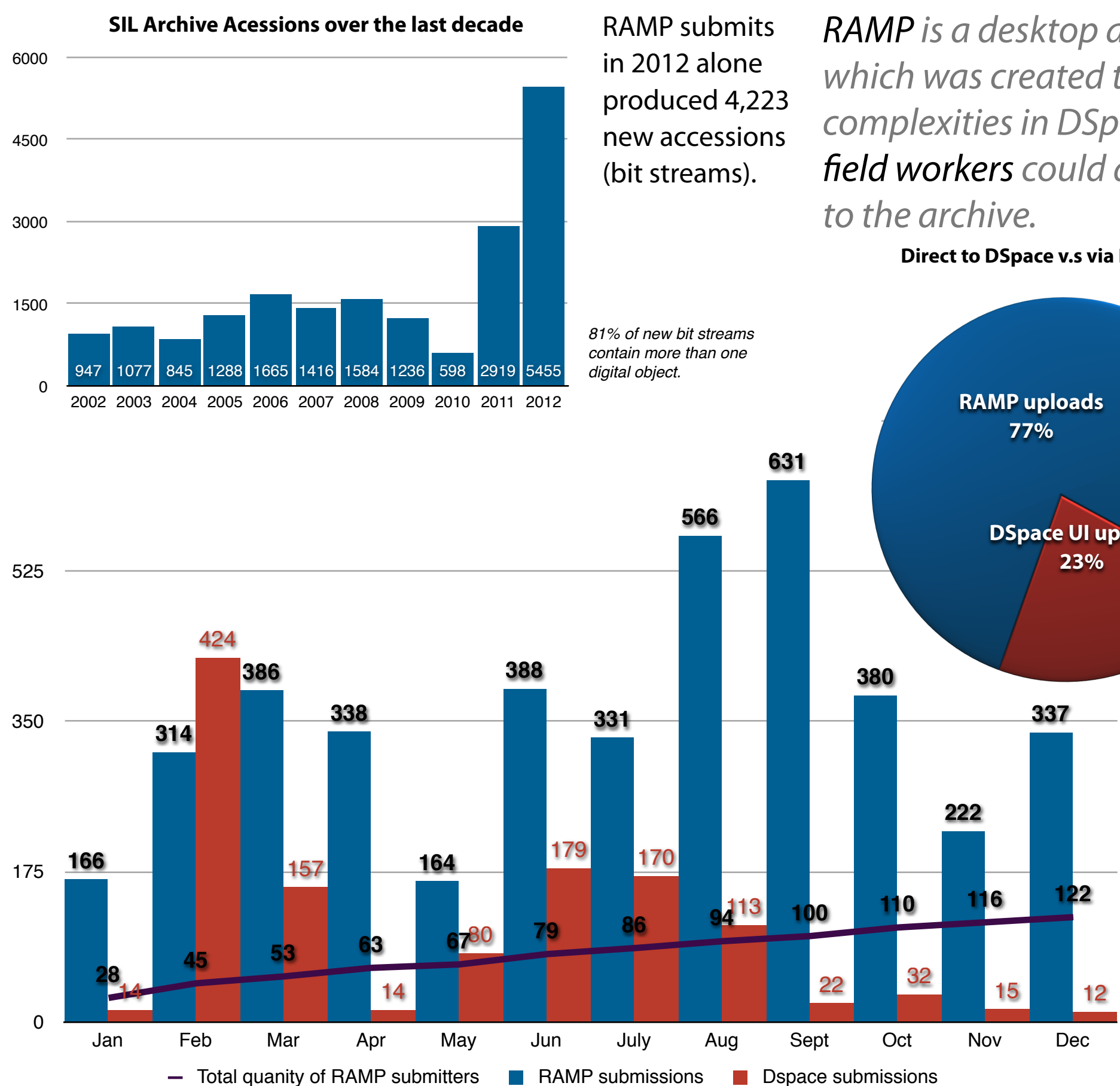


Diagram by Stephen P. Anderson

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 RAMP "effect"

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



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

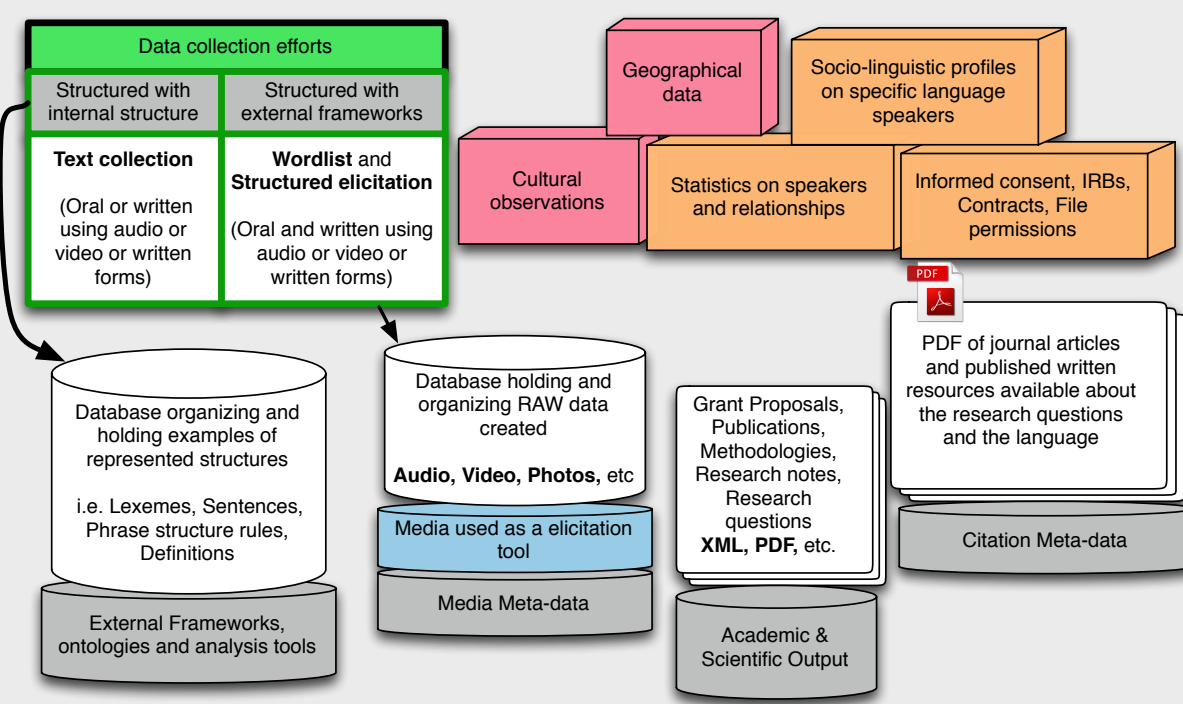
2 Social Attitudes of Linguists

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?*

Other attitudes encountered:

- The archiving institution:
 - Losses 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."



From when do these items come?

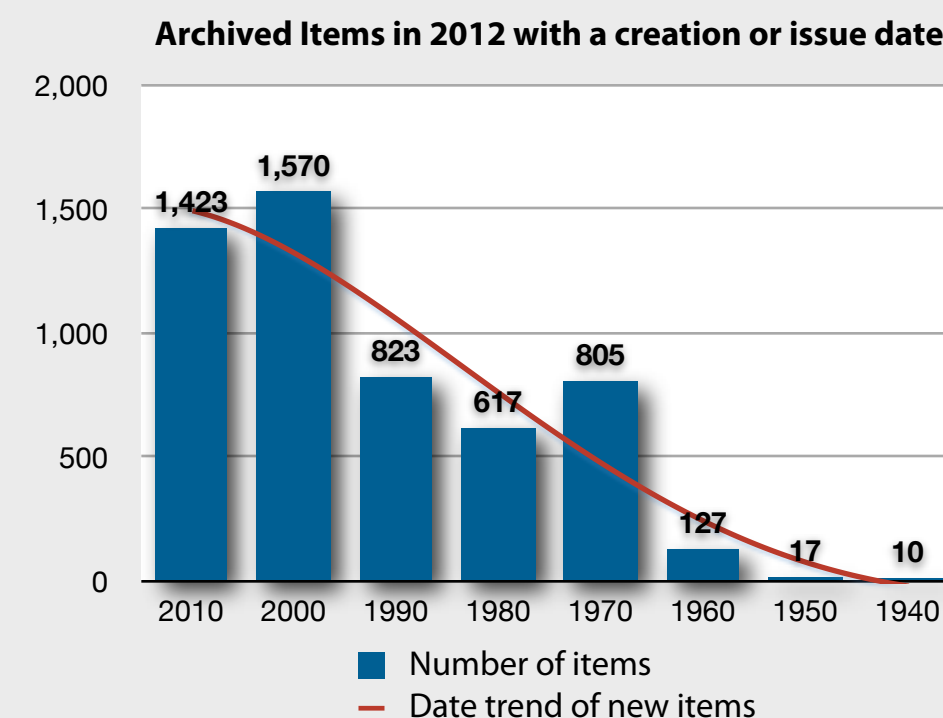
604 items archived in 2012 are known to have been created or published in 2012.

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.

Are items being accessioned in appropriate amounts of time? Or, do linguists retain the attitude: *archiving is my last task before death?*



3 Task perception

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?

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.

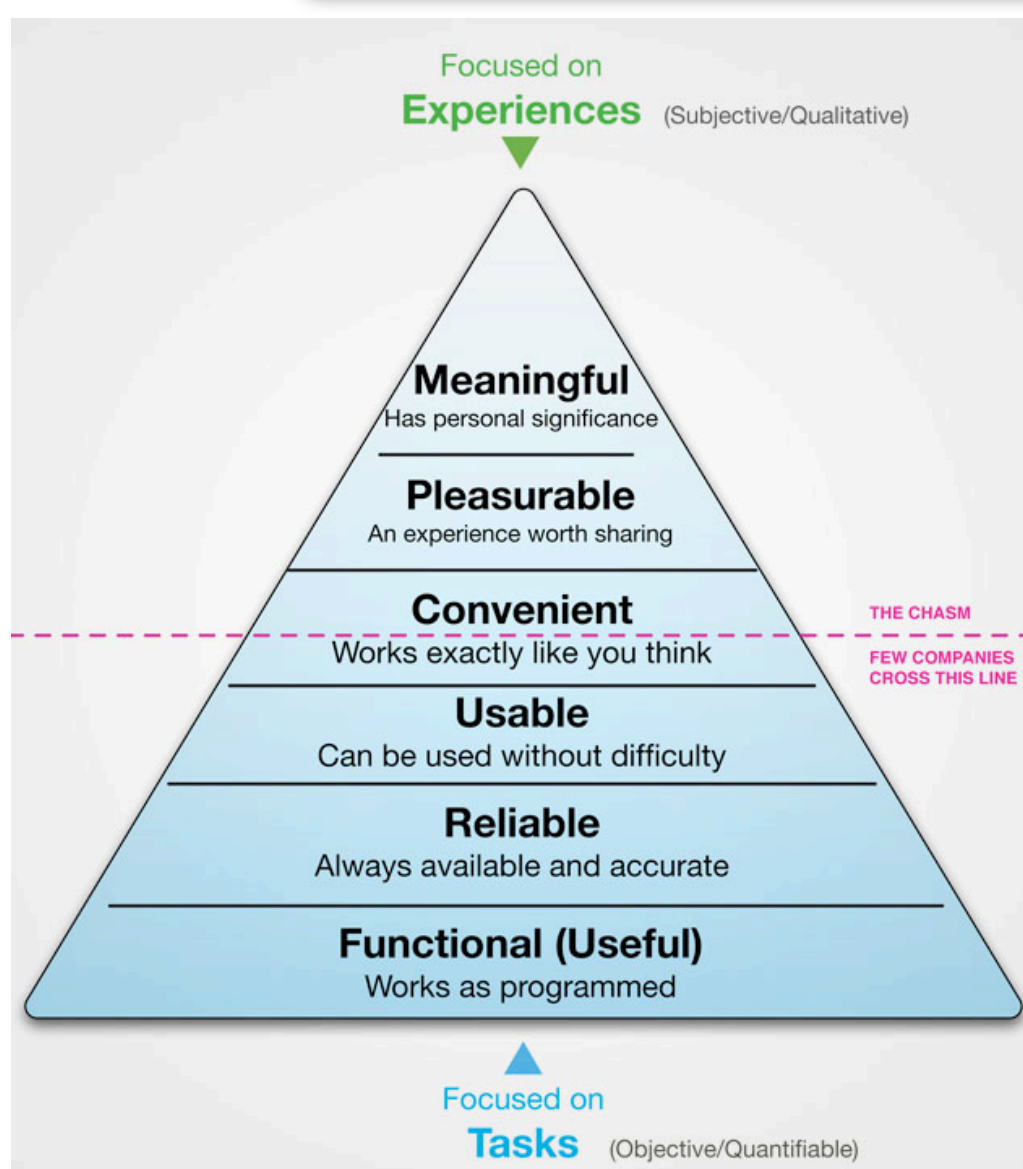
Archivist:

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.

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.



"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.

What kinds of digital objects are being submitted?

Types of items submitted in 2012	Total number of items across all submissions
Textual based objects (presentations, papers, PDFs)	5,598
Image based objects (psd, jpg, raw, tiff)	2,104
Unknown (obscure object formats, fonts, ISOs, zip)	1,484
Audio based objects (mp3, aif, wav)	1,003
Text-Data based objects (toolbox files, FLEX, xls)	67
Web formatted files (html, css)	45
Video based objects (.mov, .webm, .mp4)	32
Total Digital Objects	10,333
60% of images may be part of text based scanning of old documents to archival .tiff formats.	
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.)	

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

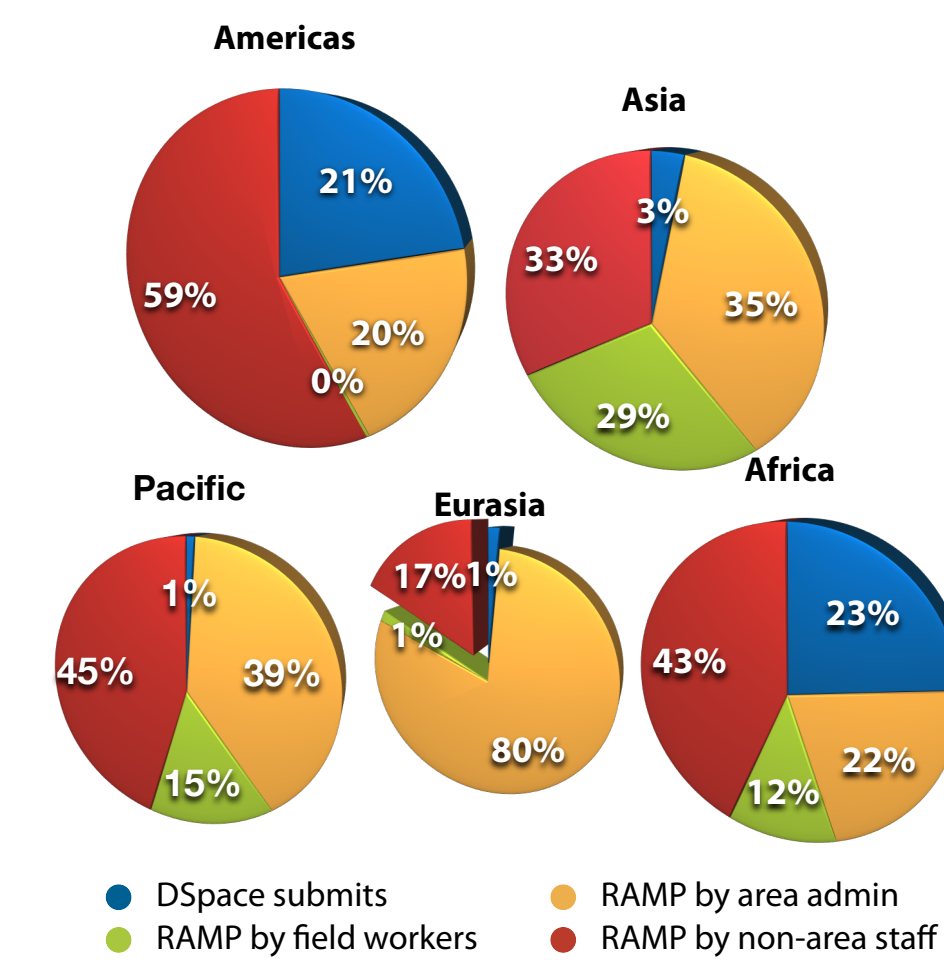
To assess, this an archive might look at how many .zip files and archive type files (.iso, .tar, .gzip, etc.) it might have accumulated. This year the SIL archive added 1,030 new .zip files. Zip files may be a reasonable transmission or storage format, but if the reason for the submission in a .zip format is because the submitter didn't want to take the time to archive each digital object independently, when the time 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.

4 Management Strategies

Submission methods used



	Africa	Americas	Asia	Eurasia	Int'l	Pacific
DSPACE submissions	255	345	40	3	586	3
RAMP submits by area non-field staff	244	317	455	163	223	129
RAMP submits by area field workers	133	4	372	3	19	50
RAMP submits by non-area staff to area related collections	481	941	422	34	82	151
	1,113	1,607	1,289	203	910	333

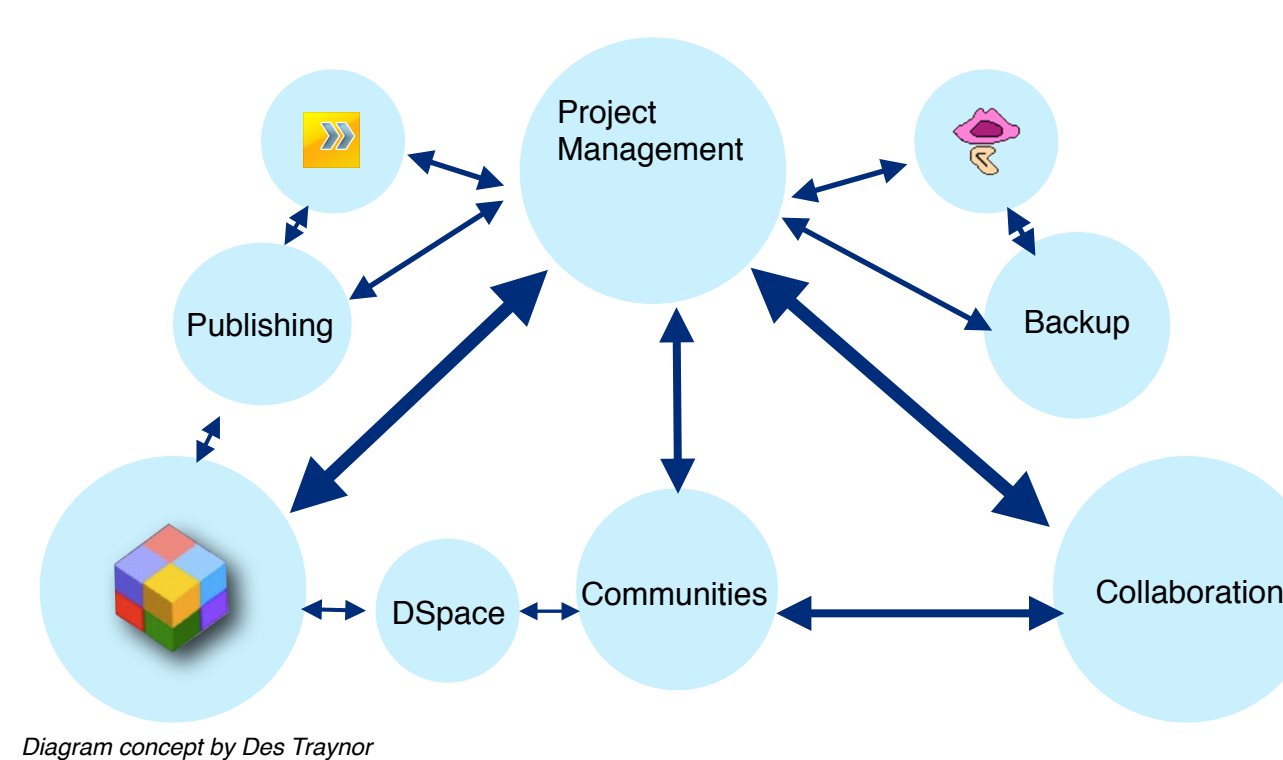
Each administrative area of SIL has different strategies for archiving content. These management strategies affect which tools are presented to various sets of linguists and therefore also who does the work related to submission to the archive.

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.

Percentage of RAMP usage by management region measured off of the collections	Africa	Americas	Asia	Eurasia	Int'l	Pacific
	77.09%	78.53%	96.90%	98.52%	35.60%	99.10%
Estimated percentage of RAMP submits by area personnel as opposed to non-area personnel contributing via RAMP	43.94%	25.44%	66.21%	83.00%	74.69%	54.24%
Estimated percentage of RAMP submits by area personnel (who have submitted materials) classified as field workers	15.50%	0.32%	29.78%	1.50%	5.86%	15.15%
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.						

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



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

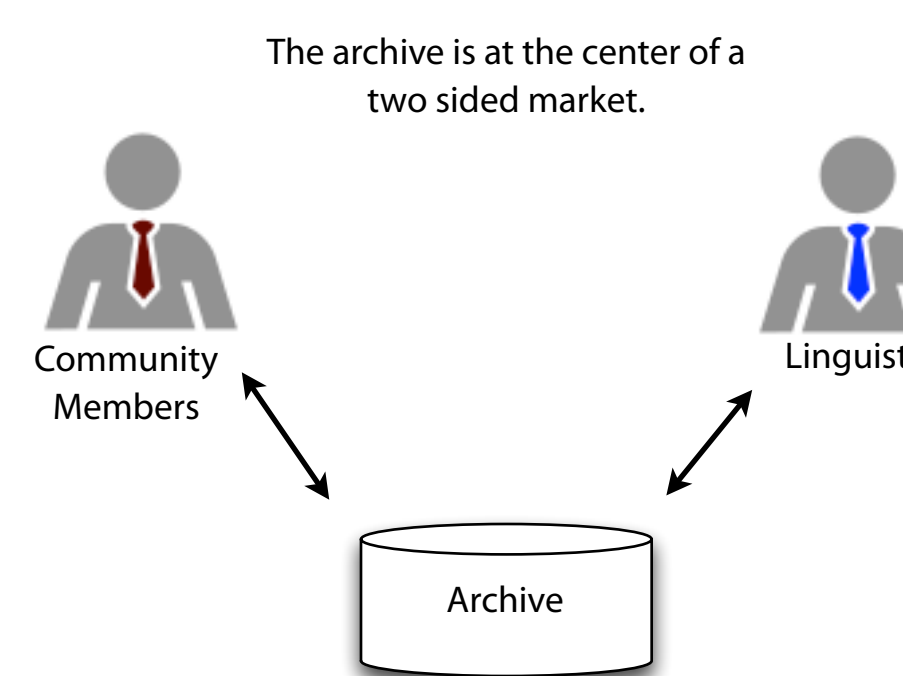
In user experience analysis we must be careful to not attribute faults of the tool to the fit, and vice versa. Because of the organizational economics of archiving, there is relatively little return on investment for linguists to archive.

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.

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.



In SIL the challenge for the archive is to prove its worth to several different user groups.

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?

5 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.

Comments invited via WordPress or YouTube:

<http://bit.ly/UOJ6UI>

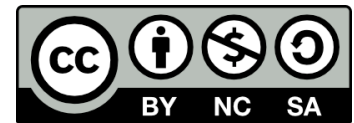
Hugh.Paterson@sil.org & Jeremy_Nordmoe@sil.org

References

- Anderson, Stephen P. 2006. Creating Pleasurable Interfaces: Getting from Tasks to Experiences (Poster). <Accessed: 18 February 2013> <http://www.artificialindustry.com/uploads/inspire/1255515166from-task-to-experience.png>
- Anderson, Stephen P. 2009. Fundamentals of Experience Design (Poster). <Accessed: 18 February 2013> <http://www.postpainter.com/thoughts/files/Fundamentals-of-Experience-Design-stephen.pdf>
- Anderson, Stephen P. 2011. Seductive interaction design: creating playful, fun, and effective user experiences. Berkeley, CA: New Riders.
- Gray, Jeremy R., Todd S. Braver & Marcus E. Raichle. 2002. Integration of emotion and cognition in the lateral prefrontal cortex. Proceedings of the National Academy of Sciences 99: 4115-20. <http://www.pnas.org/content/99/16/4115.abstract>
- Nash, Douglas. 2012. Businessperson. The Neuron Project.
- Nordmoe, Jeremy. 2011. Introducing RAMP: an application for packaging metadata and resources offline for submission to an institutional repository. In Proceedings of Workshop on Language Documentation & Archiving 18 November 2011 at SOAS, London. Edited by David Nathan. p. 27-32. <Accessed: 18 February 2013> <http://www.sil.org/acpub/abstracts/249585>
- Parker, Geoffrey & Marshall W. Van Alstyne. 2000. Information Complements, Substitutes, and Strategic Product Design (November 8, 2000). Available at SSRN: <http://ssrn.com/abstract=249585>
- Traynor, Des. 16 January 2012. Copy the fit, not the features. <Accessed: 18 February 2013> <http://blog.intercom.io/copy-the-fit-not-the-features/>

Suggested Citation:

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. Version 1.4



This work is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/3.0/us/> or send a letter to Creative Commons, 171 Second Street, Suite 300, San Francisco, California, 94105, USA.