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:
http://bit.ly/UOJ6UI
Hugh.Paterson@sil.org & Jeremy_Nordmoe@sil.org

1

Scope and Challenge

- SIL has nearly 80 years of history working with minority language communities.
- About 1 million relevant non-digital objects are estimated to exists 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: 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.

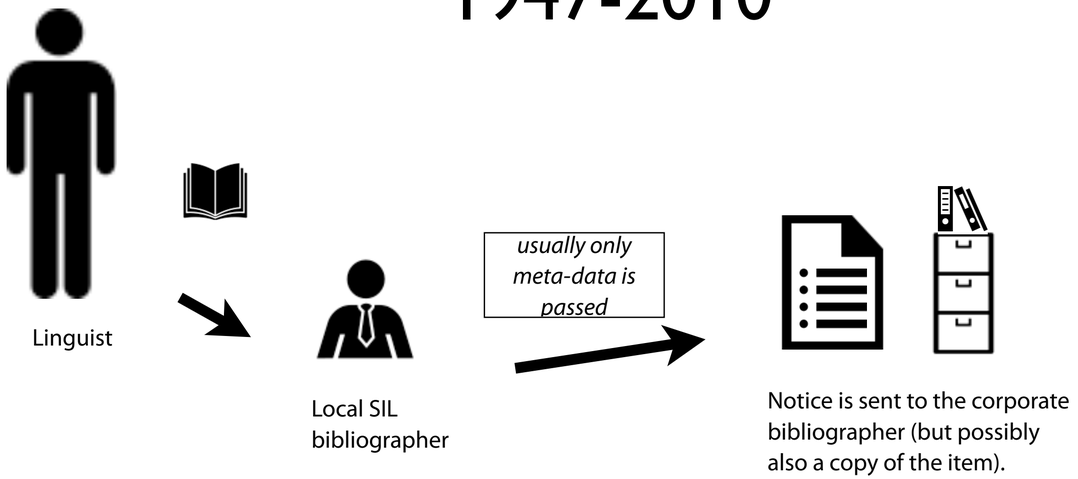
RAMP is simple compared to DSpace. - True

But do people really want to use RAMP?

RAMP and the SIL archive submission process

Previous models of submissions in SIL

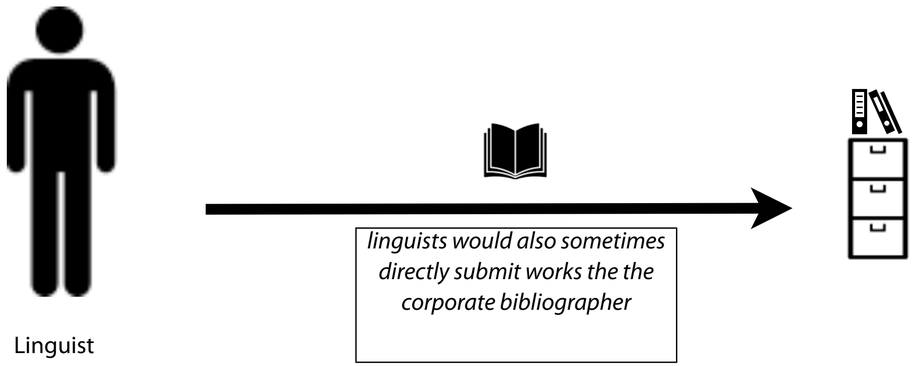
1947-2010



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

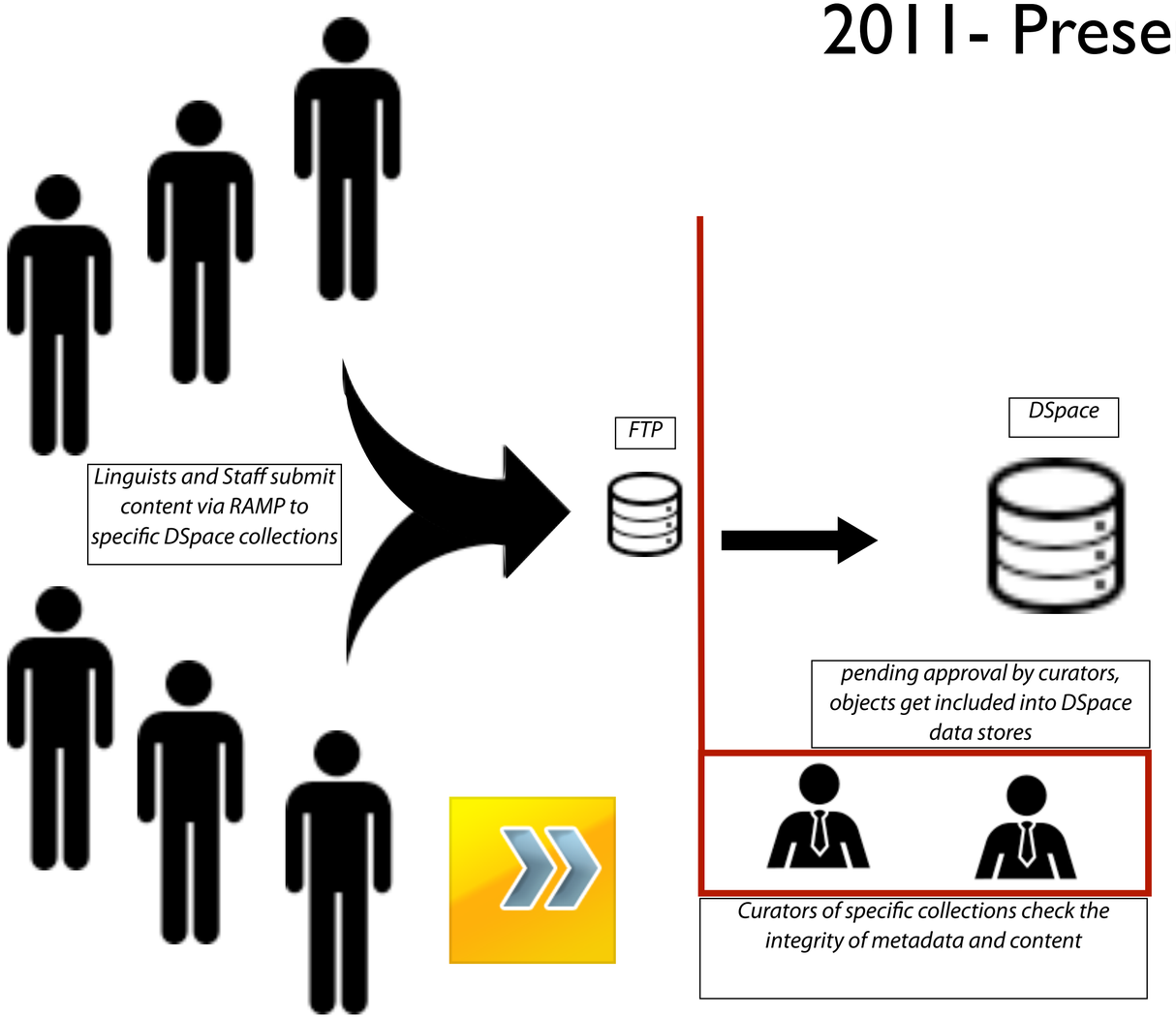
Local "archives" are functionally operated as private libraries.

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



The RAMP model of submissions

2011- Present

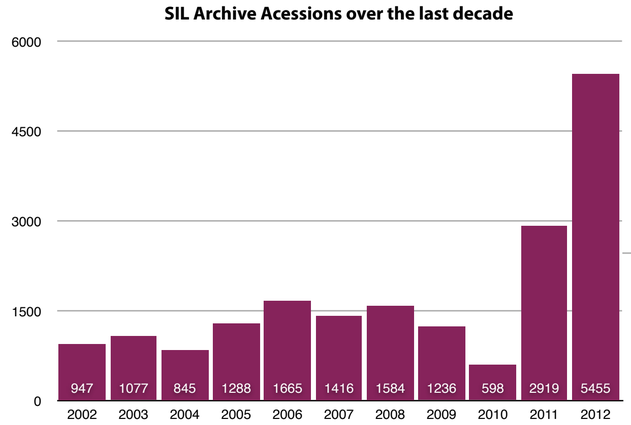


Content is then piped to appropriate access points to serve various communities

2013 marks the end of the SIL Bibliography *and the beginning of the online presence of the* SIL archive.

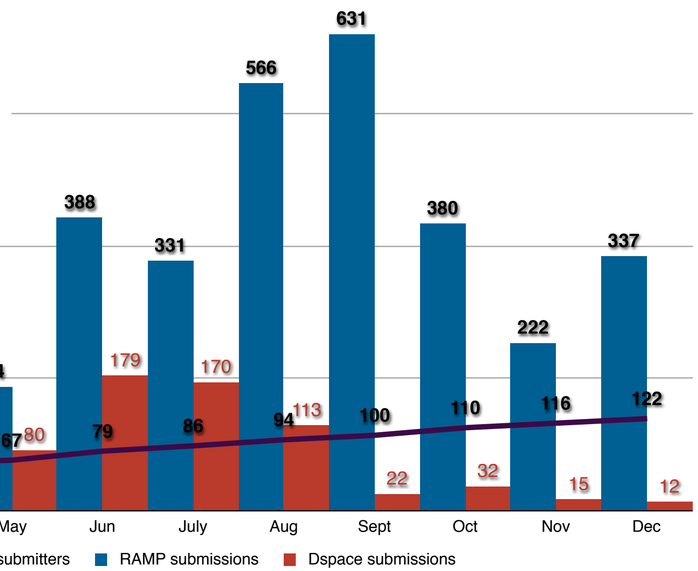
The “RAMP” Effect

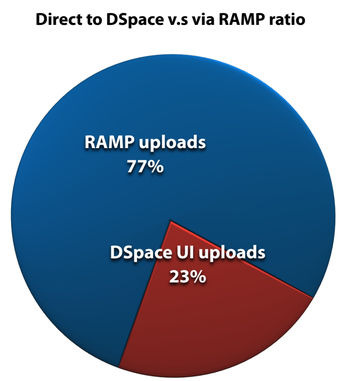
Prior accession rates over the last 10 years have averaged between 1,500 and 2,000 items per year.



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

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





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

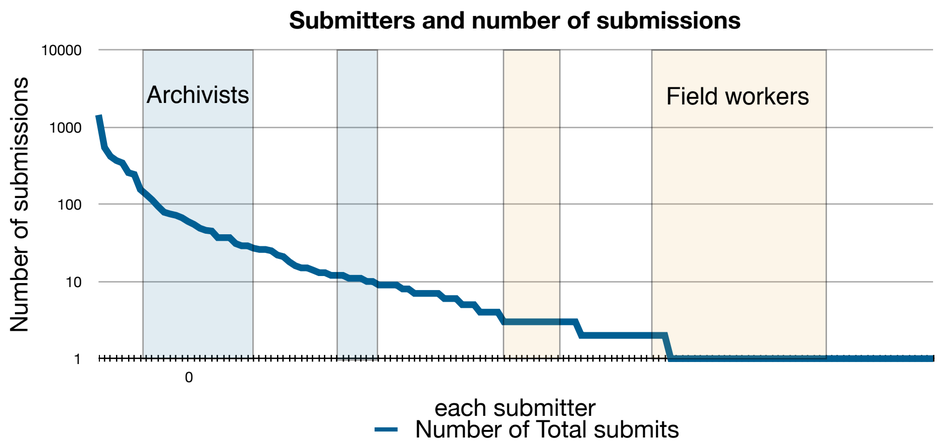
RAMP contributed an increased capacity (200%) to accession materials, but was it equally successful in reaching its intended audience?

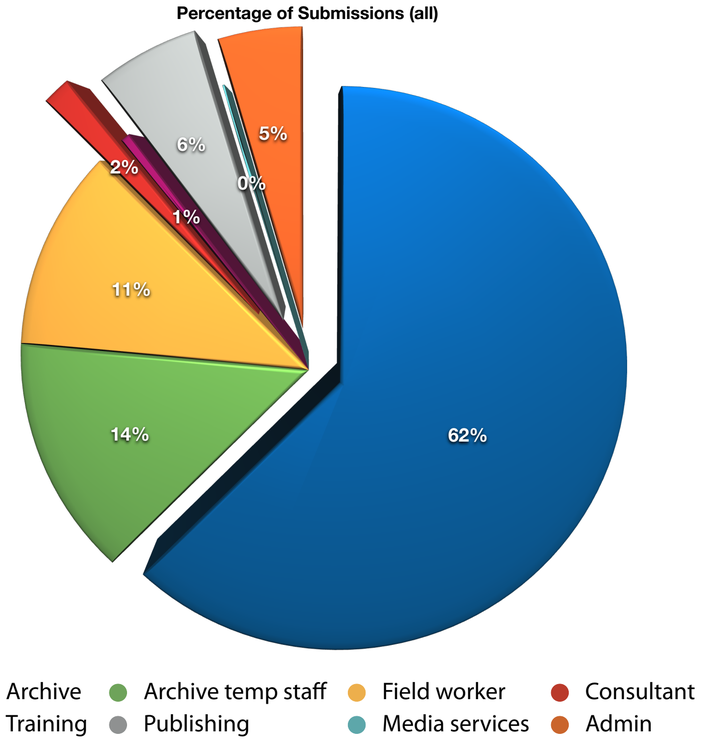
So *who* is making submissions

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?

Number of RAMP submissions	Number of bit streams	Number of people	Class of SIL staff submitting a work	Percentage of total submissions	Percentage of total submitters
2,341	3409	36	**Archive**	62.49%	25.53%
734	762	2	**Archive-temp**	13.97%	1.42%
596	610	50	**Field worker**	11.18%	35.46%
83	86	23	**Consultant**	1.58%	16.31%
2	30	4	**Training**	0.55%	2.84%
224	306	14	**Publishing**	5.61%	9.93%
5	6	2	**Media services**	0.11%	1.42%
238	246	10	**Admin**	4.51%	7.09%
Total number of submissions	5455	141	Total submitters of any kind		
Percentage of RAMP submissions submitted by field workers 14.1%	**RAMP packages**	**4223**			
	DSpace Uploads	1232			





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%.

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.

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.

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.

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.

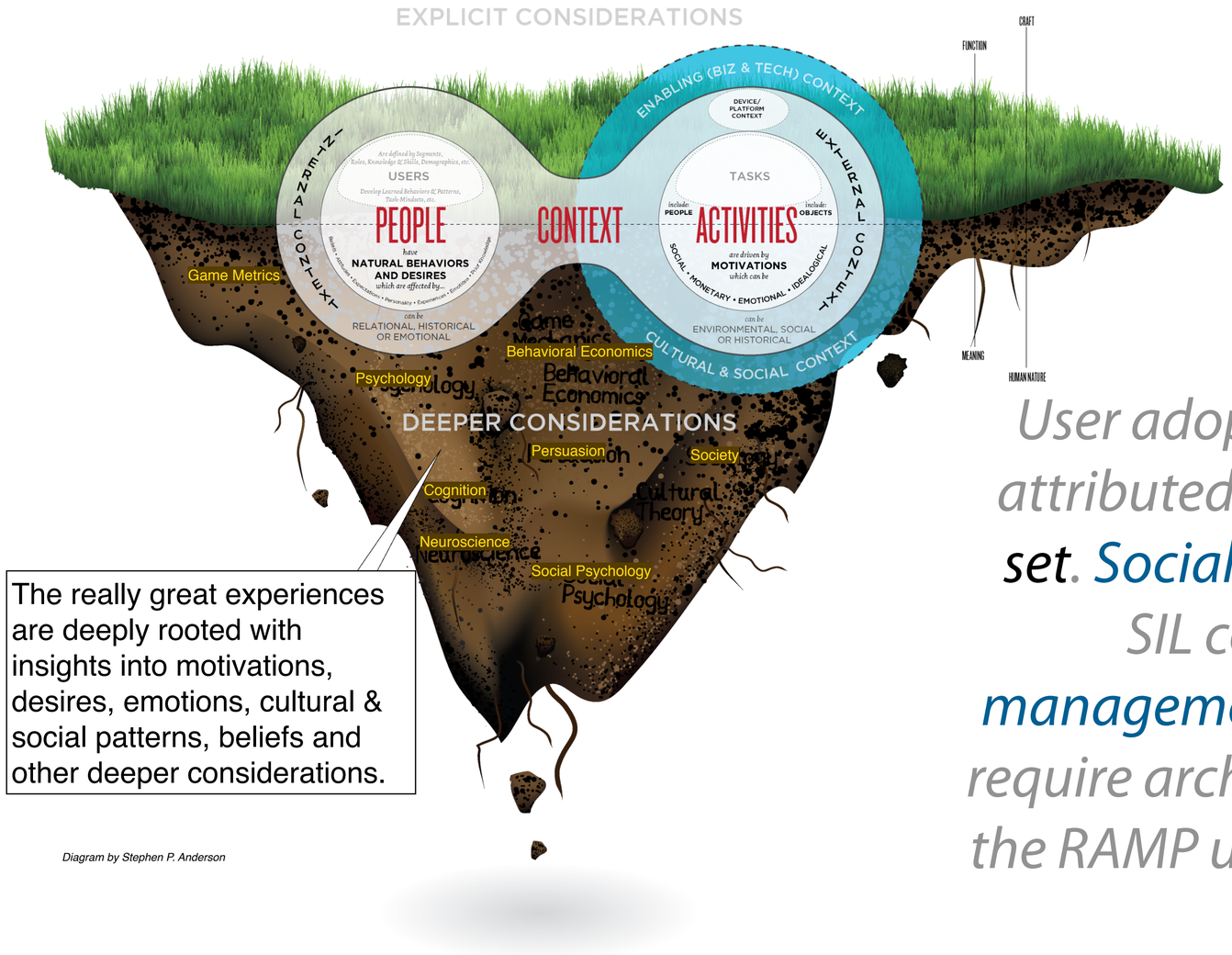
Of the top 12 repeat DSpace submitters, 10 of them have roles in archiving or publishing. Others were, 1 each: training, and fieldworker. - *Those SIL archivists who use DSpace use it a lot (or for batches)*.

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..."



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

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.*

2

Social Attitudes of Linguists

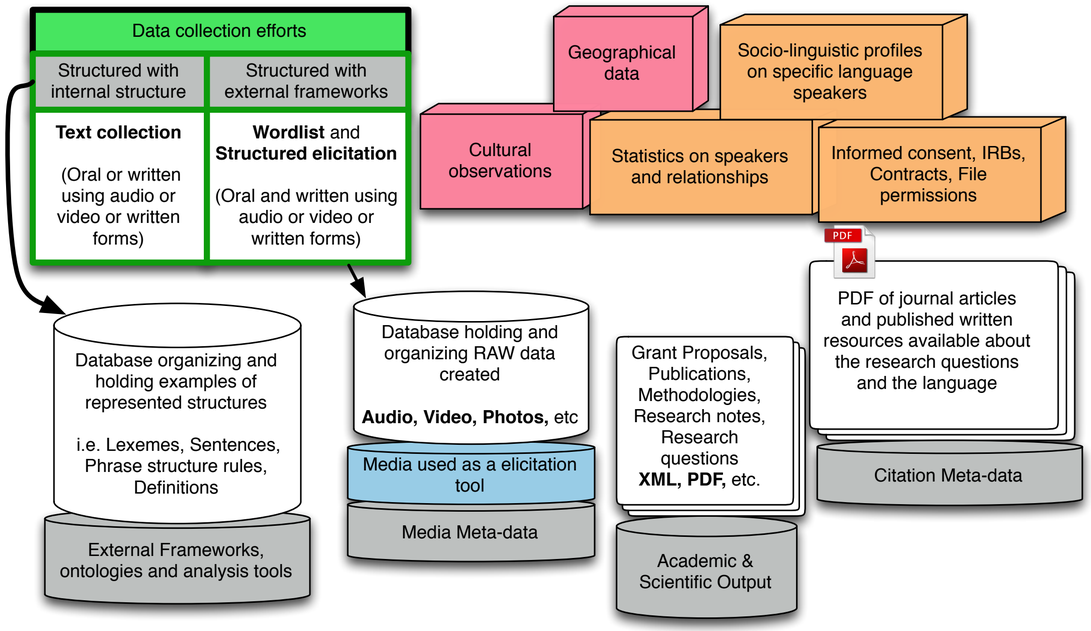
Nordmoe (2011) claims that archiving meta-schemas remain too complex for linguists… We find this objection un-grounded coming from linguists who devise meta-schemas for for describing language… (Though we make no claim that any schema is innate).

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?*

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:

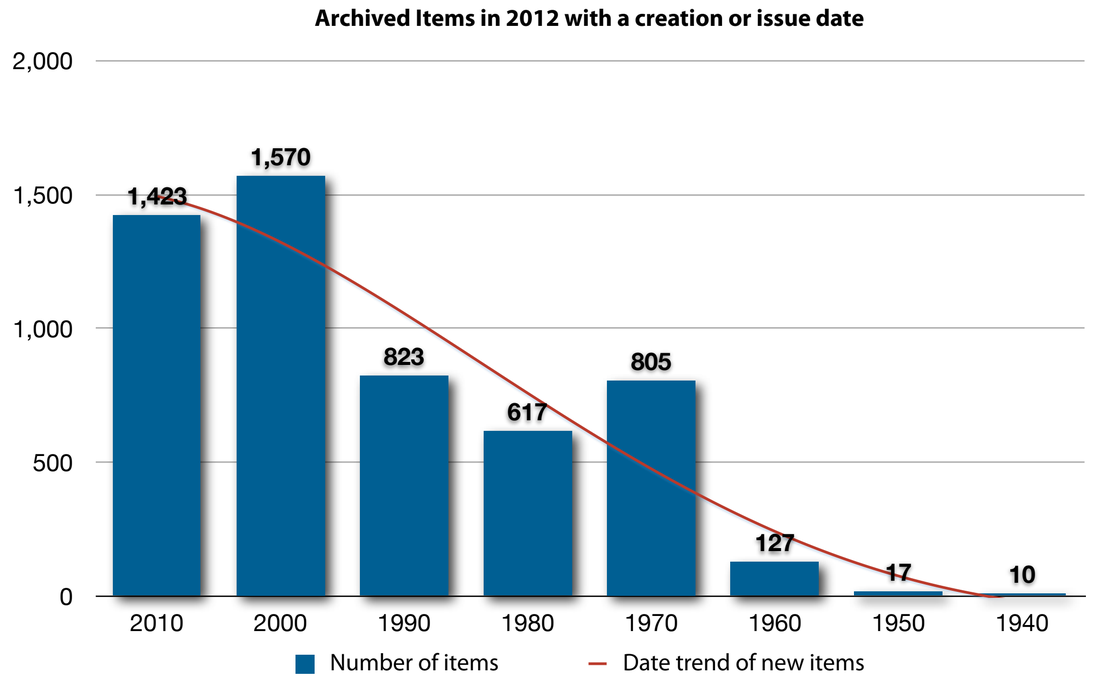
- The archiving institution:
  - loses content and materials
  - can/will not restrict content access appropriately
  - can not publish content to open access points in a timely manner
  - does not value certain types of content or will charge for access
  - does not maintain accurate records because 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."



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?*

3

Task perception

Linguist:

How does RAMP relate to my other data in my workflow?

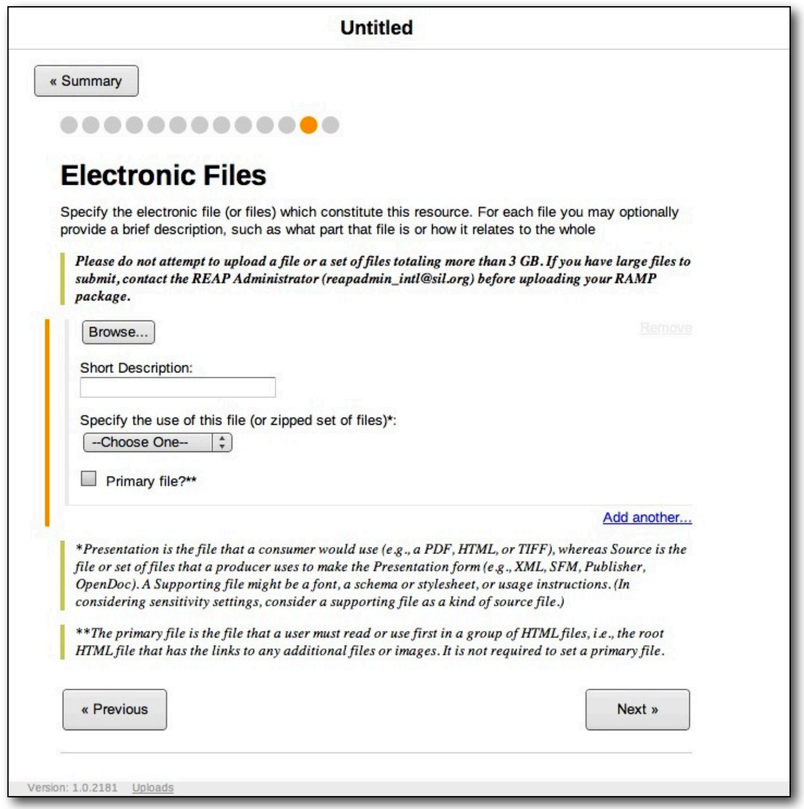
How does RAMP enable me to keep the promises I made for funding?

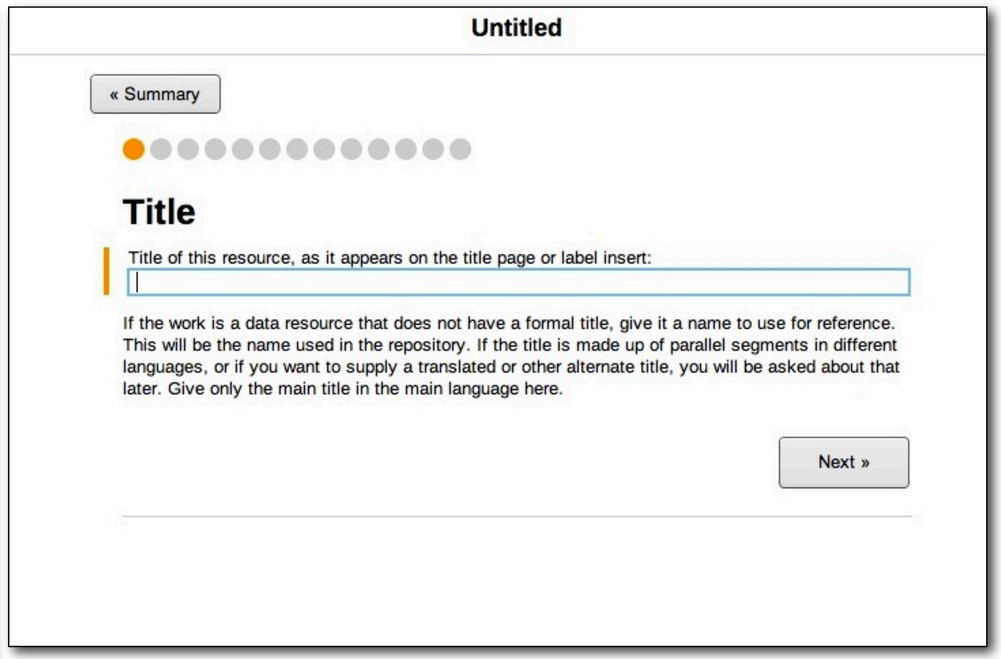
Archivist:

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.*

Task perception becomes a major issue in user interface design. The user interface also has a major role insetting the mood for the entire interaction.
An archivist wants to know what the object is that the submitter has.

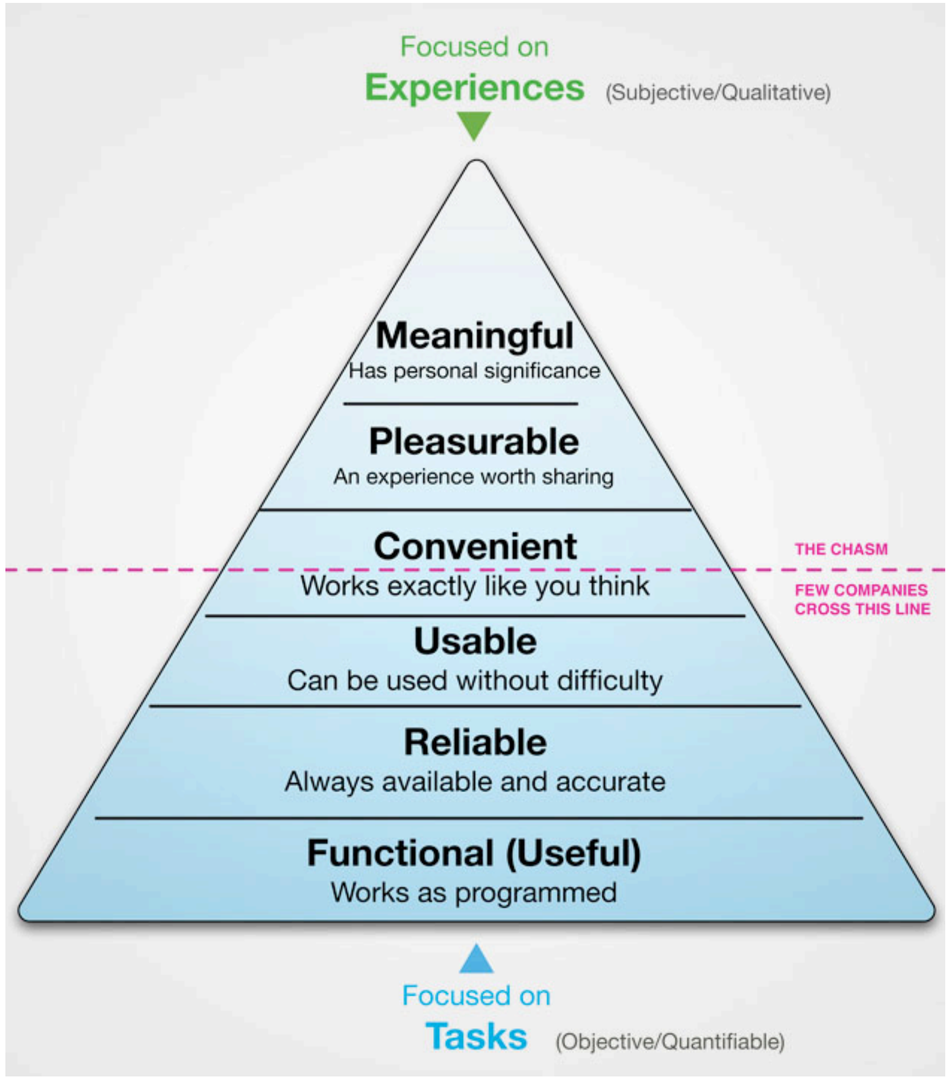




Which screen should come first?

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



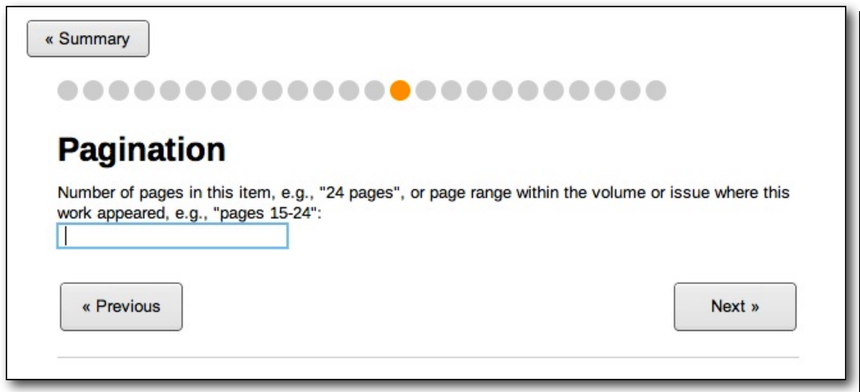
"Emotion and cognition conjointly and equally contribute to the control of thought and behavior." (Gray 2002) Often the design of linguistics based software is focused on specific tasks, not creating meaningful experiences.

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?

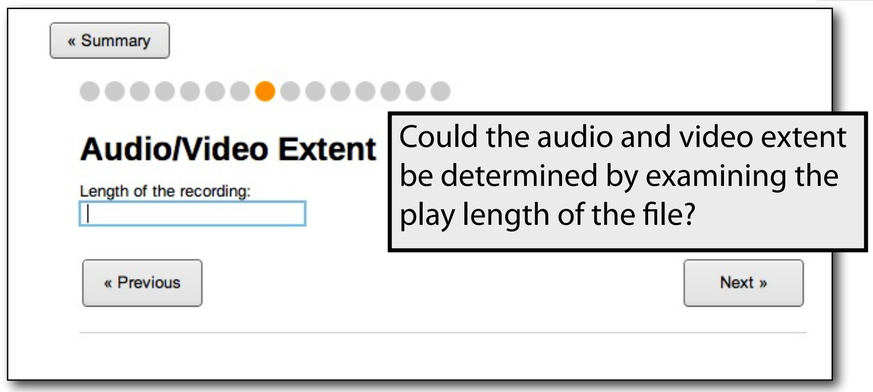
What is the emotional impact on the RAMP user's attitudes towards archiving? Is it the meta-schema which is too difficult or is it the relationship through the software? - To the RAMP user, is the experience worth repeating and telling their friends about?

Working with existing metadata

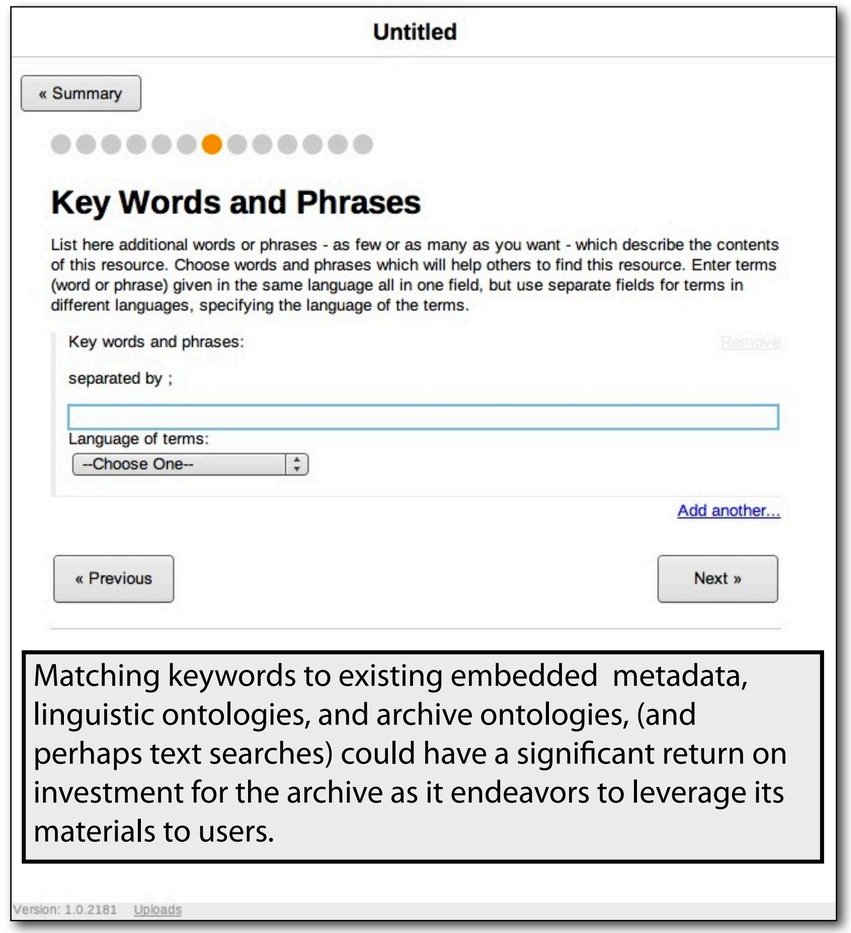
Most media files have metadata of some kind embedded in them. Working with this metadata (and allowing the user to verify it) rather than completely ignoring it would save the user effort, and in some cases time. The perceived gain would make the user appreciate the software more. They feel like they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



Pagination in text based documents (.doc, docx, PDF, multi-page .tiff) can be programmatically determined, so why ask the user?
The asking of non-pertinent questions is perceived to be time wasting.

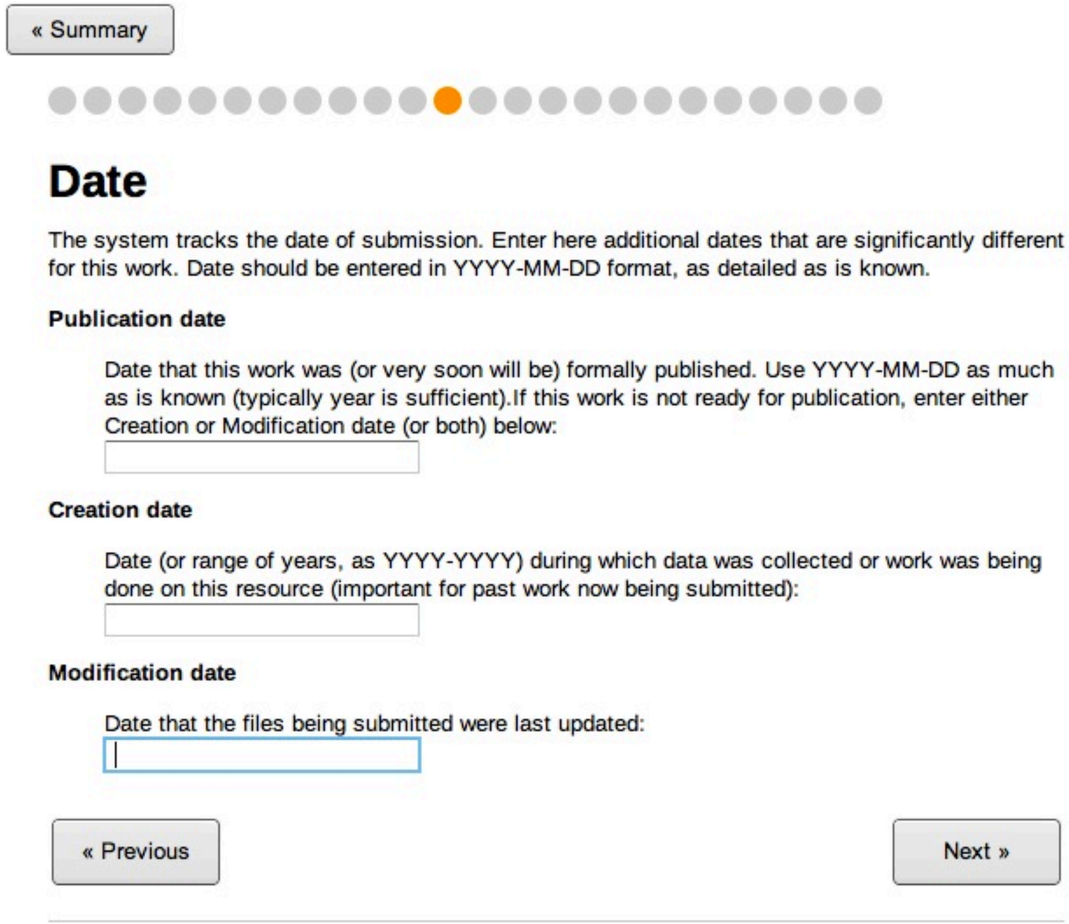


Could the audio and video extent be determined by examining the play length of the file?



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?

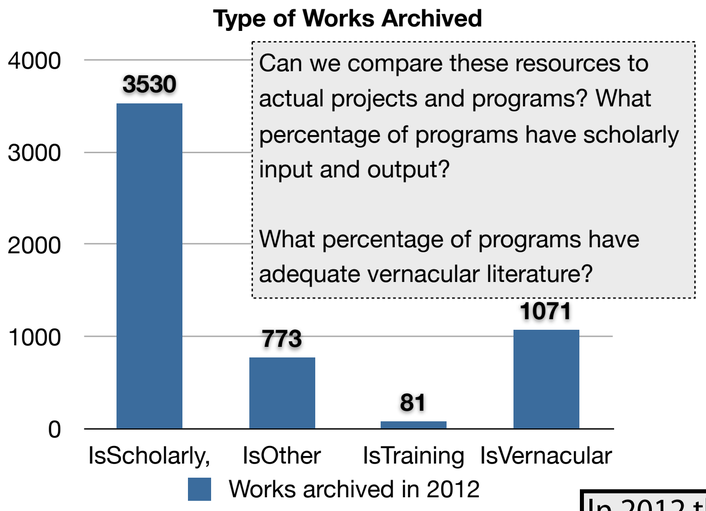
Preparing through embedding

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?



Can we compare these resources to actual projects and programs? What percentage of programs have scholarly input and output?

What percentage of programs have adequate vernacular literature?

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, .aiff .wav)	1,003
Text-Data based objects (toolbox files, FLEx, .xls)	67
Web formatted files (html, css)	45
Video based objects (.mov, .vob, .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 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.

4

Project Management

There are there are three issues which affect radically project management, two of these have drastic impact on how and where the out put 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*

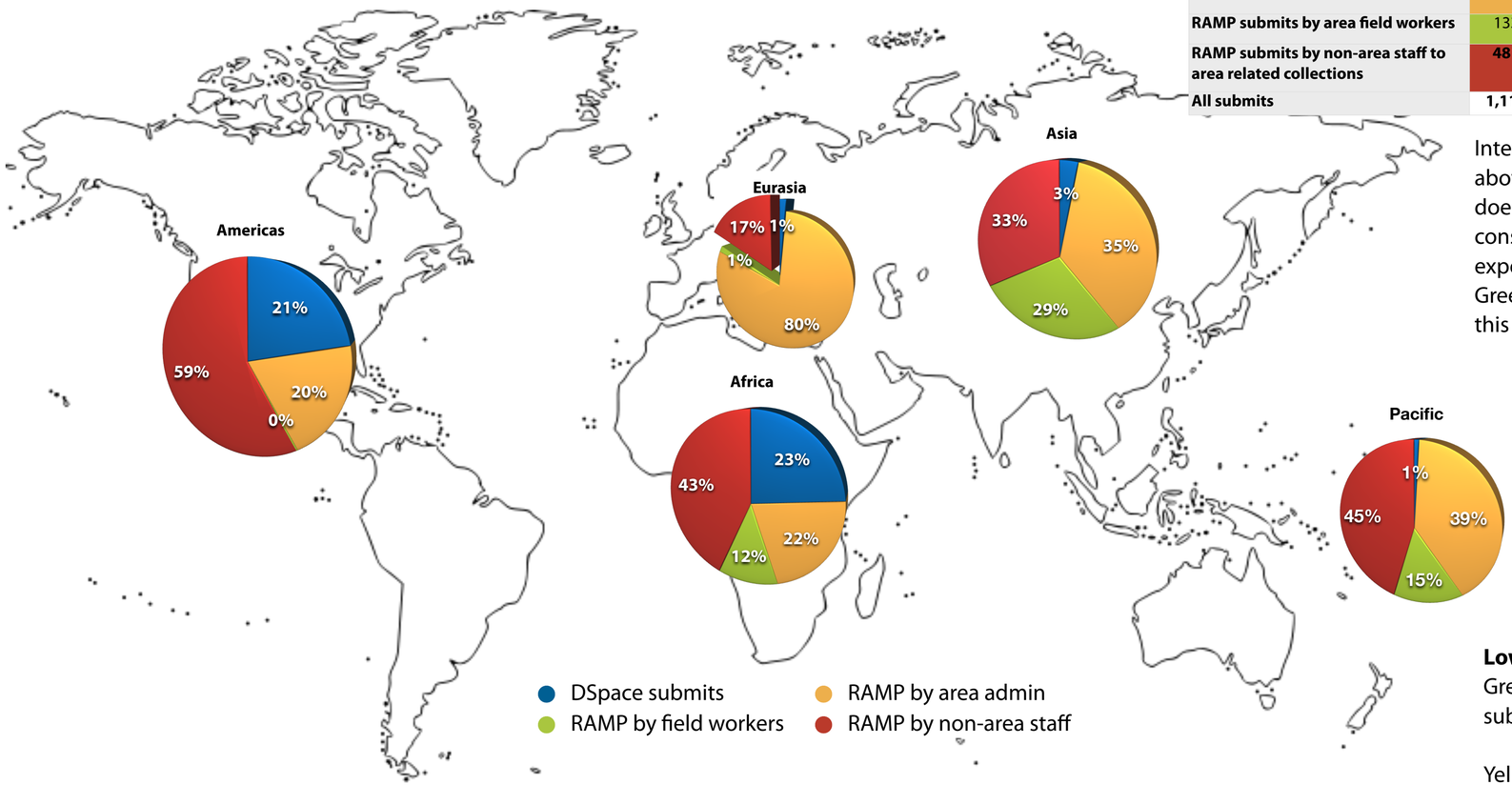
Funders of the project *- often put requirements on projects to be archived*

Institutional relationships *of the researchers often dictate where a researcher can or will not archive.*

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**.



	Africa	Americas	Asia	Eurasia	Int'nl	Pacific
DSpace submits	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**
All submits	**1,113**	**1,607**	**1,289**	**203**	**910**	**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.

	Africa	Americas	Asia	Eurasia	Int'nl	Pacific
Percentage of total submits which were submitted via RAMP.	77.09%	78.53%	96.90%	98.52%	35.60%	99.10%
Estimated percentage of RAMP submits by area personnel as apposed 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.

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.

The tool or the fit

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

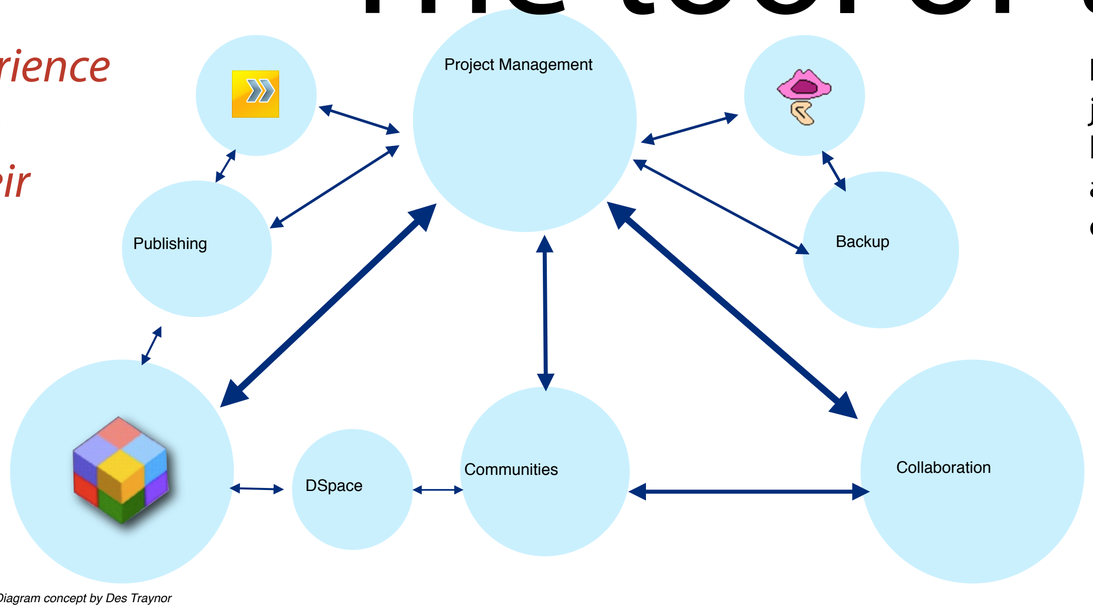


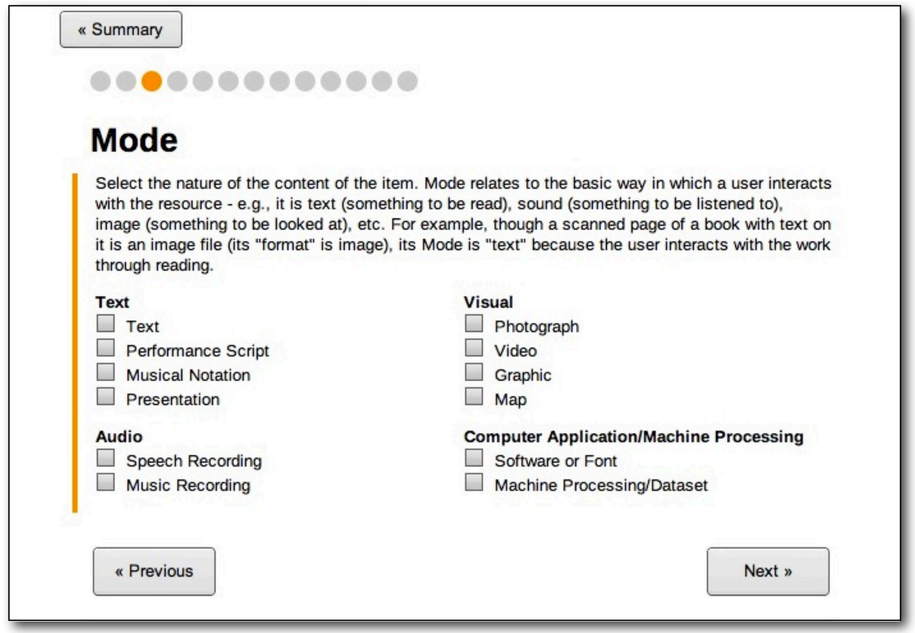
Diagram concept by Des Traynor

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

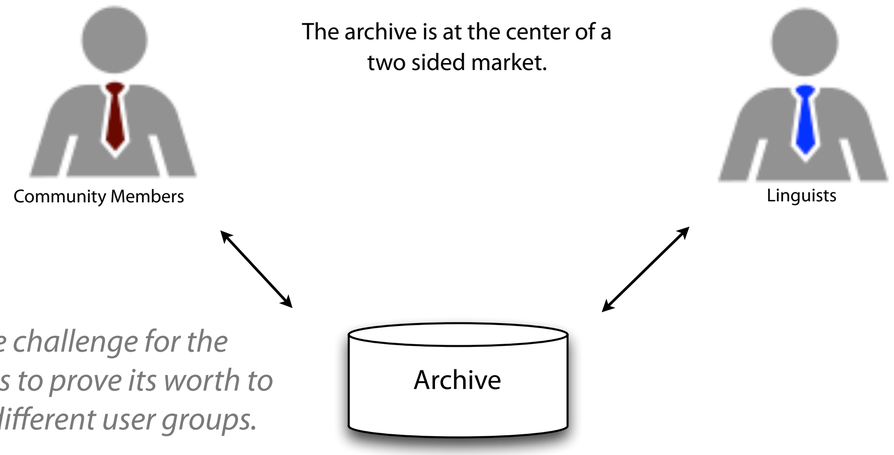
In user experience analysis we must be careful to not attribute faults of the tool to the fit, and vice versa. Because of the organizational economics of archiving, there is relatively little return on investment for linguists to archive.

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 breaths 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



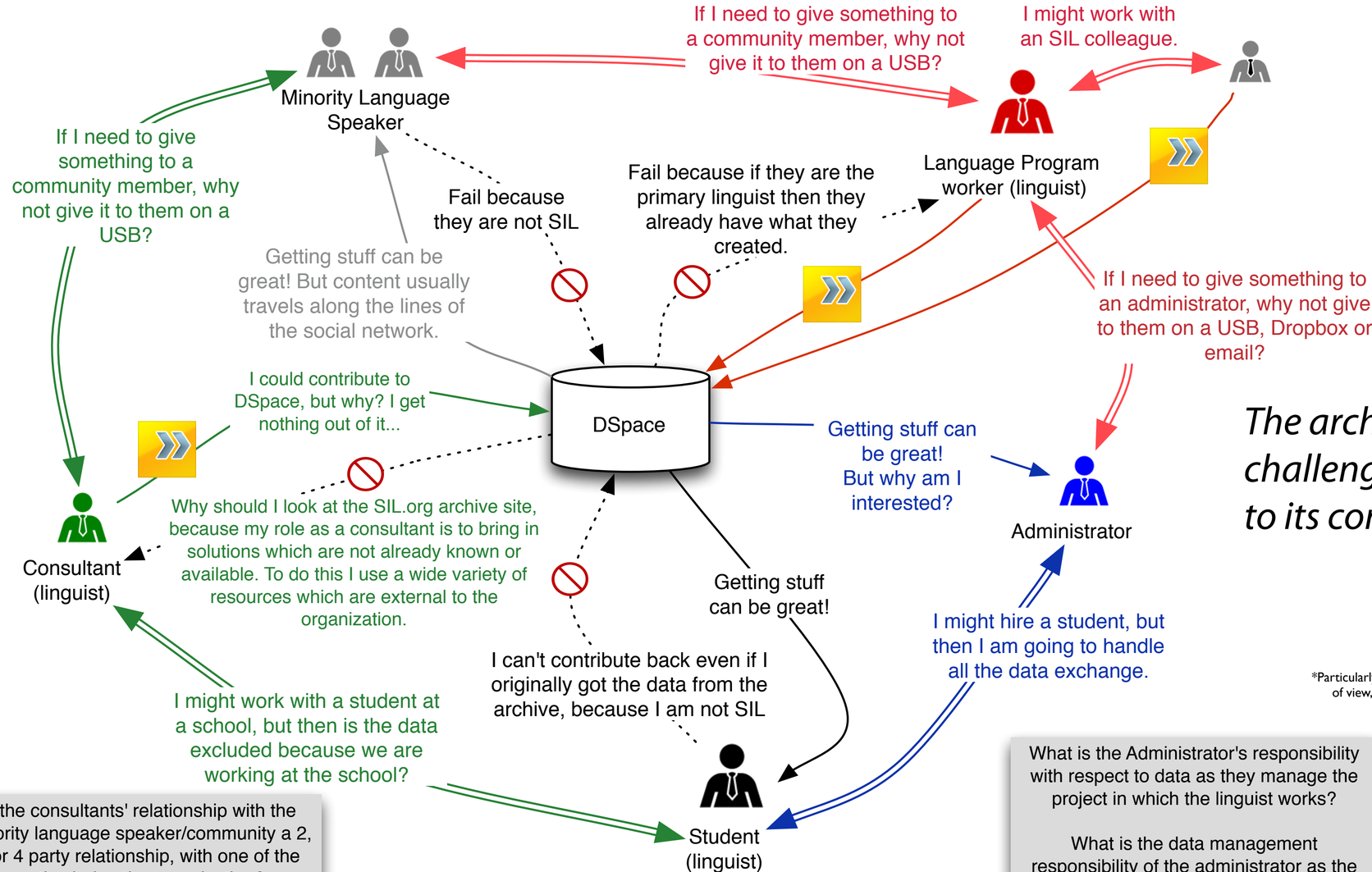
In SIL the challenge for the archive is to prove its worth to several different user groups.

- 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?

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.

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.

*Particularly diagramed from an SIL point of view, relating to the SIL archive

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?

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 the manage the relationship between the community and the organization?

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.*

References

Anderson, Stephen P. 2006. Creating Pleasurable interfaces: Getting from Tasks to Experiences (Poster). <Accessed: 18 February 2013> http://www.artificialindustry.com/uploads/inspiratie/1295515106from-task-to-experience.png

Anderson, Stephen P. 2009. Fundamentals of Experience Design (Poster). <Accessed: 18 February 2013> http://www.poetpainter.com/thoughts/files/Fundamentals-of-Experience-Design-stephenpa.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 99.6: 4115-20. http://www.pnas.org/content/99/6/4115.abstract

Nash, Douglas. 2012. Businessperson. The Noun 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/repository/LDLT3_Nordmoe_preprint.pdf

Onori, P. J. 2009. List. The Noun Project.

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 or http://dx.doi.org/10.2139/ssrn.249585

Scribner, Matt. 2012. Arrow. The Noun Project.

Shmidt, Sergey. 2013. Database. The Noun Project.

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/