Lexical dataset archiving: an assessment of practice

Hugh Paterson III

1 The *outrageous* claim

Millions of language artifacts are not being archived

Paterson and Nordmoe (2013) made the following claim:

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.

This claim was made solely on the bases of working with data in SIL International's network of staff over the course of 4-5 years. An outstanding question remains: *Can the estimated ratio be applied more generally to all linguistic researchers, or is it subject to network constraints, and therefore limited to only SIL staff?*

I seek to answer the question:
is the volume of unarchived and endangered resources a localized behavioral attribute within a particular social network or is it a more general sociological phenomenon?

2 Testing the equivalent network hypothesis

Before asserting that the archiving behavior of linguists is dependent on specific factors such as project funding requirements, or social network affiliation (the Academy vs. NGO), an assessment needed to be made. My sampling methods attempted to included a cross-network sample of linguists and language program workers. A single data type - *the lexical dataset* - was chose as "representative" of archive worthy documentary evidence (Woodbury 2003). An online questionnaire was developed and a request for voluntary participation was sent to a variety of mailing lists. ***176 people responded***; indicating knowledge about ***370 lexical data sets***. Of the respondents, ***96 were SIL staff*** and ***80 were not affiliated with SIL***.

MAILING LIST	DATE SENT
SIL-LDL Mailing list	15. November 2013
SIL-Survey Mailing list	15. November 2013
ANU Austronesian Mailing list	16. November 2013
Yahoo! Lexicography List	17. November 2013
RNLD list	18. 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 Mailing list	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
SIL Linguistics Coordinators	24. May 2014
SIL Africa Computing Mailing list	24. May 2014
SIL Malaysia Branch	20. June 2014

Questions asked

- What Lexical Database Solution do you use?
  FLEx, Toolbox, Lexus, Other:____
- What is the language of study in your Lexical Database?
  ISO 639-3 code
- Have you archived your database?
  No. - Never Archived it, with SIL, with ELAR, with TLA, with PARADISEC, other:_____.
- Email address
- Is this an SIL Project?
  Yes/No
- If you are using FLEx or ToolBox have you produced a Print publication?
- Anything else we should know?

Contribute to the Data

Fill out the questionnaire!

http://bit.ly/19QSPMb

What if there is a lexical database in an archive but no-one responded about it?

Use cases where a linguist may have already archived lexical data sets, but were disinclined to responded directly to the questionnaire were attempted to be included. On this account, three archives were contacted and given the opportunity to participate: ELAR/SOAS, PARADISEC, SIL International's Language & Culture Archive. Each archive declined to provide information, citing either: privacy concerns, lack of a detailed indexing procedures on curated archive holdings, or lack of staffing to sufficiently answer the question. All three archives suggested that the best results might be achieved through searches at OLAC - an aggregate records listing of participating language archives. ***Some 22 records were found via OLAC*** *but were not included in this presentation of the results. A manual tally of* ***some 70 records of lexical databases in the SIL archive's catalogue****, not aggregated to OLAC, were also included. This brings the cumulative token count for lexical data sets to* ***476 tokens****.*

Suggested Citation:

Paterson, Hugh J, III. 2015. *Lexical dataset archiving: an assessment of practice*. Poster presented at the 4th International Conference on Language Documentation and Conservation, at the University of Hawai'i Mānoa, Honolulu, HI. February 28 – March 3rd. *Version 0.9*

Comments invited via Hugh.Paterson@sil.org

References:

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 1.5*

Paterson, Hugh J, III. 2015. Scripts for statistic tests. available via github: https://github.com/HughP/Lexical-Database-Archiving-Stats Accessed: 11. February 2015

Woodbury, Anthony C. 2003. Defining Documentary Linguistics In Peter K. Austin (ed.), Language Documentation and Description, vol. 1, 35-51. London: SOAS.

3 Responses and analysis

370 Resources were discussed

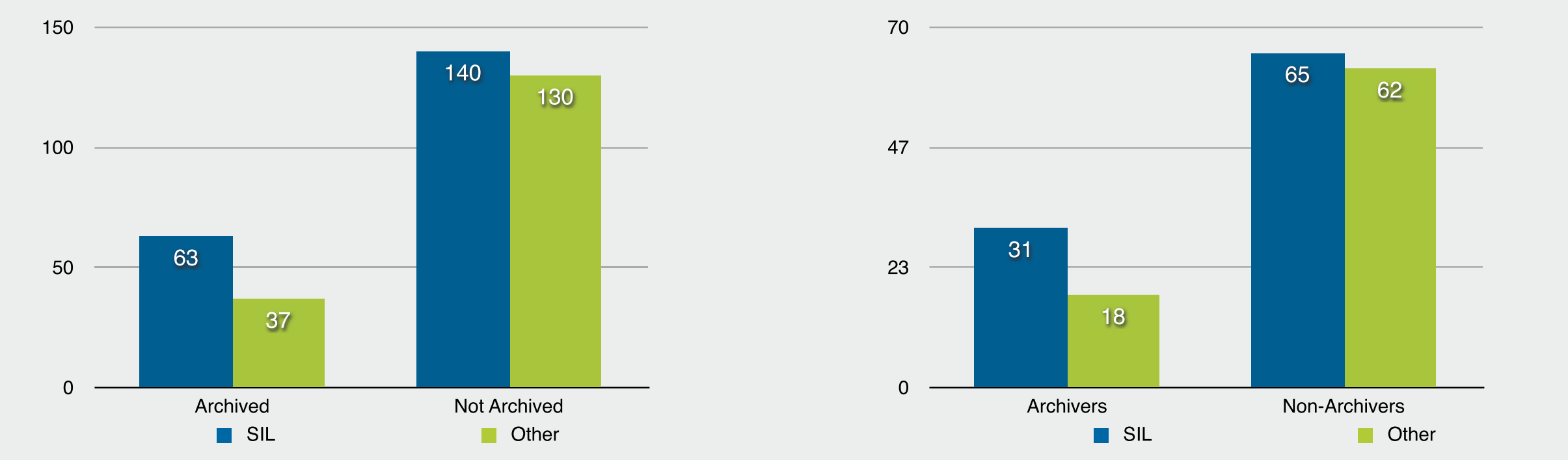
Resources	SIL	Other	*Cumulative*
Archived	**63**	**37**	*100*
Not Archived	**140**	**130**	*270*
Cumulative	*203*	*167*	

Fisher's exact test *p value: 0.06045*

176 people were involved in the discussion

Respondents	SIL	Other	*Cumulative*
Archivers	**31**	**18**	*49*
Non-Archivers	**65**	**62**	*127*
Cumulative	*96*	*80*	

Fisher's exact test *p value: 0.17773*



31% of SIL resources discussed were claimed to be archived

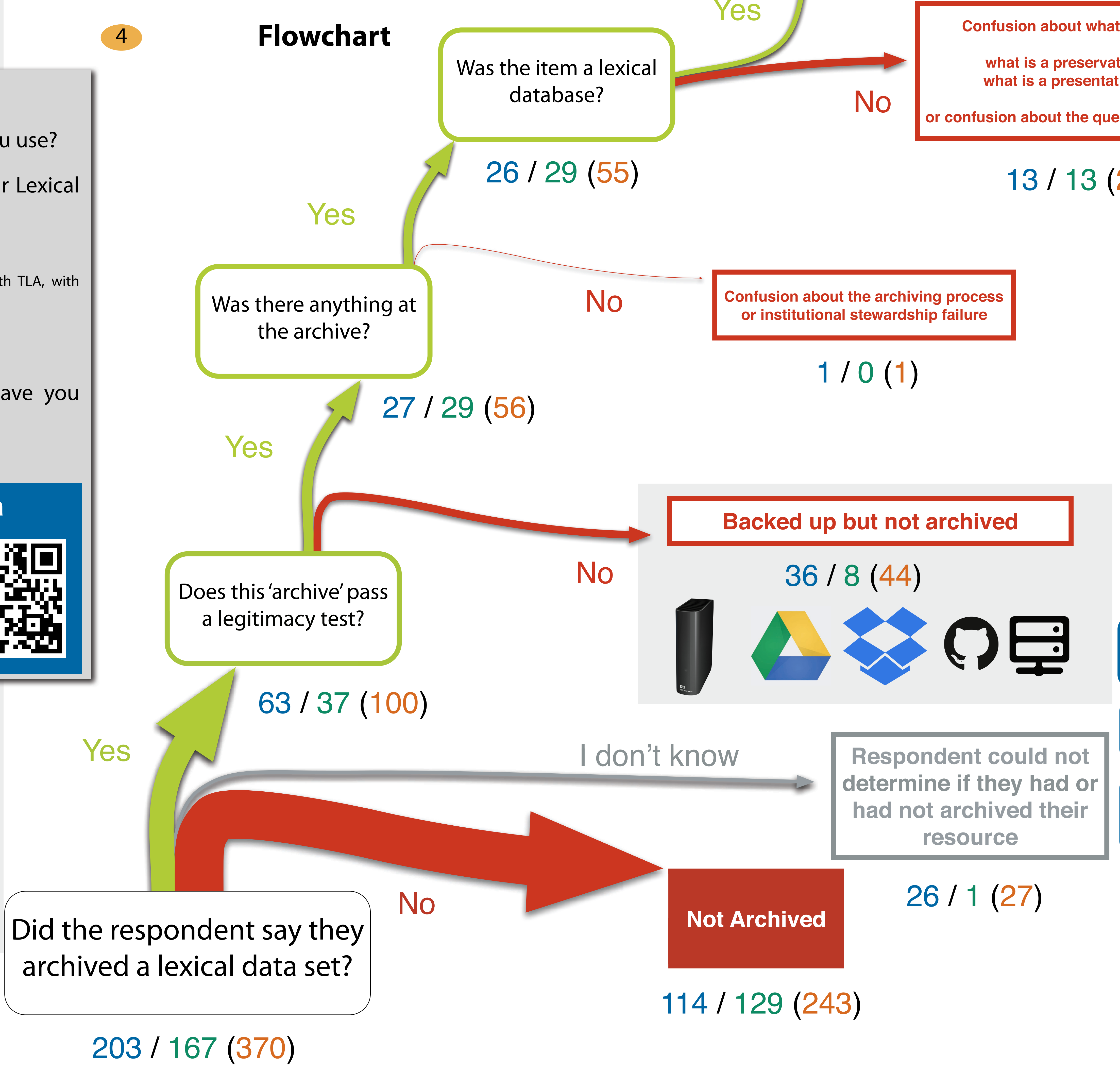
22% of Other resources discussed were claimed to be archived

32% of SIL respondents claimed to have archived

22% of Other respondents claimed to have archived

Number of respondents	134	19	8	4	1	1	1	1	1	1	1	1
Number of lexical database records provided	1	2	3	4	6	7	10	12	15	21	31	37

4 Flowchart



Did the respondent say they archived a lexical data set?

203 / 167 (370)

Not Archived

114 / 129 (243)

Respondent could not determine if they had or had not archived their resource

26 / 1 (27)

Does this 'archive' pass a legitimacy test?

63 / 37 (100)

Backed up but not archived

36 / 8 (44)

Was there anything at the archive?

27 / 29 (56)

Confusion about the archiving process or institutional stewardship failure

1 / 0 (1)

Was the item a lexical database?

26 / 29 (55)

Archived Lexical Dataset

Kinds
FLEx
.lift file
.fwdata file
Shoebox
Toolbox
MSAccess
FileMaker Pro
Other

13 / 16 (29)

Confusion about what to archive, i.e.

what is a preservation version,
what is a presentation version,

or confusion about the questions in the survey

13 / 13 (26)

Some responses pointed to an archived "dictionary". Such resources might have been the PDF of a printed book or it might have been a literacy type *picture dictionary*. These were not counted as *lexical databases*, rather are the derivative products of *lexical databases*.

Lexique Pro datasets were marked as distribution formats and not as lexical databases therefore categorically ended up here.

Is getting materials to the archive enough?

What does data ecology in linguistics call for?

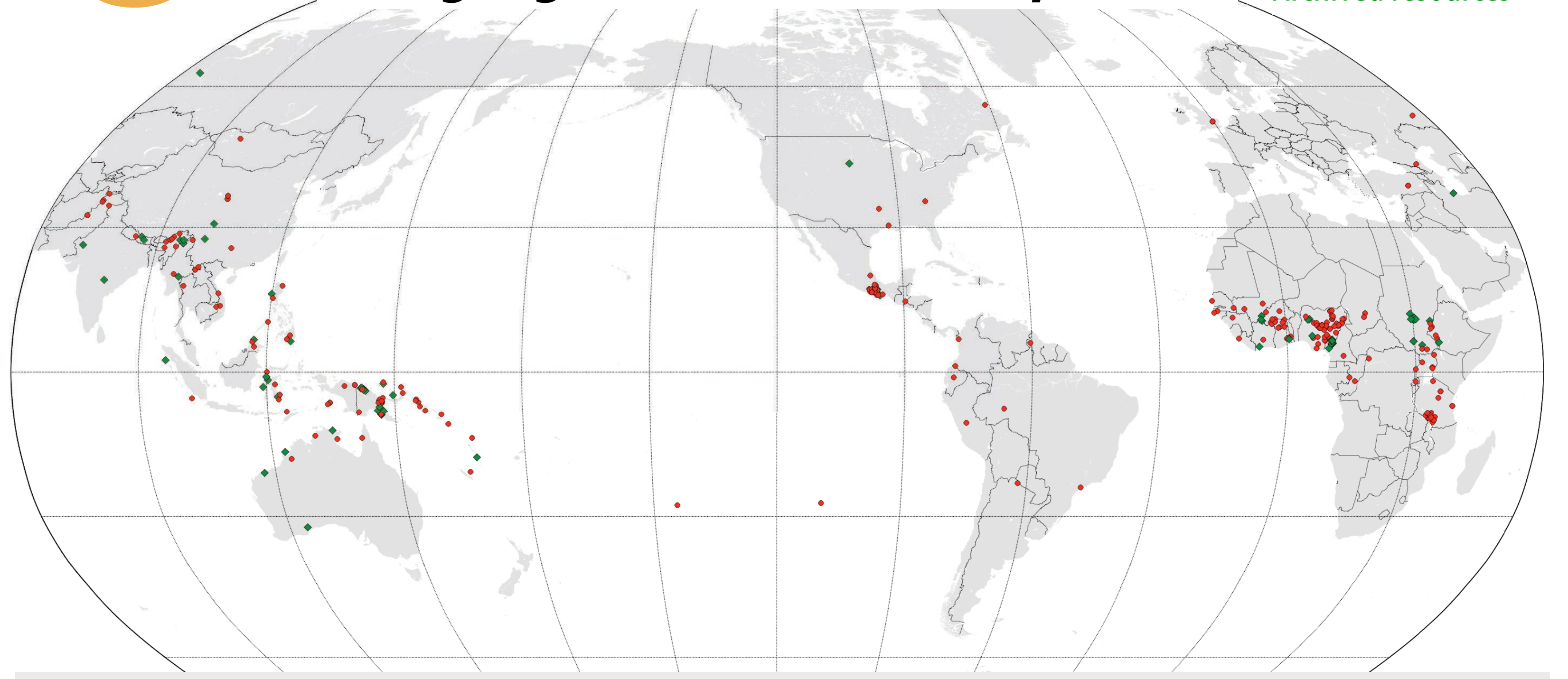
Many respondents pointed to a lack of perceived value in archiving. This may be a contributing factor in lower participation levels. It may be that this perception by potential archive users is warranted. Do we need an archiving model which incorporates the *Share - Enhance - Return* workflow?

Two outstanding issues were observed about lexical databases/ datasets which were archived:

Overall resources are not described as a Lexical database/ datasets.

Often clear paths for digital object acquisition, enhancement and return to the archive are not clear from archive catalogue listings.

5 *Languages mentioned in responses*



6 *Conclusions*

My FLEx database includes some vulgar terms that would not likely be in a public dictionary, so I would want to delete those before uploading it to a public archive.

Have I Archived?:
No (never took the time to find out how to do it)

We use languagedepot.org for project sharing. Does that count (as archiving)? - Otherwise, no (slow internet is a pretty big issue here).

Linguists are opinionated about the merits of archiving:

I think I haven't thought about it because it is still early in the project.

Even though my Toolbox file is a bit of a mess, I'd rather people in the future have a useful mess than nothing at all!

I have not put any of my work into institutional archives. I suppose that I should, but to figure out where and how to do this would take time and research that I haven't wanted to spend time on. I figure that as long as I make work that might be useful available for open access to anyone, I am doing enough. To tell the truth, I have a rather low opinion of these big archives. For example, the IMDI archive is such an incomprehensible mess that I don't see how it could be useful to anyone. I believe that my stuff is straightforward, easy to download, and free to everyone. It might seem kind of hidden away on a personal website, but people seem to find it, and my work is a niche field, so most people working in that niche know about it.

It is my conclusion that the prevalence of endangered resources is not a situation limited to a particular social network. The *p* value of the respondents suggest that the networks are different and that the "other" category is less likely to archive than the "SIL" category. Both networks exhibit similar quantities of known unarchived resources, as reported by similar numbers of non-archivers. I take this evidence to support the hypothesis that claims in Paterson and Nordmoe (2013) can be extended and applied more generally to other networks of linguists.