Beyond the Repository:
Integrating Local Preservation Systems with National Distribution Services

INSTITUTE *of*
Museum and **Library**
SERVICES

LG-72-16-0135-16

EVVIVA WEINRAUB

EVVIVA.WEINRAUB@NORTHWESTERN.EDU

LAURA ALAGNA

LAURA.ALAGNA@NORTHWESTERN.EDU

Beyond the Repository:
Goals

- Investigate common problems in digital object curation, versioning, and interoperability between local repositories and distributed preservation systems

- Identify broadly applicable use cases and design patterns

- Propose high-level technical solutions

Beyond the Repository:
People and institutions

Northwestern University
 Evviva Weinraub (PI)
 Carolyn Caizzi
 Laura Alagna
 Brendan Quinn
 Gina Petersen

University of California San Diego
 Sibyl Schaefer

Advisory Board

Mike Giarlo (Stanford)

Bert Lyons (AVPreserve)

Mary Molinaro (DPN)

Mike Ritter (University of Maryland)

Justin Simpson (Artefactual)

David Wilcox (Fedora/DuraSpace)

Andrew Woods (Fedora/DuraSpace)

Beyond the Repository:
Research questions

- How does one curate objects to ingest into a long-term dark preservation system?

- How does versioning of objects and metadata play out in long-term dark preservation systems and how to automate these actions?

- How can systems that store data differently be made more interoperable?
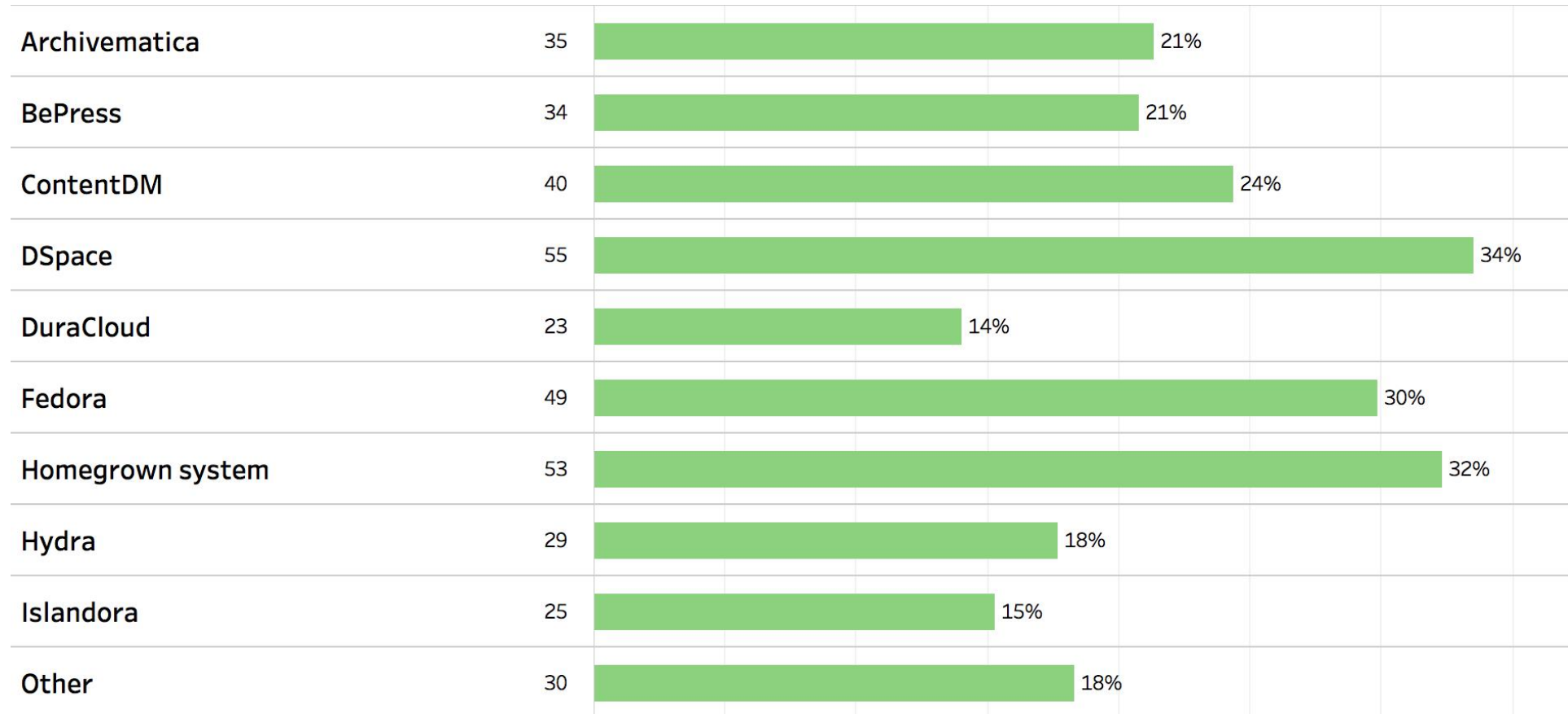
Beyond the Repository:
Methodology

1. Gather information on the first two research questions via a survey of practitioners

      a. Understand the breadth of implemented local systems

      b. Identify local workarounds and metadata fixes in place to address these issues

      c. Gather data about local preferences around versioning

      d. Identification of preservation policies and rights issues

2. Hold a series of in-depth interviews to gather additional qualitative information

3. Using this data, work with the Advisory Board to design high-level requirements for increased interoperability between local and distributed systems

4. Disseminate findings

Results: survey metrics

- 170 valid responses

- 65% have collected 10 TB or more

- More than 80% expected their content to grow by at least 10 TB in the coming year

- Wide geographic distribution represented, including 15 international responses

- Mostly academic libraries (77%)

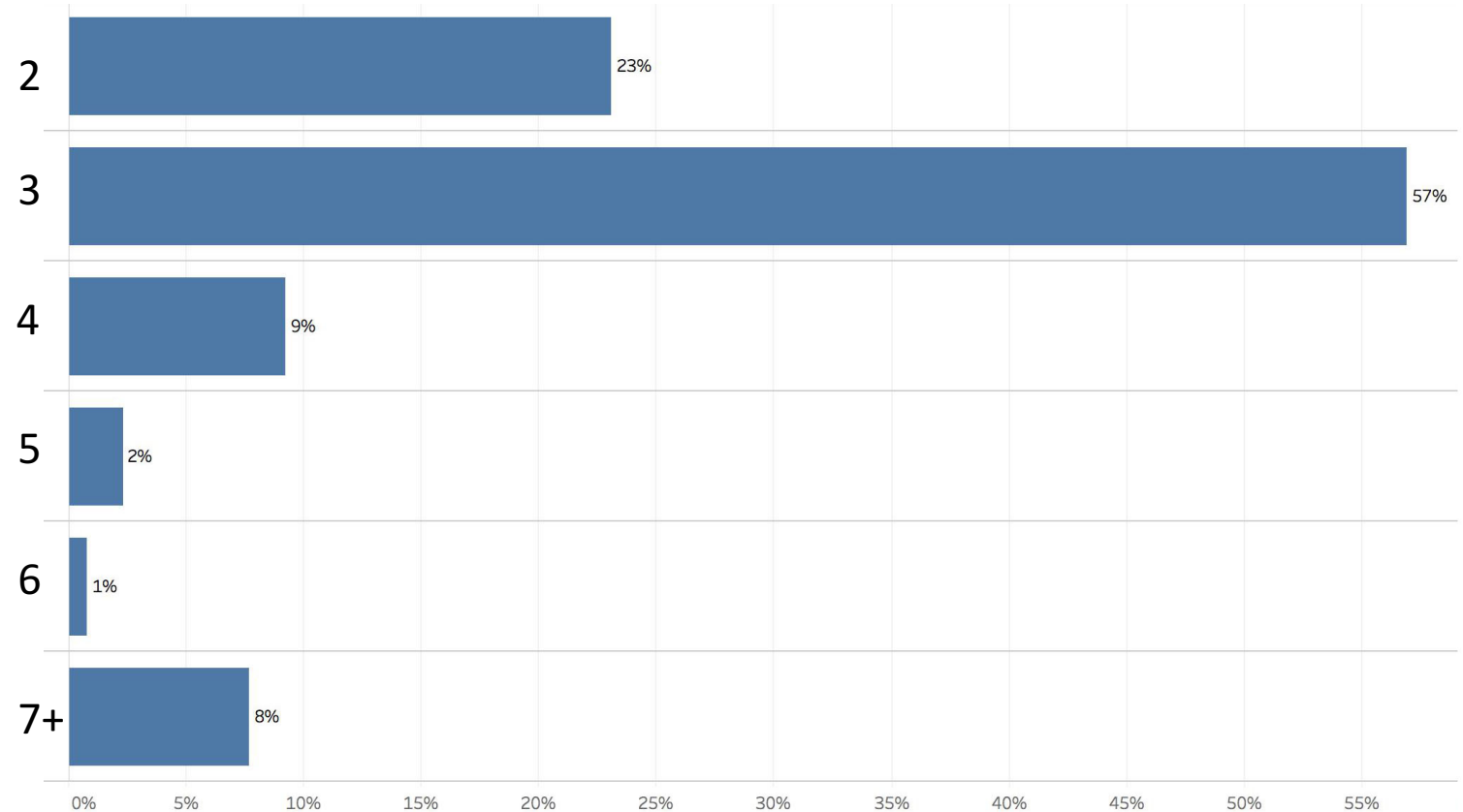- 73 people were willing to discuss further with us

Survey results:
Systems used

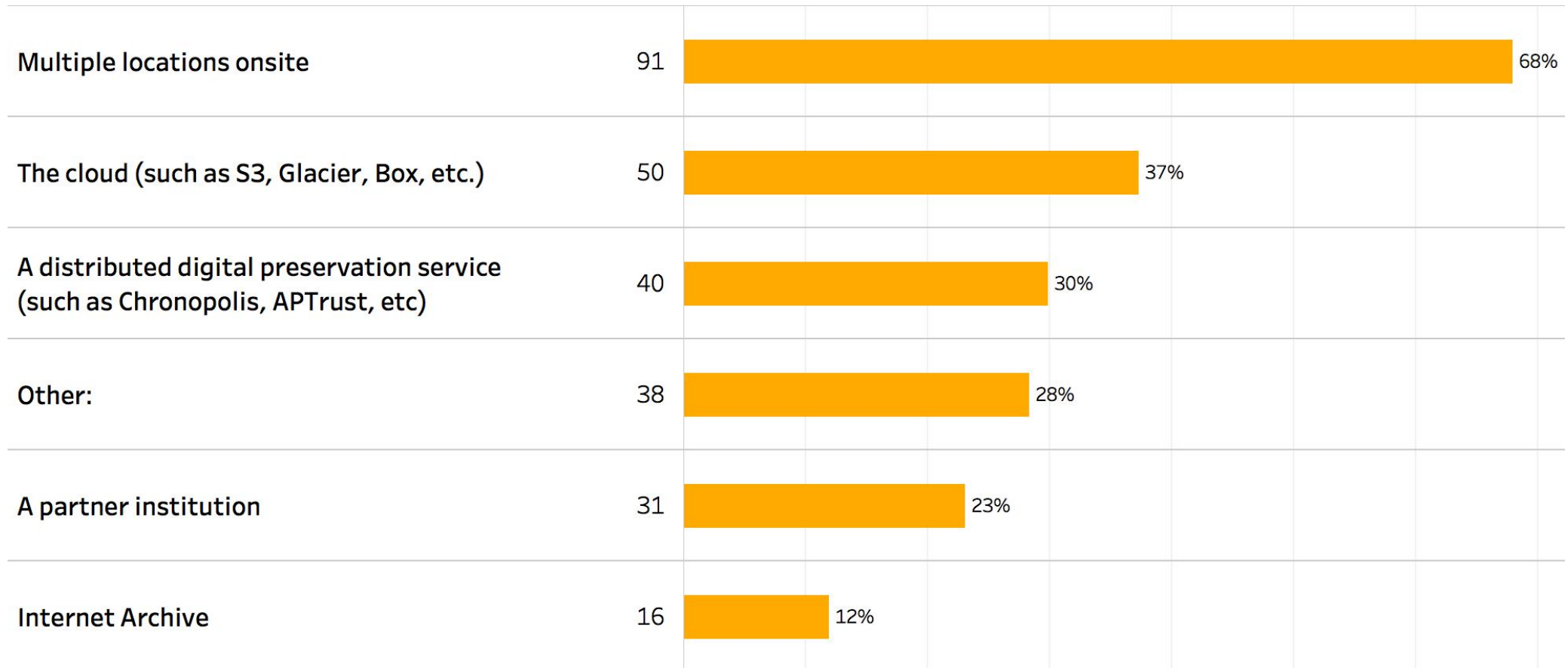| System	Count		Percentage
Archivematica	35		21%
BePress	34		21%
ContentDM	40		24%
DSpace	55		34%
DuraCloud	23		14%
Fedora	49		30%
Homegrown system	53		32%
Hydra	29		18%
Islandora	25		15%
Other	30		18%

Survey results:
Distributed storage & number of copies

- Respondents who reported not keeping multiple copies cited funding as the most common barrier
- 85% of respondents reported keeping multiple copies in multiple locations
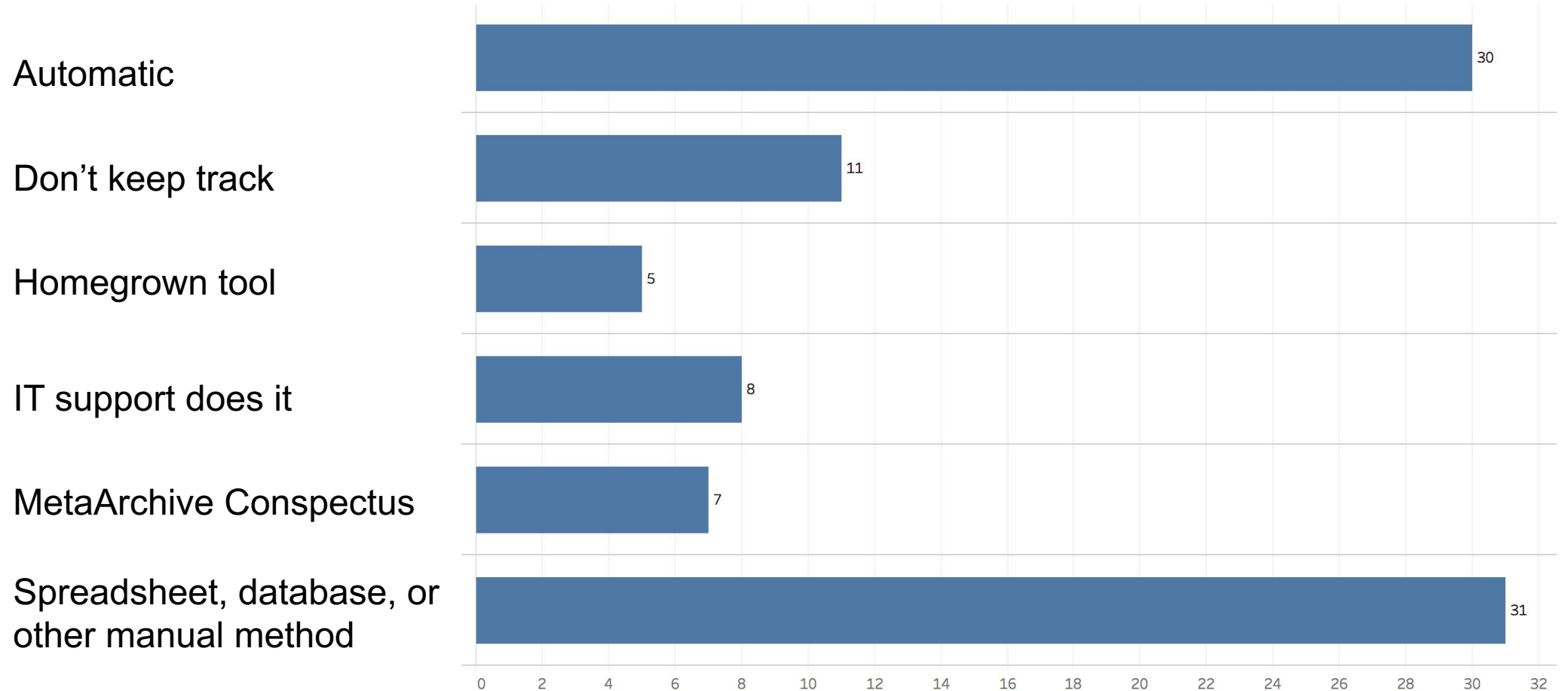- Of these, the vast majority keep three copies

Survey results:
Where copies of data are stored

Multiple locations onsite	91	68%
The cloud (such as S3, Glacier, Box, etc.)	50	37%
A distributed digital preservation service (such as Chronopolis, APTrust, etc)	40	30%
Other:	38	28%
A partner institution	31	23%
Internet Archive	16	12%

Survey results:
How copies are tracked

Category	Value
Automatic	30
Don't keep track	11
Homegrown tool	5
IT support does it	8
MetaArchive Conspectus	7
Spreadsheet, database, or other manual method	31

Survey results:
Versioning & curation

When versioning distributed copies:
- 85% of respondents reported keeping all versions
- 20% reported only keeping the newest version
- 20% were unsure
- Many indicated that versioning practices are dependent on the type of materials

In terms of selection:
- 48% of respondents say they select a subset of materials to go to a distributed repository
- The top two selection criteria for these materials were:
  - Mandate (legal, grant, or other)
  - Intrinsic value

Interviews: a snapshot

- 12 institutions:
  - 6 public university libraries
  - 2 private university libraries
  - 2 museums
  - 1 public library
  - 1 government archives

- Interviewees collectively use 8 different local repository systems and four different distributed digital preservation systems

Interview trends:
Versioning & curation

"I think our versioning has been somewhat haphazard rather than deliberate."

"It's this real manual versioning going on, but it's not really even true versioning. It's not recording exactly what was changed."

"We can't rely on the curators yet to help us with those value choices... it kind of falls to us to make some of those decisions, and we don't feel qualified to know what's more valuable, so it's kind of messy right now, and it probably is going to need some coordination in the organization to sort of get that right."

Interview trends: interoperability

"I think interoperability itself is the main challenge that we're facing, to be able to get these different systems to work together, whether it's our descriptive systems or preservation."

"Right now, nothing is actually interacting together."

"In a sense, our workarounds are just doing things manually."

Interview trends:
Brutal honesty

"In terms of any sort of catastrophic event, we're toast pretty much."

"We've been around since 1849 and this is the first time the institution has acknowledged that preservation is worthy of a full time position."

"It's really hard to convince stakeholders that [digital preservation] is something that's worth spending money on. It's not glamorous, it's invisible…there's just so many other competing things that are flashier things to spend money on."

Next steps

September/October: Report writing

October: Advisory board meeting

December: Report dissemination

Thank you

INSTITUTE *of*
Museum and **Library**
SERVICES

LG-72-16-0135-16