Beyond the Repository:

Integrating Local Preservation Systems with National Distribution Services

LG-72-16-0135-16

SIBYL SCHAEFER
SSCHAEFER@UCSD.EDU

EVVIVA WEINRAUB
EVVIVA.WEINRAUB@NORTHWESTERN.EDU

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

People and Institutions

University of California San Diego
Sibyl Schaefer

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

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)

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?

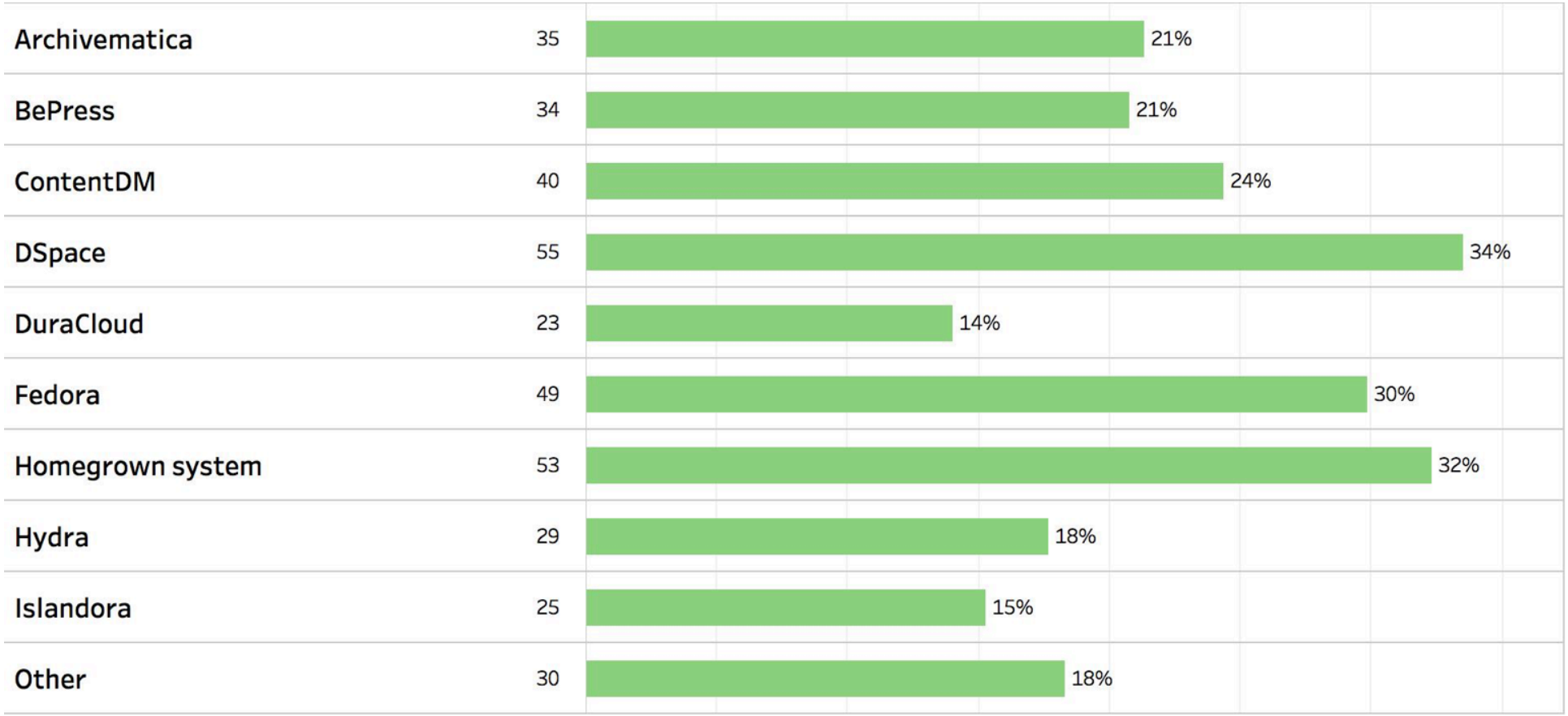
Methodology

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 and curation
        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 and recommendation

Survey Results

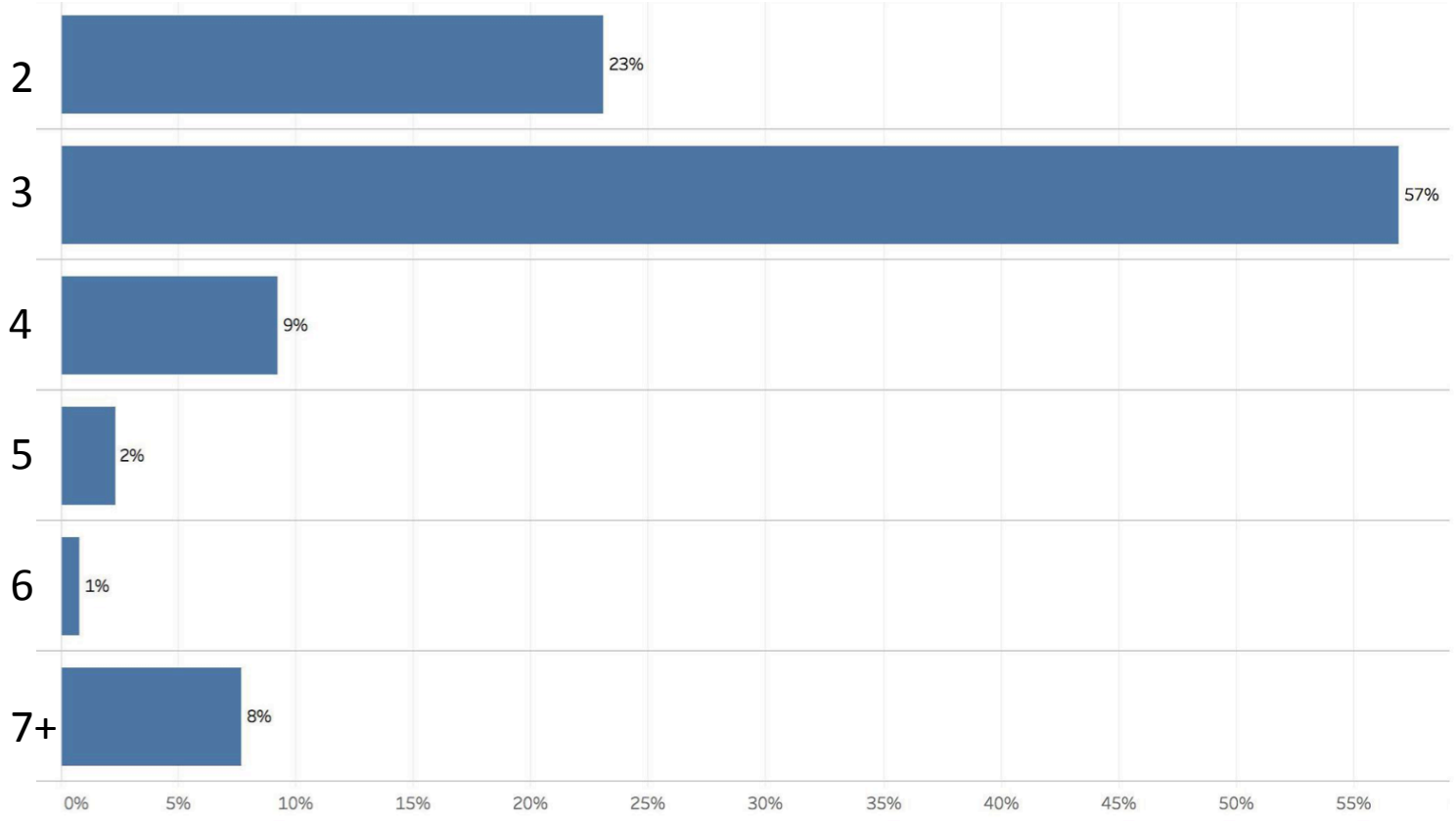
- 170 valid responses
- 65% have collected 10 TB or more
- More than 80% expected their content to grow by at least 10 TB in the coming year
- Wide geographic distribution represented, including 15 international responses
- Mostly academic libraries (77%)
- 73 people were willing to discuss further with us

Systems used

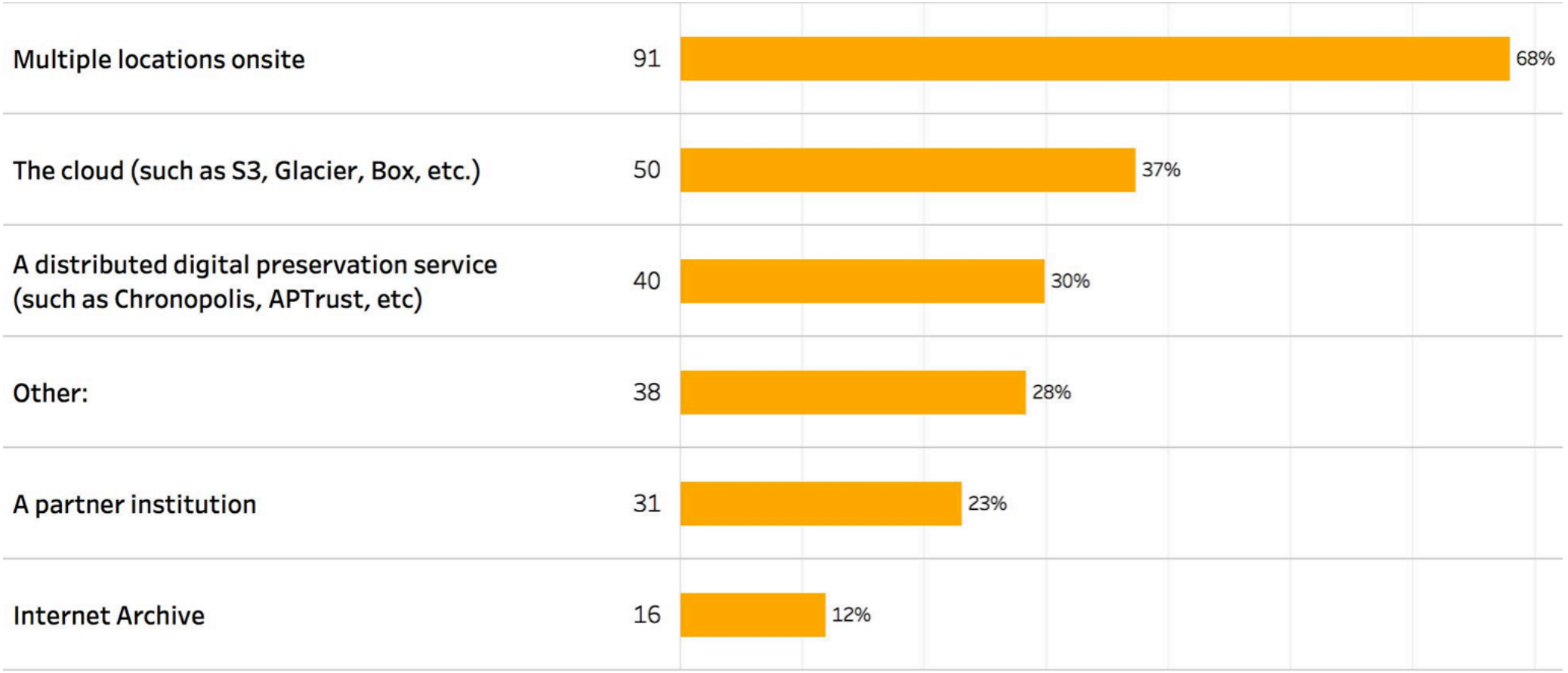


Distributed storage and number of copies kept

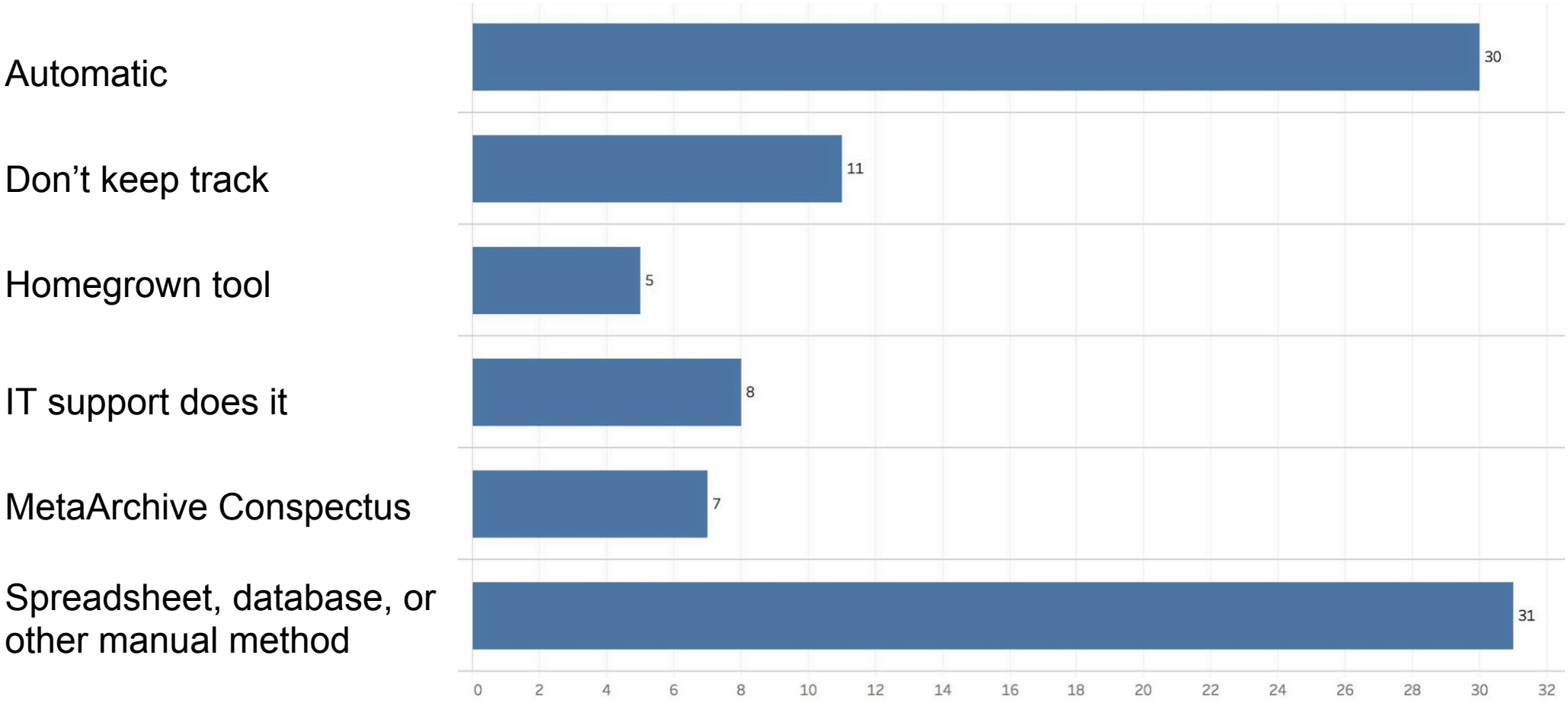
- 85% of respondents reported keeping multiple copies in multiple locations
- Of these, the vast majority keep three copies
- Funding cited as the most common barrier for not keeping multiple copies



Where copies are stored



How copies are tracked



Versioning

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

Curation

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

12 institutions:

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

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

Interviews: Curation

"... well if the risk involved in losing that content would mean losing that information entirely, then I want that content to be preserved in MetaArchive."

"… is it born digital? Is it something that has preservation needs? So if we scan something that is super fragile, or that is in an unstable format, or if it's something that we pay for a vendor to reformat, like a reel to reel tape or something, those are all going to be things we do full-fledged, full-on digital preservation for. But if …. we're just scanning to make available online for researchers, we're not going to do full on digital preservation because it's too expensive."

Interviews: Versioning

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

"...what are the real use cases? For special collection stuff, all versions are first class objects. That's almost the whole point of some of these things. If somebody has five manuscripts of somebody's book, well they're all equally important. There's no question of versioning. On the respective digitization side, it's way too expensive to go back and produce another TIFF, if that ever happens, because the first group of TIFFs were a mistake, you don't want to keep them. For digital archives, versioning is moot. For digital collections, it's too expensive."

Interviews: 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."

Interviews: 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.”

Conclusions

- How does one curate objects to ingest into a long-term dark preservation system?
  - Mandate and Intrinsic Value are important
  - Some consensus that there are different tiers of digital preservation - not all material is worth preserving at the highest tier and that the highest tier of preservation is a distributed digital preservation service.

- How does versioning of objects and metadata play out in long-term dark preservation systems and how to automate these actions?
  - No consistent versioning practices locally, let alone in distributed systems

- How can systems that store data differently be made more interoperable?
  - Bags are most common form of data packaging

Recommendations

- Shared Bagit specifications for local repository systems
- Standardized API for distributed digital preservation services
- Curation decision tools

Next steps

2017:	Report writing
2018:	Report dissemination
2018:	Apply for an IMLS Implementation Grant

Thank you!

LG-72-16-0135-16