Beyond the Repository:

Integrating Local Preservation Systems with National Distribution Services

LG-72-16-0135-16

LAURA ALAGNA

LAURA.ALAGNA@NORTHWESTERN.EDU

@DIGITIZED_LAURA

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?

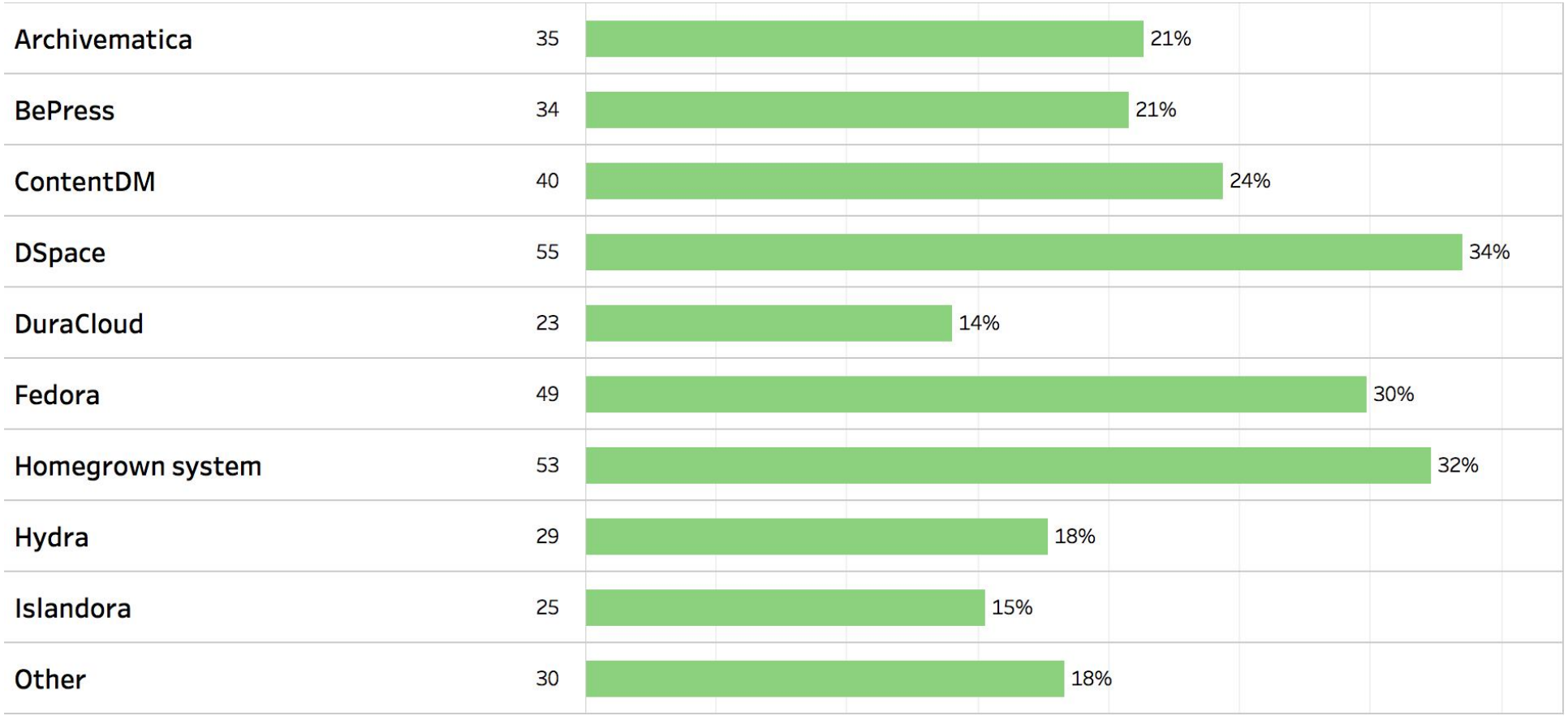
Beyond the Repository:
Methodology

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

Preliminary results: survey metrics

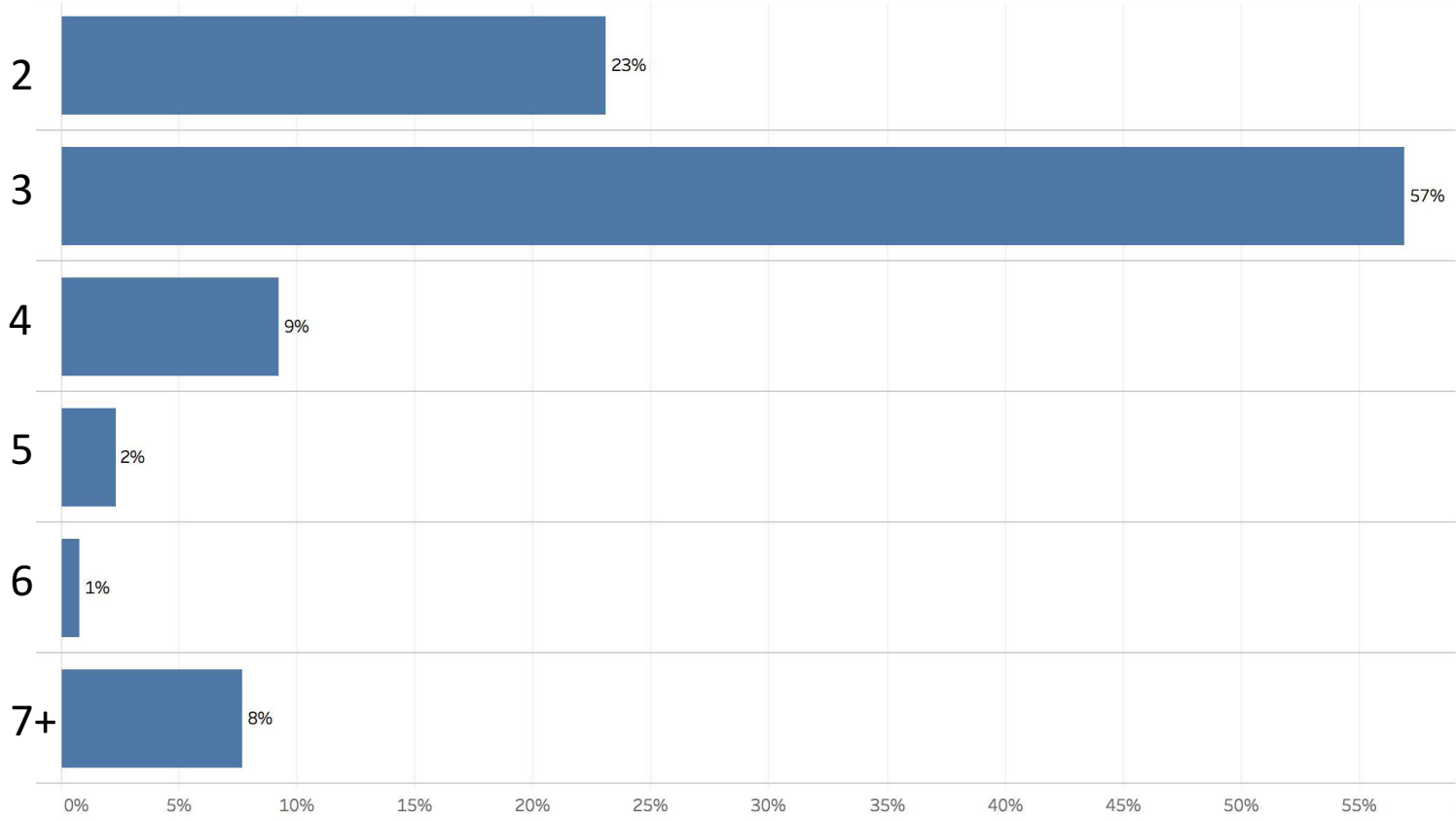
- 170 valid responses
- 65% have collected 10 TB or more
- More than 80% expected their content to grow by at least 10 TB in the coming year
- Wide geographic distribution represented, including 15 international responses
- Mostly academic libraries (77%)
- 73 people were willing to discuss further with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

Survey results:
Systems used

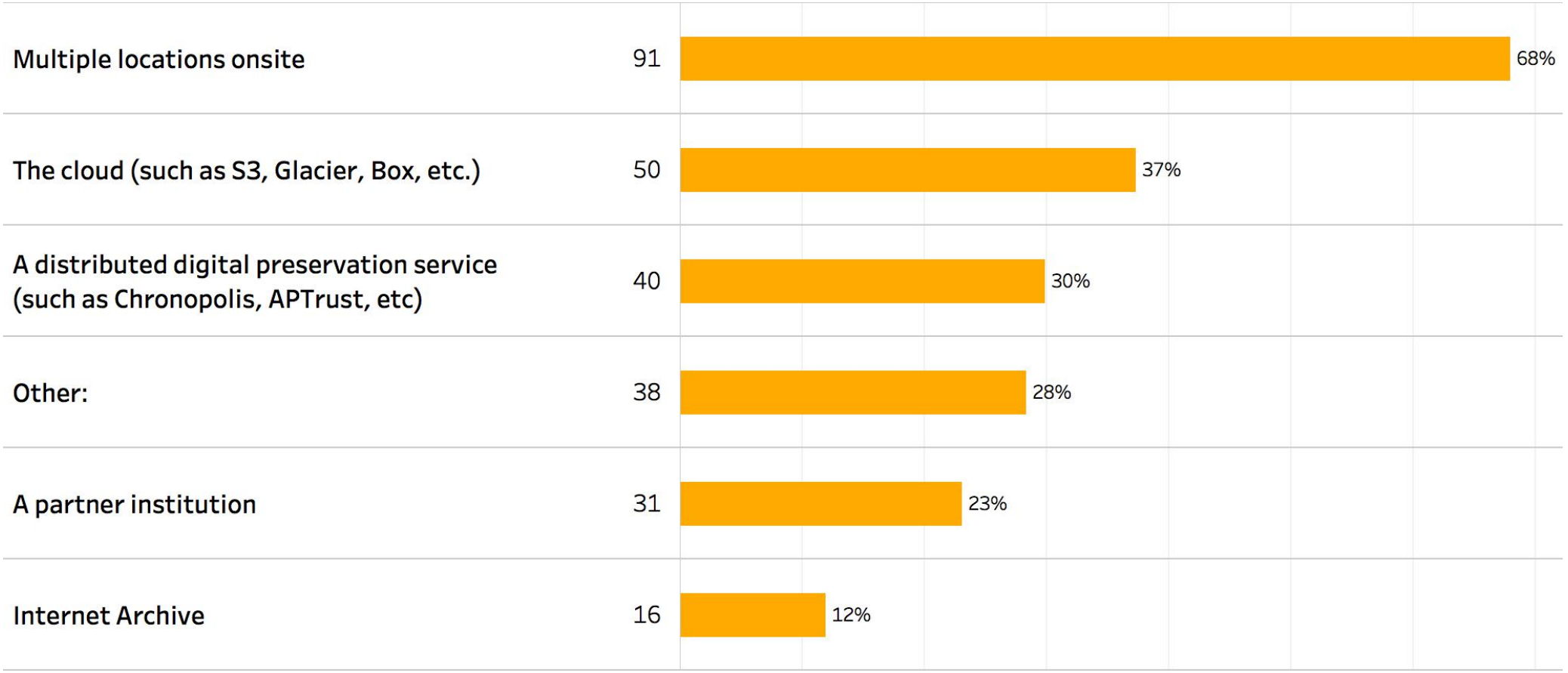


Survey results:
Distributed storage & number of copies

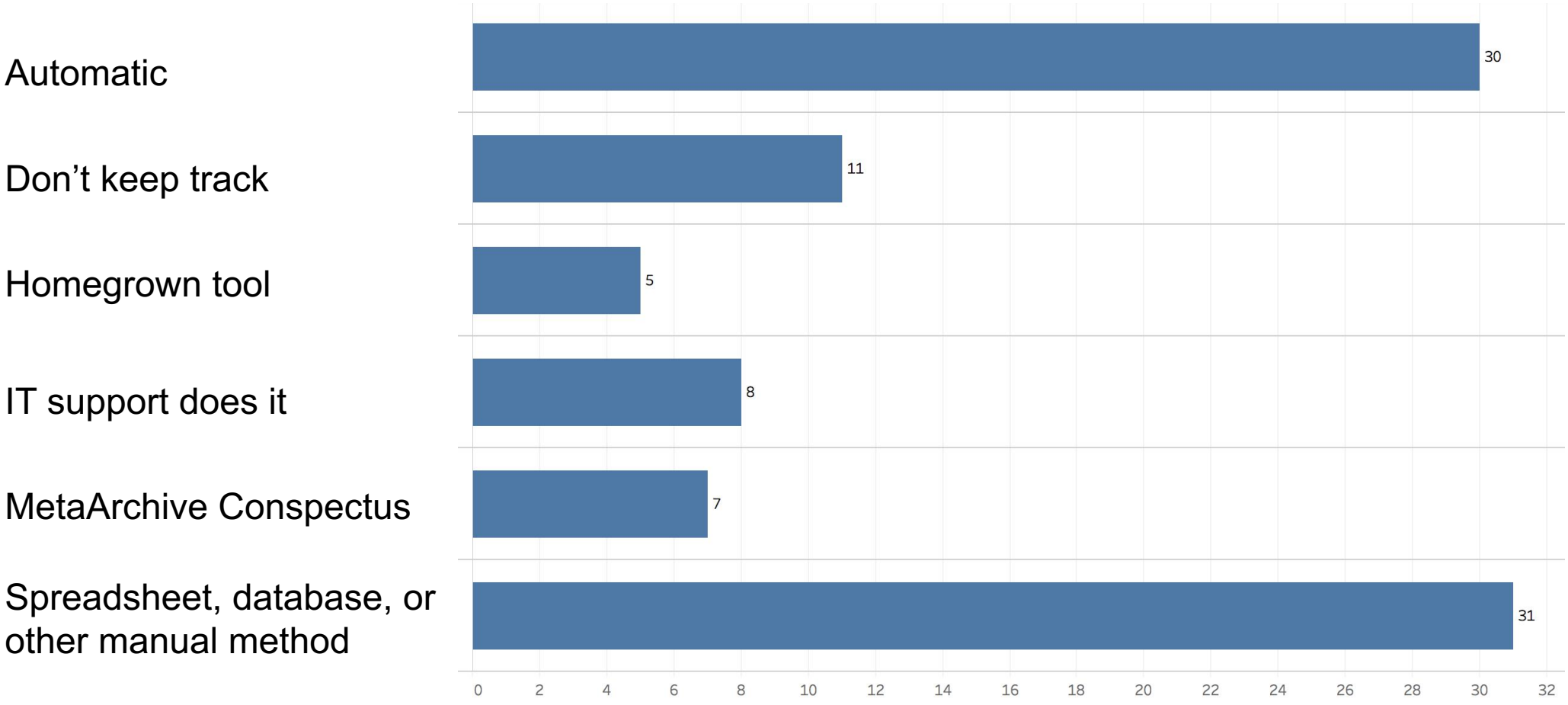
- Respondents who reported not keeping multiple copies cited funding as the most common barrier
- 85% of respondents reported keeping multiple copies in multiple locations
- Of these, the vast majority keep three copies



Survey results:
Where copies of data are stored



Survey results:
How copies are tracked



Survey results:
Versioning & curatorial decisions

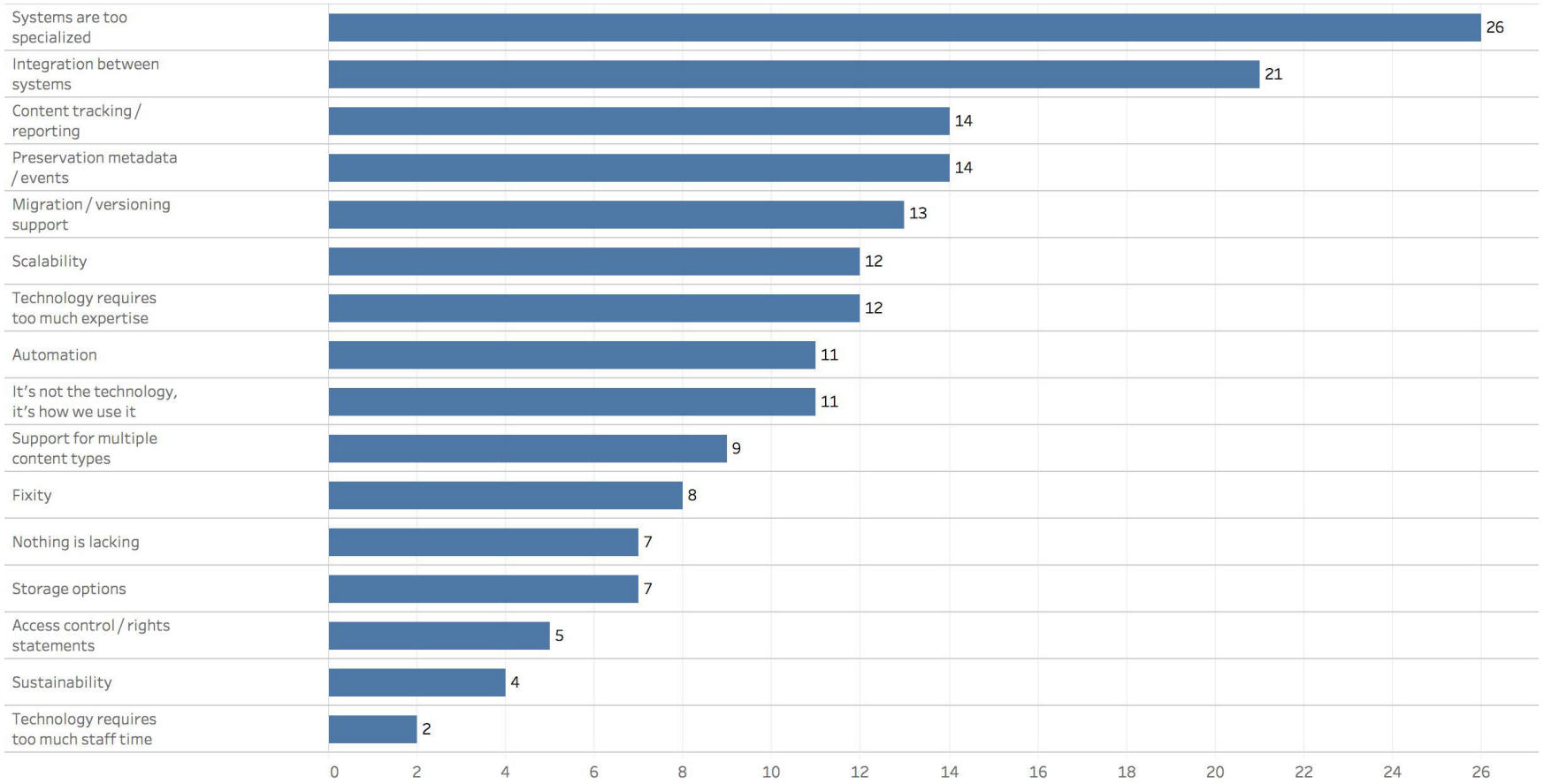
When versioning distributed copies:

- 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

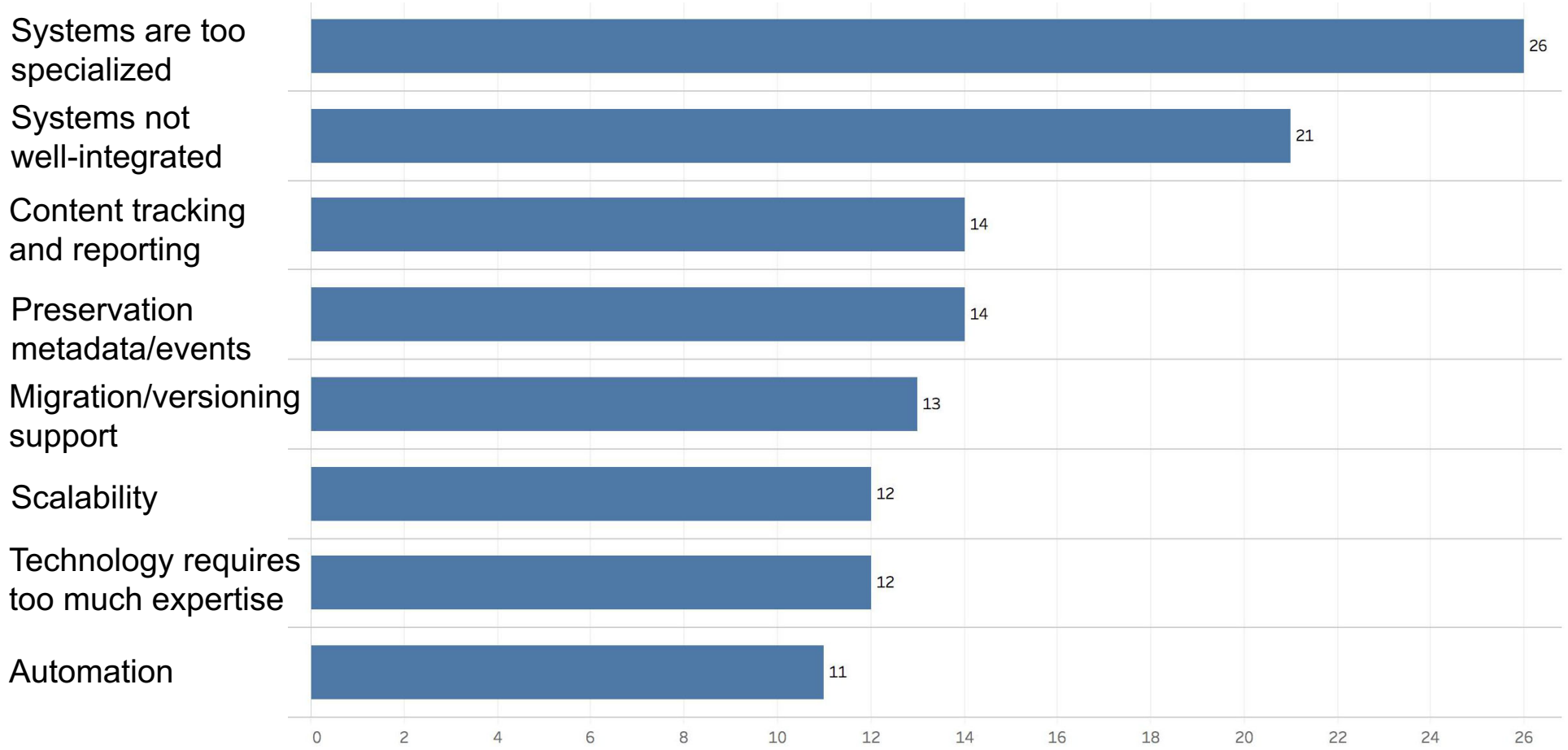
In terms of selection:

- 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

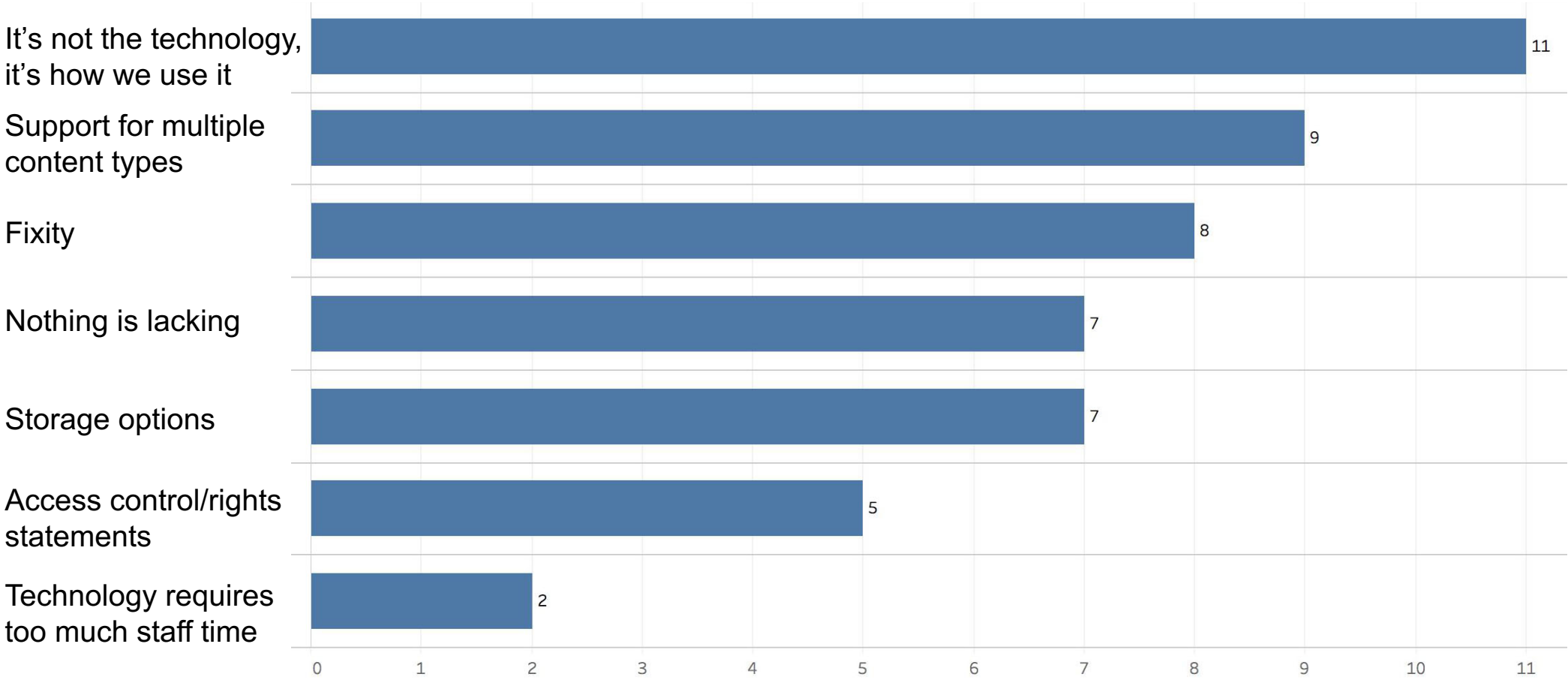
Survey results:
What is lacking in current tools and services



Survey results:
What is lacking in current tools and services



Survey results:
What is lacking in current tools and services



Next steps

July / August:	Interviews
August:	Interview analysis
September/October:	Report writing
October:	Advisory board meeting
December:	Report dissemination

Thank you

LG-72-16-0135-16